Quasi-parametric recovery of Hammerstein system nonlinearity by smart model selection

Zygmunt Hasiewicz and Grzegorz Mzyk[1] and Przemysław Śliwiński

Institute of Computer Engineering, Control and Robotics
Wrocław University of Technology
Janiszewskiego 11/17, 50-372 Wrocław, Poland
[1]e-mail: grzegorz.mzyk@pwr.wroc.pl

Abstract. In the paper we recover a Hammerstein system nonlinearity. Hammerstein systems, incorporating nonlinearity and dynamics, play an important role in various applications, and effective algorithms determining their characteristics are not only of theoretical but also of practical interest. The proposed algorithm is quasi-parametric, that is, there are several parametric model candidates and we assume that the target nonlinearity belongs to the one of the classes represented by the models. The algorithm has two stages. In the first, the neural network is used to recursively filter (estimate) the nonlinearity from the noisy measurements. The network serves as a teacher/trainer for the model candidates, and the appropriate model is selected in a simple tournament-like routine. The main advantage of the algorithm over a traditional one stage approach (in which models are determined directly from measurements), is its small computational overhead (as computational complexity and memory occupation are both greatly reduced).
Keywords: system identification, structure detection, Hammerstein system, wavelet neural network

1 Introduction

1.1 Types of knowledge. Classification of approaches

In the paper we propose the cooperation between parametric and nonparametric methods for system modeling. The term 'parametric' means that estimated nonlinearity can be described with the use of finite number of parameters. The nonparametric approach is applied for smart selection of the best parametric model of nonlinear characteristic from a given finite class of models. In section 2 the problem is formulated in detail and the relation between the regression function and the static characteristic of Hammerstein system is discussed. In section 4 we present the nonparametric method for regression estimation, and next in section 3 the parametric nonlinear least-squares for the regression approximation in Hammerstein system is introduced. The above two different approaches are combined in section 5. First, nonparametric estimates are continuously (recursively) computed from the learning pairs on the grid of N_0 input points and

support selection of one from competing models. The parameters of the best model in the selected class are then obtained by the nonlinear least squares, and broad variety of optimization algorithms (also soft methods, e.g. genetic, tabu search, simulated annealing, particle swarming) can be applied in this stage, depending on the specifics of the optimization criterion. If the resulting parametric approximation is not satisfying, and the number of measurements is large enough, the residuum between the system output and the model output is used for its nonparametric refinement.

1.2 Contribution

The contribution of the paper is the following.

- the idea of the *regression-based* approach to *parametric* identification of nonlinear characteristic in Hammerstein system is introduced, and models which are *not linear in the parameters* are admitted in general,
- the proposed identification methods work under *uncertain* knowledge about the parametric representation of the static nonlinear characteristic,
- the method of fast *recursive* nonparametric *recognition/selection* of the true formula from a given set of parametric models is proposed.

2 Statement of the problem

2.1 Class of systems

The Hammerstein system is built of a static non-linearity, $\mu()$, and a linear dynamics, with the impulse response $\{\gamma_i\}_{i=0}^{\infty}$, connected in a cascade and described by the following set of equations: $y_k = v_k + z_k$, $v_k = \sum_{i=0}^{\infty} \gamma_i w_{k-i}$, $w_k = \mu(u_k)$, or equivalently

$$y_k = \sum_{i=0}^{\infty} \gamma_i \mu(u_{k-i}) + z_k, \tag{1}$$

where u_k and y_k denote the system input and output at time k, respectively, and z_k is the output noise (see Fig.1)

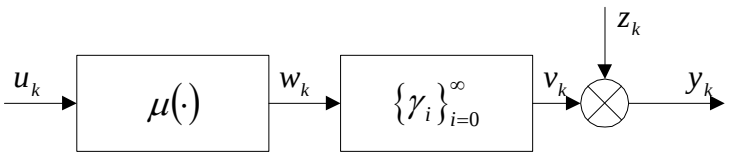

Fig. 1. The identified Hammerstein system

2.2 Assumptions / a priori knowledge

1. The input signal $\{u_k\}$ is for $k = ..., -1, 0, 1, ...$ an i.i.d. bounded random process $|u_k| \leqslant u_{\max}$, some $u_{\max} > 0$, and there exists a probability density of u_k, say $\nu(u)$.
2. The nonlinear characteristic $\mu(u)$ is a bounded function on the interval $[-u_{\max}, u_{\max}]$, i.e.
$$|\mu(u)| \leqslant w_{\max} \tag{2}$$
where $w_{\max}$ is some positive constant.
3. The linear dynamics is an asymptotically stable *IIR* filter
$$v_k = \sum_{i=0}^{\infty} \gamma_i w_{k-i} \tag{3}$$
with the unknown impulse response $\{\gamma_i\}_{i=0}^{\infty}$ (such that $\sum_{i=0}^{\infty} |\gamma_i| < \infty$).
4. The output noise $\{z_k\}$ is a random, arbitrarily correlated process, governed by the general equation
$$z_k = \sum_{i=0}^{\infty} \omega_i \varepsilon_{k-i} \tag{4}$$
where $\{\varepsilon_k\}$, $k = ..., -1, 0, 1, ...$, is a bounded stationary zero-mean white noise ($E\varepsilon_k = 0$, $|\varepsilon_k| \leqslant \varepsilon_{\max}$), independent of the input signal $\{u_k\}$, and $\{\omega_i\}_{i=0}^{\infty}$ is unknown; $\sum_{i=0}^{\infty} |\omega_i| < \infty$. Hence the noise $\{z_k\}$ is a stationary zero-mean and bounded process $|z_k| \leqslant z_{\max}$, where $z_{\max} = \varepsilon_{\max} \sum_{i=0}^{\infty} |\omega_i|$.
5. $\mu(u_0)$ is known at some point u_0 and $\gamma_0 = 1$.

As it was explained in detail in [1] and [2], the input-output pair $(u_0, \mu(u_0))$ assumed to be known can refer to arbitrary $u_0 \in [-u_{\max}, u_{\max}]$, and hence we shall further assume for convenience that $u_0 = 0$ and $\mu(0) = 0$, without loss of generality.

2.3 Preliminaries

A fundamental meaning for the methods elaborated in this paper has the dependence between the regression and the nonlinear characteristic

$$R(u) = E\{y_k | u_k = u\} = \gamma_0 \mu(u) + d \text{ where } d = E\mu(u_k) \cdot \sum_{i>0} \gamma_i = \text{const}\,.$$

Since under assumption $\mu(0) = 0$, it holds that $R(0) = d$ and

$$R(u) - R(0) = \gamma_0 \mu(u) \tag{5}$$

The equivalence (5) allows to recover the nonlinear characteristic $\mu()$ under lack of prior knowledge about the linear dynamic subsystem. This observation was successfully utilized in eighties by Greblicki and Pawlak in nonparametric

methods, when the parametric form of $\mu()$ is also unknown, but was never explored in parametric identification framework, when some prior knowledge of only $\mu()$ is given.

In this paper we show that the regression-based approach to nonlinearity recovering in Hammerstein system can also be applied when some, even uncertain, parametric prior knowledge of the static characteristic is given.

3 Parametric approximation of the regression function

The methods presented in section 4 recover the true regression function from the input-output measurements. Here we accept the parametric class of model and try to find the best approximation in this class.

3.1 Regression-based parametric approach

In the traditional (purely parametric) approach we suspect that the nonlinear characteristic $\mu(u)$ can be well approximated by the model from the given class

$$\overline{\mu}(u, c), \tag{6}$$

where $c = (c_1, c_2, ..., c_m)^T$ includes finite number of parameters. The function $\overline{\mu}(u, c)$ is assumed to be *differentiable* with respect to c. Let $c^* = (c_1^*, c_2^*, ..., c_m^*)^T$ be the best choice of c in the sense that

$$c^* = \arg\min_c E\left(\mu(u) - \overline{\mu}(u, c)\right)^2. \tag{7}$$

Further, we will explore the generalized version of (6)

$$\overline{\mu}(u, \vartheta) = c_\alpha \overline{\mu}(u, c) + c_\beta,$$

where $\vartheta = \left(c^T, c_\alpha, c_\beta\right)^T$ is the extended model vector enriched with the scale c_α and the offset c_β. Obviously for $c_\alpha = 1$ and $c_\beta = 0$ equation (7) can be rewritten in the form

$$\arg\min_\vartheta E\left(\mu(u) - \overline{\mu}(u, \vartheta)\right)^2 = \left(c^{*T}, 1, 0\right)^T \triangleq \vartheta^*. \tag{8}$$

Remark 1. For functions which are linear in the parameters and has additive constant the classes $\overline{\mu}(u, c)$ and $\overline{\mu}(u, \vartheta)$ are indistinguishable. For example the polynomial model of order m (see [3])

$$\overline{\mu}(u, c) = c_m u^{m-1} + ... + c_2 u + c_1$$

leads to the same class of

$$\overline{\mu}(u, \vartheta) = c_\alpha c_m u^{m-1} + ... + c_\alpha c_2 u + c_\alpha c_1 + c_\beta.$$

Remark 2. If $E\left(\mu(u) - \overline{\mu}(u, \vartheta)\right)^2$ is minimized by $\left(c^{*T}, 1, 0\right)^T$, then $E\left(R(u) - \overline{R}(u, \theta)\right)^2$ is minimized by $\left(c^{*T}, \gamma_0, d\right)^T$.

3.2 Approximation

By rewriting (1) in the form $y_k = \gamma_0\mu(u_k) + \sum_{i=1}^{\infty}\gamma_i\mu(u_{k-i}) + z_k$, and taking into account that $R(u_k) = \gamma_0\mu(u_k) + \sum_{i=1}^{\infty}\gamma_i E\mu(u_1)$, we obtain the equivalent (cardinal) description $y_k = R(u_k) + \delta_k$ of the Hammerstein system, in which the total noise

$$\delta_k \triangleq y_k - R(u_k) = \sum_{i=1}^{\infty}\gamma_i\left(\mu(u_{k-i}) - E\mu(u_1)\right) + z_k$$

is zero-mean ($E\delta_k = 0$) and independent of u_k. We want to find the vector θ^* for which the model $\overline{R}(u_k,\theta)$ fits to data the best, in the sense of the following criterion

$$E\left(y_k - \overline{R}(u_k,\theta)\right)^2 = \operatorname{var}\delta_k + E\left(R(u_k) - \overline{R}(u_k,\theta)\right)^2. \tag{9}$$

From (9) we conclude that

$$\arg\min_{\theta} E\left(y_k - \overline{R}(u_k,\theta)\right)^2 = \arg\min_{\theta} E\left(R(u_k) - \overline{R}(u_k,\theta)\right)^2,$$

which is fundamental for the least squares approximation

$$\widehat{\theta}_N = \arg\min_{\theta}\sum_{k=1}^{N}\left(y_k - \overline{R}(u_k,\theta)\right)^2.$$

4 Nonparametric neural network trainer/teacher

To evaluate $\widehat{R}_N(\overline{u})$ we use a neural network, denoted further by $R_k(u)$, and based on either radial basis [4–6], [7, Ch. 17] and [8, 9], or wavelet [10–12, 7, Ch. 18], or classic kernel [13, 7, Ch. 5] regression function estimates.

For each pair of the *learning sequence,* (u_k, y_k), $k = 1, 2, \ldots$, and for each point u from some *training set* $\{u_e\}_{e=1}^{N_0}$, the network learning formula is given recursively as:

$$\hat{R}_k(u) = \hat{R}_{k-1}(u) + \overbrace{\underbrace{\kappa_k(u)}_{\text{weight}}\cdot\underbrace{\left[y_k - \hat{R}_k(u)\right]}_{\text{error}}}^{\text{correction}},\ \text{for all } u = u_e, \tag{10}$$

where (we take $0/0 = 0$ when necessary)

$$\kappa_k(u) = \frac{\phi_k(u)}{\hat{f}_k(u)} \text{ with } \hat{f}_k(u) = \hat{f}_{k-1}(u) + \phi_k(u)$$

and where $\phi_k(u)$ is a shorthand of the selected kernel function $\phi_{K(k)}(u, u_k)$, and where, finally, $\hat{f}_k(u)$ is the recursive estimate of the density of the inputs in the learning sequence. The initial conditions are $\hat{R}_0(u) = \hat{f}_0(u) = 0$). The following lemma characterizes the limit properties of the proposed neural network; *cf.* [14].

Lemma 1. *Let the nonlinearity $R(u)$ and the input signal density function have $\lfloor \nu_R \rfloor$ and $\lfloor \nu_f \rfloor$ derivatives, respectively (for some $\nu_R, \nu_f > 0$). If the kernel function $\phi_{K(k)}(u, v)$ has p vanishing moments and its bandwidth parameter is governed by the rule $K(k) = (2\nu + 1)^{-1} \log_2 k$, where $\nu = \min\{\nu_R, \nu_f, p\}$, then, in all* training points $u \in \{u_e\}_{e=1}^{N_0}$ *the network error vanishes and*

$$\left| \hat{R}_k(u) - R(u) \right| = \mathcal{O}\left(k^{-\nu/(2\nu+1)} \right), \text{ in probability.} \tag{11}$$

Proof. See [15] for the proof in case of wavelet network and [14] for the proof for other networks.

5 Model training and competition

Let's split the set of *training set* $\{u_e\}$ into two disjoint parts containing $\{u_l\}$ and $\{u_t\}$, being the *learning* and *testing* points, respectively. After each new measurement arrival and application of the recursive update procedure (10)) the training/testing routine is performed on the models $\left\{ \overline{R}^{(l)}(u, \theta_l) \right\}$.

In this phase, the models are trained, *i.e.* their parameters $\{\theta_M\}$ are evaluated using learning part, $\{u_l\}$, of the training points set $\{u_e\}$.

Remark 3. The evaluation routine is particularly simple when the models are linear in parameters and the functions the model is built upon are pairwise orthogonal. Then it actually reduces to solving the linear equation system.

Each model $\overline{R}^{(l)}(u, \theta_l)$ collects its own number of wins W_l. The following two strategies can be now used to model the nonlinearity $\mu(u)$:

1. the *winner-take-all* approach, in which the model with the largest number of wins W_l is selected as the model of the nonlinearity (the "winner-takes-all" approach), *i.e.*

$$R(u, \theta) = \overline{R}^{(l_{\max})}(u, \theta_{l_{\max}}), \text{ such that } l_{\max} = \max_l \{W_l\}$$

   or
2. the *soft* approach, in which a *convex combination* of the models is used as the nonlinearity model, *i.e.*

$$R(u, \theta) = \sum_{l=1}^{M} w_l \cdot \overline{R}^{(l)}(u, \theta_l)$$

   where $w_l = W_l / k$. Clearly, $\sum_{l=1}^{M} w_l = 1$.

The following theorem describes the limit properties of the proposed model selection algorithm:

Theorem 1. *If the neural network trainer $\hat{R}_k(u)$ converges to the nonlinearity $R(u)$ in all points of the training set, $\{x_e\}$, then the proposed smart model selection algorithm picks (in probability) eventually the best model of the nonlinearity.*

Proof. The proof is immediate. The neural network trainer $\hat{R}_k(u)$ approaches the actual nonlinearity $R(u)$ by virtue of the Lemma 1. Since each model $\overline{R}^{(l)}(u, \theta_l)$ is evaluated to minimize the distance (the error) between models and the estimate, then the convergence rate is determined by the estimate rate, otherwise, the best model converges to the best approximation of the nonlinearity with the same rate.

The theorem says that in both *winner-takes-all* and *soft* strategy, the best nonlinearity model is chosen. Indeed, one can expect that with the growing size of the processed learning sequence pairs (u_k, y_k) the neural network trainer becomes the better-and-better estimate of the unknown nonlinearity, and hence the model closest to the estimate is simultaneously the closest to the nonlinearity, and eventually it 'overwhelms' its rivals. With the number of learning pairs growing to infinity, the weight, w_l, of this model tends to 1 (and the weights of other vanishes).

Remark 4. Clearly, the training set needs to be split into learning and testing parts. If, for instance, we had used use the same set for learning and testing we would have obtained the zero error (the best match) for all models. Consider for example three models utilizing the first $N_0 = 2^\eta$, $\eta = 1, 2, \ldots$, terms of the Haar wavelet, Fourier trigonometric or Legendre polynomial orthogonal series. Assume now that the $\{x_l\}_{l=1}^{N_0}$ are equidistant, *i.e.* they form a binary grid. In such a setting, all these series are orthogonal bases on such a discrete grid and hence are able to represent exactly (recover) any function defined in training points $\{x_l\}$.

6 Simulation example

The experiment illustrated the algorithm performance for small samples sizes. As the Hammerstein system nonlinearity $\mu(u)$, a polynomial of order eight was selected. The dynamic system was a discrete damped oscillator with the impulse response $\lambda_i = (-1/2)^{-i}$. Both the system input and the external noise was uniform. The input signal range was $[0, 1]$ and the noise amplitude was set so that $SNR = \max_k |z_k| / \max_{u\in[0,1]} |\mu(u)| = 10\%$.

The network trainer $\hat{\mu}_k(u)$ was based on Daubechies wavelets with $p = 5$ vanishing moments and used the practical bandwidth selection rule $K(k) = 1/3 \log_2 k$; *cf.* 1. Three classes – based on either the Haar, Fourier and Legendre functions (with eight parameters each) were employed. During the 256 tournaments (performed after each arrival of packages of 32 learning pairs (u_k, y_k)), the models were trained in $N_0 = 64$ equidistant points. The competition phase took place in another $N_t = 64$ points.

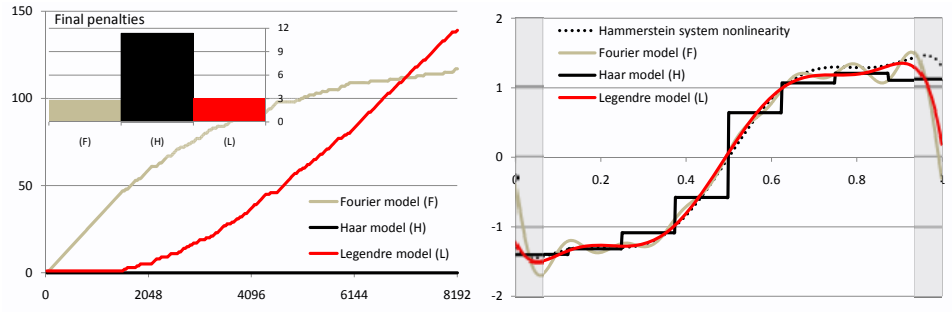

Fig. 2. The results of the competition between Fourier, Haar and Legendre models. Note that the winner (the Legendre model) had the higher final penalty than the Fourier one (left). The nonlinearity diagram against the models (right). The gray strips show the 'unfair' boundaries excluded from the competitions to avoid the boundary problem influence

The experiments confirm in general our theoretical findings showing also some (natural) disparities between the limit properties and the small sample size ones. In particular, they reveal that:

- The neural network estimate should have the shape resembling neither of the model shapes. This is because, for the small learning data sets, the nonlinearity approximation abilities of the estimate are quite poor (as only a few expansion terms are active). Hence, initially, the estimate exhibits the shapes of its basis functions rather than the actual shape of the nonlinearity, *viz.*, the models "learn" the estimate artifacts in lieu of the nonlinearity shape.
- The above argument suggest also that, especially when k is small or moderate, one should prefer the approximation-based (averaging) schemes rather than interpolation-based ones to train the models.
- If the model classes are similar (like, in our example, the Fourier and Legendre ones), then, initially, the discrimination can be difficult as well.

Remark 5. Instead of collecting the models winnings, one can sum up their training errors (penalties). However, when k is small, the models errors are typically variance-induced and greater by orders of magnitude than their small and mainly of an approximation nature errors occurring for large k, and this model selection algorithm can longer than the proposed one be misled by the erroneous initial results.

7 Final remarks

In the paper we proposed the new algorithms which combines the *parametric* and *nonparametric* approaches to recover the Hammerstein system nonlinearity under *quasi-parametric* prior knowledge. The nonparametric neural network

trainer is used first to filter (smooth) the noisy learning sequence, and then to train the model candidates. In the competition phase, the winner is the model which matches best the neural network trainer.

It is shown that in the proposed *winner-take-all* approach, the model which collects the largest number of wins, is selected as the nonlinearity model. Otherwise, in the alternative *soft* strategy, the convex combination of all competitor models is taken as the model. Note, however, that asymptotically both strategies lead to the selection of the single model.

One can point out the following advantages of the algorithm:

1. No need of storing the measurements in memory.
2. Fast model training – the random learning points are replaced by the deterministic ones, *i.e.*, the active experiment techniques (*e.g.* orthogonal plans) can be used in model training in place of the passive ones.
3. Flexible model selection routine – the examine routines, *i.e.* the model competitions can be performed in the user-defined regions of interests, *e.g.* in the working points.
4. The list of model candidates can be open – the new models can join the competition at any time (with a "*wild card*"), and win, if they are actually the proper ones – since all the information about the nonlinearity used to train the models is maintained by the neural network trainer.

Clearly, there are also some weaknesses:

1. The main disadvantage of the proposal consists in its slower convergence rate. It is of nonparametric order, $\mathcal{O}(k^{-1/2+\gamma})$, where $\gamma = 1/2\,(2\nu+1)$, and in fact, is a toll we pay for a smaller prior knowledge. Recall however (*cf.* (11) in the Lemma 1) that, in general, ν grows with the smoothness of the nonlinearity. That is, the smoother the nonlinearity, the smaller γ, and the convergence rate is closer to the typical parametric rate $\mathcal{O}(k^{-1/2})$.
2. The set of model class candidates has to be complete in the sense that the nonlinearity has to belong to the one of them. Otherwise, the nonlinearity can be of any shape and neither model considered in the section 3.2 can be its reasonable approximation.

If the a priori knowledge is nonparametric we cannot guarantee that any model class matches the target nonlinearity. Then one should use a *semiparametric approach*, in which the true nonlinearity can be recovered even if the prior model is incorrect; see [3, 16]. This technique allows avoiding the most pessimistic scenario, when *e.g.* the nonlinearity is orthogonal to each model classes, and its representations (even the best approximation; *cf.* Section 3.2) in theses classes are merely worthless.

Remark 6. The algorithm can also be seen as a pattern recognition algorithm classifying the system nonlinearity to the one of the predefined model classes. The wavelet neural network trainer plays there a role of the raw learning data preprocessor and the tournament routine is an implementation of a nearest-neighbor algorithm; *cf. [17].*

References

1. Hasiewicz, Z., Mzyk, G.: Combined parametric-nonparametric identification of Hammerstein systems. IEEE Transactions on Automatic Control **49**(8) (2004) 1370–1375
2. Hasiewicz, Z., Mzyk, G.: Hammerstein system identification by nonparametric instrumental variables. International Journal of Control **82**(3) (2009) 440–455
3. Śliwiński, P., Rozenblit, J., Marcellin, M.W., Klempous, R.: Wavelet amendment of polynomial models in nonlinear system identification. IEEE Transactions on Automatic Control **54**(4) (2009) 820–825
4. Krzyzak, A., Linder, T., Lugosi, C.: Nonparametric estimation and classification using radial basis function nets and empirical risk minimization. IEEE Transactions on Neural Networks **7**(2) (1996) 475–487
5. Kamiński, W., Strumiłło, P.: Kernel orthonormalization in radial basis function neural networks. IEEE Transactions on Neural Networks **8**(5) (1997) 1177–1183
6. Krzyzak, A., Linder, T.: Radial basis function networks and complexity regularization in function learning. IEEE Transactions on Neural Networks **9**(2) (1998) 247–256
7. Györfi, L., Kohler, M., A. Krzyżak, Walk, H.: A Distribution-Free Theory of Nonparametric Regression. Springer-Verlag, New York (2002)
8. Buhmann, M.D.: Radial Basis Functions: Theory and Implementations. Cambridge University Press, Cambridge (2003)
9. Ferrari, S., Maggioni, M., Borghese, N.A.: Multiscale approximation with hierarchical radial basis functions networks. IEEE Transaction one Neural Networks **15**(1) (2004)
10. Zhang, Q., Benveniste, A.: Wavelet networks. IEEE Transactions on Neural Networks **3** (1992) 889–898
11. Zhang, J., Walter, G., Miao, Y., Lee, W.N.: Wavelet neural networks for function learning. IEEE Transactions on Signal Processing **43** (1995) 1485–1497
12. Hasiewicz, Z.: Modular neural networks for non-linearity recovering by the Haar approximation. Neural Networks **13** (2000) 1107–1133
13. Greblicki, W., Pawlak, M.: Identification of discrete Hammerstein system using kernel regression estimates. IEEE Transactions on Automatic Control **31** (1986) 74–77
14. Rutkowski, L.: Generalized regression neural networks in time-varying environment. IEEE Transactions on Neural Networks **15**(3) (2004) 576 – 596
15. Śliwiński, P., Hasiewicz, Z.: Recursive wavelet estimation of Hammerstein systems nonlinearity. International Journal of Control (2010) In preaparation.
16. Greblicki, W., Mzyk, G.: Semiparametric approach to Hammerstein system identification. In: Proceedings of the 15th IFAC Symposium on System Identification, Saint-Malo, France (2009) 1680–1685
17. Rutkowski, L.: Adaptive probabilistic neural networks for pattern classification in time-varying environment. IEEE Transactions on Neural Networks **15**(4) (2004) 811 – 827