JAISCR, 2021, Vol. 11, No. 3, pp. 181– 194
10.2478/jaiscr-2021-0011

BANDWIDTH SELECTION FOR KERNEL GENERALIZED REGRESSION NEURAL NETWORKS IN IDENTIFICATION OF HAMMERSTEIN SYSTEMS

Jiaqing Lv[1], Mirosław Pawlak[1,2,*]

[1]*Department of Electrical & Computer Engineering*
University of Manitoba, Canada

[2]*Information Technology Institute,*
University of Social Sciences, Lodz, Poland

[*]*E-mail: miroslaw.pawlak@umanitoba.ca*

Submitted: 5th August 2020; Accepted: 19th January 2021

Abstract

This paper addresses the issue of data-driven smoothing parameter (bandwidth) selection in the context of nonparametric system identification of dynamic systems. In particular, we examine the identification problem of the block-oriented Hammerstein cascade system. A class of kernel-type Generalized Regression Neural Networks (GRNN) is employed as the identification algorithm. The statistical accuracy of the kernel GRNN estimate is critically influenced by the choice of the bandwidth. Given the need of data-driven bandwidth specification we propose several automatic selection methods that are compared by means of simulation studies. Our experiments reveal that the method referred to as the partitioned cross-validation algorithm can be recommended as the practical procedure for the bandwidth choice for the kernel GRNN estimate in terms of its statistical accuracy and implementation aspects.

Keywords: Generalized regression neural networks, nonparametric estimation, bandwidth, data-driven selection, nonlinear systems, Hammerstein systems.

1 Introduction

The goal of system identification is to build mathematical models of dynamic systems from observed input-output data. This fundamental modeling problem has found applications in various fields of science and engineering, e.g., communication, signal processing, control systems, power engineering, biomedical engineering, chemical processes, and financial modeling [1, 2, 3, 4, 5]. For the comprehensive overview of the field we refer to [6, 7, 8, 9, 10, 11].

The accuracy of an identification algorithm critically depends on the assumed class of models and the size and quality of the observed data. There are two distinct strategies to specify a class of models, i.e., the parametric specification and the nonparametric one. In the latter case, one makes no functional assumptions on the system characteristics and as a result the identification procedure must be conducted in the infinite dimensional space [6, 12, 13]. In contrast, in the parametric approach the identification procedure is performed in the finite dimensional space spanned by a vector of unknown parameters [11]. These two strategies are particularly manifested in the field of nonlinear dynamic system identification [11], where one encounters a wide class of nonlinearities and memory structures. The choice between parametric and nonparametric approaches to system identification depends on the data size, dimensionality and memory length.

A parsimonious strategy for nonlinear system identification is based on the concept of semiparametric block-oriented models which are characterized by the separation between static nonparametric nonlinearities and linear parametric dynamical systems [8]. These models often reflect the physical nature of the examined system, where one encounters nonlinear sensors and actuators. There is a large number of combinations of nonlinear/linear elements that define block-oriented dynamic models. The series/parallel structures are the most popular choices with numerous applications in control, signal processing and biological systems [14, 15, 16, 17, 18, 19, 20]. This includes popular cascade models as Hammerstein, Wiener, sandwich, and their parallel counterparts.

In this paper we examine the commonly used Hammerstein block-oriented system being a series connection of a nonlinear static characteristic followed by a linear dynamical system. This configuration defines the basic building block for other more involved block-oriented models. The structure of the Hammerstein system is depicted in Figure 1.

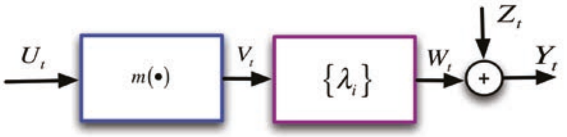

Figure 1. Hammerstein system with input U_t and output Y_t

The system consists of the cascade of a memoryless nonlinearity $m(\cdot)$ and a linear time invariant subsystem with the impulse response $\{\lambda_i\}$. The signal $V_t = m(U_t)$ defines the input signal to the linear subsystem, where $\{U_t\}$ is the input signal assumed to be the *iid* process with the density $f(\cdot)$. Hence, we have

$$W_t = \sum_{i=0}^{\infty} \lambda_i V_{t-i}, \tag{1}$$

where we assume that the liner sub-system is stable, i.e., $\sum_{i=0}^{\infty} |\lambda_i| < \infty$.
The output W_t of the linear part is disturbed by the additive white noise Z_t, i.e., we measure the output signal $Y_t = W_t + Z_t$.

The identification problem for the Hammerstein system is to recover both nonlinear and linear parts of the system from the measured input-output training data

$$\mathcal{T}_n = \{(U_1,Y_1),\ldots,(U_n,Y_n)\}. \tag{2}$$

It is worth noting that the signals V_t and W_t are unobservable. The critical part of the Hammerstein system is its nonlinearity $m(\cdot)$ which, in this paper, is assumed to be unknown and of the nonparametric form. This calls for nonparametric identification algorithms that can consistently recover $m(\cdot)$.

It is worth mentioning that the cascade structure of the Hammerstein system implies that $m(\cdot)$ can be estimated up to some scaling and additive parameters, i.e., one can only recover $\mu(u) = am(u) + b$ for some unknown constants a, b. This is easily seen by writing the input-output relationship as follows

$$Y_t = \lambda_0 m(U_t) + \sum_{i=1}^{\infty} \lambda_i m(U_{t-i}) + Z_t. \tag{3}$$

Then, owing to the *iid* nature of the input signal we obtain

$$\mathbb{E}[Y_t|U_t = u] = am(u) + b, \tag{4}$$

where $a = \lambda_0$ and $b = \mathbb{E}[m(U)]\sum_{i=1}^{\infty} \lambda_i$. Therefore, if the following assumption

$$\mathbb{E}[m(U)] = 0 \quad \text{and} \quad \lambda_0 = 1 \tag{5}$$

holds then $a = 1$ and $b = 0$.
In this paper we are interested in the problem of the bandwidth selection for a class of Generalized Regression Neural Networks estimates of $m(\cdot)$ and the multiplicative/additive scaling has no influence on the accuracy of the examined methods. Hence, without loss of generality, the assumption in (5) will be assumed to hold. Consequently, we have $\mathbb{E}[Y_t|U_t = u] = m(u)$, i.e., the system nonlinearity $m(u)$ can be uniquely recovered from the regression function $\mathbb{E}[Y_t|U_t = u]$ of the output signal on the input one. This takes place for any linear subsystem characterized by the impulse response $\{\lambda_i\}$. The issue of estimating the nonparametric nonlinearity $m(\cdot)$ is discussed in the next section.

2 The Kernel GRNN Estimate

Without imposing any a priori information on the shape of $m(\cdot)$ one can apply the existing nonparametric regression estimates in order to recover the system nonlinearity $m(\cdot)$. Various nonparametric estimation techniques can be applied in this context including kernel, orthogonal series and nearest neighbor estimates [6, 21, 12].

In this paper we utilize a class of kernel Generalized Regression Neural Networks (GRNN) originally introduced in [22], see also [23] for some more recent results. The kernel GRNN estimate is a single-layer feed-forward neural network that uses the normalized kernels in the hidden layer as activation functions. Moreover, each activation function is scaled by the smoothing parameter (bandwidth) that controls the network input-output mapping complexity. In fact, the large value of bandwidth leads to a linear model, whereas the small bandwidth results in the highly complex input-output mapping. The kernel GRNN structure does not require any back-propagation learning algorithm as it is given in the following explicit form

$$\widehat{m}_h(u) = \frac{\sum_{i=1}^{n} K_h(u-U_i)Y_i}{\sum_{j=1}^{n} K_h(u-U_j)}, \tag{6}$$

where $K_h(\cdot) = h^{-1}K(\cdot/h)$ is the scaled version of the kernel function $K(\cdot)$. The kernel function $K(\cdot)$ can be selected as any symmetric density function. The parameter h (bandwidth) must be properly selected as it plays the critical role in the statistical accuracy of the estimate $\widehat{m}_h(u)$. In fact, if h is too small then a few data points will be effectively included into the local averaging and the estimate $\widehat{m}_h(u)$ would have too many spikes being the illustration of the large estimate variance. On the other hand, if h is too large then too many local data points would be included into the local averaging and the estimate $\widehat{m}_h(u)$ would be too smooth reflecting the large estimate bias. Both of these two scenarios would lead to imprecise models with high prediction and estimation errors. Therefore, the proper selection of h is crucial for the practical use of the kernel GRNN estimate. To quantitatively explain this variance/bias dilemma let us consider the mean integrated squared error

$$MISE(h) = \mathbb{E}\int_{S}[\widehat{m}_h(u) - m(u)]^2 w(u)du, \tag{7}$$

where $w(u)$ is the weight function and $\mathcal{S}$ is the support of the input density $f(u)$. Some standard algebra, see [6], yields the following asymptotic decomposition of $MISE(h)$

$$\frac{\sigma^2}{nh}k_1\int_{S}\frac{w(u)}{f(u)}du + \frac{h^4}{4}k_2^2\int_{S}\varphi^2(u)w(u)du, \tag{8}$$

where $\sigma^2 = \mathbb{E}[m^2(U)]\sum_{i=1}^{\infty}\lambda_i^2 + \sigma_Z^2$. Here σ_Z^2 is the variance of the additive external noise Z_t. Also $k_1 = \int K^2(u)du$, $k_2 = \int u^2K(u)du$ and

$$\varphi(u) = \frac{m^{(2)}(u)f(u) + 2m^{(1)}(u)f^{(1)}(u)}{f(u)}.$$

In the formula in (8) the first terms represents the asymptotic variance, whereas the second one is the asymptotic bias. It is clear that the variance term is the decreasing function of h, while the bias is a increasing function of h confirming the aforementioned discussion on the variance/bias tradeoff. The direct minimization of (8) shows that the asymptotic optimal value of h is

$$h_{opt} = Cn^{-1/5}, \tag{9}$$

where C is the unknown constant depending on the Hammerstein system characteristics, i.e., $m(\cdot)$, $\{\lambda_i\}$, σ_Z^2 and the input density $f(\cdot)$. Hence, h_{opt} cannot be used in practical applications.
The data-driven choices of h rely on various resampling techniques that are used for estimating the prediction error. Note that if $w(u) = f(u)$ in (7) then we can re-write $MISE(h)$ as follows

$$MISE(h) = \mathbb{E}[\widehat{m}_h(U_t) - m(U_t)]^2, \tag{10}$$

for some U_t being independent of the training set $\mathcal{T}_n$ in (2). The ideal value of h would be the one that minimizes $MISE(h)$ in (10). By virtue of (3) and (5) we can write

$$Y_t = m(U_t) + \varepsilon_t, \tag{11}$$

where $\varepsilon_t = \sum_{i=1}^{\infty}\lambda_i m(U_{t-i}) + Z_t$ represents the overall noise added to the system nonlinearity. This noise has a complex correlation structure as it depends on the linear and nonlinear parts of the system. Since ε_t is independent on U_t, we observe that the minimization of $MISE(h)$ in (10) is equivalent to the minimization of the following prediction error

$$CMISE(h) = \mathbb{E}[\widehat{m}_h(U_t) - Y_t]^2, \tag{12}$$

where the input-output pair (U_t, Y_t) is independent of the training set $\mathcal{T}_n$ in (2). Clearly, the minimization of $CMISE(h)$ is impossible since we do not know the distribution of the input-output data. The naive estimate of $CMISE(h)$ based on $\mathcal{T}_n$ would be

$$RE(h) = \frac{1}{n}\sum_{i=1}^{n}[Y_i - \widehat{m}_h(U_i)]^2. \qquad (13)$$

This is the residual square error that, once it is minimized, produces the bandwidth value yielding the kernel GRNN estimate with very large variance.

The problem of the data-driven bandwidth selection for standard kernel estimates has been extensively examined in the statistical literature [24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34]. In these contributions the classical regression model

$$Y_t = m(U_t) + \varepsilon_t \qquad (14)$$

with the *iid* noise $\{\varepsilon_t\}$ has been mostly studied. Various cross-validation (CV) methods have been established with some proved favourable statistical properties.

In the case when the noise process $\{\varepsilon_t\}$ is correlated the proposed methods fail to estimate the proper bandwidth value. In fact, the obtained bandwidth is either too large or too small depending on the sign of the correlation function of $\{\varepsilon_t\}$. To fix the aforementioned deficiency several modified CV algorithms have been proposed that accommodate the correlation nature of the noise process [28, 29, 30, 31, 32, 33, 34]. However, in most of these contributions, the input signal is assumed to be non-random (fix design regression) and the noise process $\{\varepsilon_t\}$ meets some mixing properties. These two assumptions do not hold in the context of the Hammerstein system. In fact, due to (11) the Hammerstein system is of the form as in (14). However, the error term $\varepsilon_t = \sum_{i=1}^{\infty}\lambda_i m(U_{t-i}) + Z_t$ is correlated and depends on both the linear and nonlinear parts of the system as well as on the external additive noise Z_t. Moreover, it can be shown, see [13], that the noise process $\{\varepsilon_t\}$ appearing in the Hammerstein system is not necessarily mixing. Thus, we are faced with much more complicated data structure compared to the aforementioned cases studied in the statistical literature [28, 29, 30, 31, 32, 33, 34].

In this paper, we examine the three carefully selected cross-validation methods for the data-driven bandwidth selection. The accuracy of these methods will be evaluated by mens of simulation studies for various combinations of nonlinear and linear components of the Hammerstein system. The theoretical properties of the methods will be examined elsewhere.

The rest of the paper is organized as follows. In Section 3 we introduce the data-driven algorithms for the bandwidth selection. Section 4 summarizes our simulation studies for a large combination of linear and nonlinear characteristics of the Hammerstein system. In Section 5 we summarize our results, whereas in Section 6 further extensions are discussed.

3 Data-Driven Bandwidth Selectors

A natural modification of the naive estimate in (13) is the classical leave-one-out selector, i.e., a data-driven h is found by minimizing the following criterion

$$CV(h) = n^{-1}\sum_{i=1}^{n}[Y_i - \widehat{m}_{h,-i}(U_i)]^2, \qquad (15)$$

where

$$\widehat{m}_{h,-i}(u) = \frac{\sum_{j=1, j\neq i}^{n} Y_j K_h(u - U_j)}{\sum_{j=1, j\neq i}^{n} K_h(u - U_j)}. \qquad (16)$$

Note that $\widehat{m}_{h,-i}(u)$ is the leave-one-out version of the estimator $\widehat{m}_h(u)$ in (6) where the i-th observation pair (U_i, Y_i) is omitted. The criterion $CV(h)$ is based on the partition principle that data used for forming the estimate $\widehat{m}_{h,-i}(u)$ and those used for averaging should be separated as much as possible. The leave-one-out criterion is universal but is not taking into account the fact that we are dealing with dependent data. We will use $CV(h)$ as the reference selector with which we will compare other data-driven methods studied in this paper. This includes, modified, partitioned and corrected cross-validation procedures.

3.1 Modified Cross-Validation

The selector relying on $CV(h)$ is efficient if Y_i and $\widehat{m}_{h,-i}(U_i)$ are statistically independent. This takes place if the noise process $\{\varepsilon_i\}$ is white. This clearly is not the case in the context of the Hammerstein system.

In order to include the dependence structure of the observed data let us define the so-called the Modified Cross-Validation (MCV) criterion. Hence, let us define the following selection criterion

$$MCV(h) = n^{-1}\sum_{i=1}^{n}[Y_i - \widehat{m}_{MCV}(U_i;l)]^2, \qquad (17)$$

where

$$\widehat{m}_{MCV}(U_i;l) = \frac{\sum_{j=1:|j-i|>l}^{n} Y_j K_h(U_i - U_j)}{\sum_{j=1:|j-i|>l}^{n} K_h(U_i - U_j)}. \qquad (18)$$

Note that in the formation of $\widehat{m}_{MCV}(U_i;l)$, only data points that are the l units apart from the observation pair (U_i,Y_i) are taken into account. Hence, $\widehat{m}_{MCV}(u;l)$ is the leave-($2l$+1)-out version of the kernel estimator. The value of the parameter l is directly related to the memory size of the linear dynamical subsystem. For instance, if it is known that the linear part in (1) is of order p, i.e.,

$$W_t = \sum_{i=0}^{p}\lambda_i V_{t-i} \qquad (19)$$

then we should choose $l = p+1$. In other more general cases one should try a several integer values of l. The range of l is not large since for a wide class of stable linear subsystems the impulse response $\{\lambda_i\}$ decays exponentially fast to zero, i.e., physical systems reveal short memory. Note that $l = 0$ corresponds to the classical leave-one-out criterion in (15). We refer to [29] for a discussion of the MCV criterion in the context of the standard regression analysis.

3.2 Partitioned Cross-Validation

Yet another data-driven procedure of the bandwidth choice in the presence of correlated errors is the Partitioned Cross-Validation (PCV) method originally introduced in [28, 29]. Here, for any $g \geq 1$, one splits the data set into disjoint g subgroups such that observations in the given group are apart from each other by the distance g. Hence, the training set $\mathcal{T}_n$ in (2) is divided into subsets $\{\mathcal{T}_{n,k}, k = 1,\ldots,g\}$, where

$$\mathcal{T}_{n,k} = \{(U_{(i-1)g+k}, Y_{(i-1)g+k}), i = 1,\ldots,n/g\}$$

is the k-th training data subgroup of the size n/g. Here, without loss of generality, we assume that n/g is an integer value. Next within each subsample $\mathcal{T}_{n,k}$ the leave-one-out criterion defined in (15) is used. Let us denote the $CV(h)$ criterion applied to the k-th subgroup as $CV_k(h)$. Consequently, the PCV selection rule is defined as follows

$$PCV(h) = g^{-1}\sum_{k=1}^{g} CV_k(h). \qquad (20)$$

Let $\widehat{h}'_{PCV}$ be the bandwidth that minimizes (20). It is important to note that $\widehat{h}'_{PCV}$ is the bandwidth corresponding to the reduced sample size n/g. Since the optimal bandwidth should be of the form $Cn^{-1/5}$, see (9), then one should correct $\widehat{h}'_{PCV}$ to adapt to the original data size n. Hence, the final PCV bandwidth choice is

$$\widehat{h}_{PCV} = g^{-1/5}\widehat{h}'_{PCV}, \qquad (21)$$

where $\widehat{h}'_{PCV}$ is the minimizer of (20). The parameter g should be obtained by some priori knowledge of the dynamical linear subsystem memory size. For the finite memory subsystem in (19) one should specify $g = p+1$. This makes the data within each subgroup $\mathcal{T}_{n,k}$ independent improving greatly the accuracy of the selected bandwidth $\widehat{h}_{PCV}$. Note finally that for $g = 1$ the PCV selector is equivalent to the ordinary leave-one-out criterion.

3.3 Corrected Cross-Validation

In [30] a class of the so-called corrected cross-validation methods has been developed. These techniques rely on certain transformations of the standard $CV(h)$ criterion in (15) and the estimated noise residuals in order to compensate the correlation present in the data set. The common characteristic of the corrected cross-validation techniques is that they explicitly employ the correlation structure of the observed data. Let us first recall, see (11), that the Hammerstein system can be written as

$$Y_t = m(U_t) + \varepsilon_t, \qquad (22)$$

with $\varepsilon_t = \sum_{i=1}^{\infty}\lambda_i m(U_{t-i}) + Z_t$. The dependent noise ε_t has the following covariance structure

$$\mathrm{Cov}[\varepsilon_t, \varepsilon_{t+l}] = \mathbb{E}[m^2(U)]\sum_{j=1}^{\infty}\lambda_j\lambda_{j+l} \qquad (23)$$

for $l \neq 0$ and $\mathrm{Var}[\varepsilon_t] = \mathbb{E}[m^2(U)]\sum_{j=1}^{\infty}\lambda_j^2 + \sigma_Z^2$. Let us denote $\mathrm{Cov}[\varepsilon_t, \varepsilon_{t+l}]$ as $\rho(l)$. Note that $\rho(l)$ is the correlation of the residual noise ε_t in (22). In order to estimate $\rho(l)$ we may use the estimated residuals

$$\widehat{\varepsilon}_i = Y_i - \widehat{m}_h(U_i), \quad (24)$$

for $i = 1, \ldots, n$, where $\widehat{m}_h(u)$ is defined in (6). In fact,

$$\widehat{\rho}_n(l) = \frac{1}{n-l} \sum_{i=1}^{n-l} \widehat{\varepsilon}_i \widehat{\varepsilon}_{i+l} \quad (25)$$

may serve as a consistent estimate of $\rho(l)$. This allows us to define the first corrected cross-validation criterion named in [30] as the direct method

$$DCV(h) = \sum_{i=1}^{n} \frac{\widehat{\varepsilon}_i^2}{(1 - \sum_{j=1}^{n} W_{h,j}(U_i)\widehat{\rho}_n(|j-i|))^2}. \quad (26)$$

Here

$$W_{h,j}(u) = \frac{K_h(u - U_j)}{\sum_{i=1}^{n} K_h(u - U_i)},$$

for $j = 1, \ldots, n$ are the weights corresponding to the kernel GRNN estimate in (6).

The method related to the direct corrected cross-validation criterion in (26) is the cross-validation technique based on the generalized cross-validation (GCV) procedure. Here, the selection criterion takes the following form

$$GDCV(h) = \sum_{i=1}^{n} \frac{\widehat{\varepsilon}_i^2}{(1 - n^{-1} tr(W_n R_n))^2}, \quad (27)$$

where W_n is the $n \times n$ matrix with the (i, j)-th component being $W_{h,j}(U_i)$, whereas R_n is the $n \times n$ matrix with the (i, j)-th component equal to $\widehat{\rho}_n(|j-i|)$.

In [30] yet another class of bandwidth selection techniques has been proposed referred to as the indirect corrected cross-validation method. Here, one is transforming the residual process $\{\widehat{\varepsilon}_i\}$ in (24) rather than the $CV(h)$ criterion as it was done in the direct method. The indirect corrected cross-validation criterion reads as

$$ICV(h) = \sum_{i=1}^{n} \frac{\widehat{\varepsilon'}_i^2}{(1 - W_{h,i}(U_i))^2}, \quad (28)$$

where $\{\widehat{\varepsilon}'_i\}$ are the transformed residuals being the components of $R_n^{-1/2}(\widehat{\varepsilon}_1, \cdots, \widehat{\varepsilon}_n)^T$. The latter is the empirical counterpart of the classical whitening transformation.

The counterpart of the GCV criterion, see (27), in the indirect strategy context is the following selection criterion

$$GICV(h) = \sum_{i=1}^{n} \frac{\widehat{\varepsilon'}_i^2}{(1 - n^{-1} \sum_{j=1}^{n} W_{h,j}(U_i))^2}. \quad (29)$$

4 Simulation Studies

In this section we evaluate the accuracy of the aforementioned bandwidth selection methods in the context of identification of the Hammerstein system. This is done for various choices of the linear and nonlinear characteristics of the system.

Throughout our studies we assume that the input signals $\{U_i\}$ is the white Gaussian process with zero mean and unit variance. The following types of nonlinear characteristics appearing in the the Hammerstein system, see Figure 1, are taken into account

– (N1) Polynomial: $m(u) = 0.5976(u^3 + u)$.

– (N2) Deadzone: $m(u) = 1.2866((u - 0.2)1(u \geq 0.2) + (u + 0.2)1(u \leq -0.2))$.

– (N3) ArcTan: $m(u) = 1.0808 \arctan(2u)$.

– (N4) ArcTan: $m(u) = 0.6995 \arctan(20u)$.

– (N5) Piecewise Constant: $m(u) = 1.1599[0.75 \cdot 1(0.2 \leq u \leq 0.5) + 1(u > 0.5) - 0.75 \cdot 1(-0.5 \leq u \leq -0.2) - 1(u \leq -0.5)]$,

where $1(A)$ is an indicator function, i.e., $1(A) = 1$ if $u \in A$ and $1(A) = 0$ otherwise.

The above nonlinearities reveal various degrees of smoothness and variability. Hence, the nonlinearity $N5$ has jump discontinuities, the nonlinearities $N3$ and $N4$ are continuous with the rapid change at $u = 0$. The characteristic $N2$ is piecewise linear, whereas $N1$ is the polynomial nonlinearity. Note that $N1$ is the most commonly used nonlinearity in applications. It should be noted that all nonlinearities are normalized, i.e., we have $\mathbb{E}[m(U)] = 0$ and $\mathbb{E}[m^2(U)] = 1$. The examined nonlinearities are depicted in Figure 2.

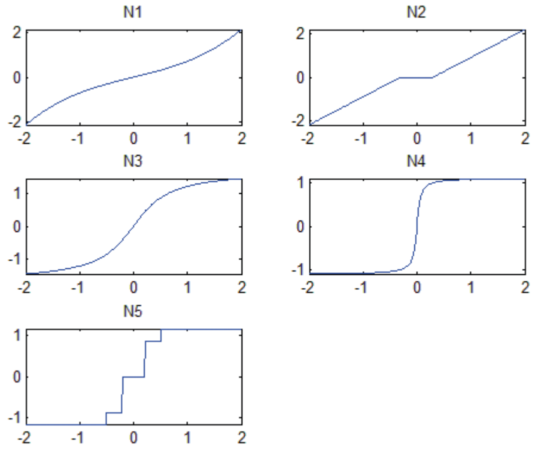

Figure 2. Nonlinear characteristics used in simulation studies.

Regarding the liner dynamical part of the Hammerstein system we will employ the following models

- (L1) AR(1): $W_t = \rho W_{t-1} + V_t$ with $\rho = -0.8$.
- (L2) AR(1): $W_t = \rho W_{t-1} + V_t$ with $\rho = -0.3$.
- (L3) AR(1): $W_t = \rho W_{t-1} + V_t$ with $\rho = 0$.
- (L4) AR(1): $W_t = \rho W_{t-1} + V_t$ with $\rho = 0.3$.
- (L5) AR(1): $W_t = \rho W_{t-1} + V_t$ with $\rho = 0.8$.
- (L6) The 3rd-order low-pass Butterworth filter with cut-off frequency 0.5π rad/sample.
  This filter has the following transfer function.
  $$H(z) = \frac{0.25 + 0.75z^{-1} + 0.75z^{-2} + 0.25z^{-3}}{1 + 0.3333z^{-2}}.$$
- (L7) The 3rd-order high-pass Butterworth filter with cut-off frequency 0.6π rad/sample. This filter is characterized by the following transfer function.
  $$H(z) = \frac{L(z)}{M(z)},$$
  where $L(z) = -0.388 + 1.164z^{-1} - 1.164z^{-2} + 0.388z^{-3}$ and $M(z) = 1 + 0.5772z^{-1} + 0.4218z^{-2} + 0.0563z^{-3}$.

We should note that the case $L3$ corresponds to the memoryless version of the Hammerstein system. The characteristics $L1$-$L5$ define the autoregressive linear models (AR) of order 1. Furthermore, all linear characteristics were normalized such that $\lambda_0 = 1$ in the convolution representation.

We will consider the Hammerstein system with all possible combinations of the aforementioned nonlinear/linear subsystems characteristics. For the purpose of making each experiment comparable, we set the noise $\{Z_i\}$ to be the *iid* Gaussian. Also to control the signal-to-noise ratio (SNR) we set $\frac{\mathbb{E}[Z^2]}{\mathbb{E}[W^2]} = 0.01$, which corresponds to the 20 dB SNR.

The data set $\mathcal{T}_n = \{(U_1, Y_1), \ldots, (U_n, Y_n)\}$ of the size $n = 200$ is used from which the kernel GRNN estimate in (6) is formed. The quadratic kernel function $K(u) = \frac{15}{16}(1 - u^2)^2 1(|u| < 1)$ has been employed.

The following data-driven bandwidth selection methods were chosen in our experimental studies.

- (M1) Leave-one-out CV.
- (M2) $MCV(l)$ with different values of l.
- (M3) $PCV(g)$ with different values of g.
- (M4) GDCV.
- (M5) GICV.

We denote the bandwidth selected by these methods by h_{CV}, $h_{MCV(l)}$, $h_{PCV(g)}$, h_{GDCV} and h_{GICV}, respectively.

The criterion chosen for measuring the accuracy of the given bandwidth selector is the $MISE(h)$ measure defined in (10). Note that $MISE(h)$ involves the averaging with respect the future test data as well as the averaging with respect to all possible training sets of the size n. To emphasize the latter dependence let us write $\widehat{m}_h(u; \mathcal{T}_n)$ for the kernel GRNN estimate determined from the training data $\mathcal{T}_n$ and tuned by the bandwidth h. In our simulation experiments we generate the (independent of the training set $\mathcal{T}_n$) test set

$$\{(U_1^T, Y_1^T), \cdots, (U_N^T, Y_N^T)\},$$

where N is very large integer being set to $N = 2000$ in our experiments. Next we generate a collection of training sets $\{\mathcal{T}_n^{[s]}, s = 1, \ldots, L\}$ of the same size n. In our experiments we set $L = 500$ and $n = 200$. These simulated test and training data sets allow us to experimentally evaluate the $MISE(h)$ in (10) as follows

$$\mathrm{MISE}(h) = \frac{1}{NL} \sum_{i=1}^{N} \sum_{s=1}^{L} \left(m(U_i^T) - \widehat{m}_h(U_i^T; \mathcal{T}_n^{[s]}) \right)^2, \tag{30}$$

where we some abuse of notation we use the symbol $MISE(h)$ for the simulation based version of the true $MISE(h)$ in (10).

Let us denote the ideal bandwidth value that minimizes the approximated $MISE(h)$ criterion in (30) by h^*. Also let $\widehat{h}$ refers to the one of the proposed methods for the data-driven bandwidth selection, i.e., h_{CV}, $h_{MCV(l)}$, $h_{PCV(g)}$, h_{GDCV} and h_{GICV}. In our first simulation experiment we determine the values of $\widehat{h}$, h^* and the corresponding estimation errors, i.e., $MISE(\widehat{h})$, $MISE(h^*)$. This is done for the aforementioned combinations of the nonlinear and linear characteristics of the Hammerstein system. In the case of the methods $MCV(l)$ and $PCV(g)$ we use several choices for l and g ranging from small to large values.

Tables 1-5 show the optimal bandwidth h^* and the bandwidth specified by the examined methods, i.e., h_{CV}, $h_{MCV(l)}$, $h_{PCV(g)}$, h_{GDCV} and h_{GICV}. The corresponding values of $MISE(h)$ are also shown in the square brackets.

We observe that the $PCV(g)$ method can greatly decrease the estimation error. In most cases, the $MCV(l)$ technique can also increase the estimation accuracy but not as much as the $PCV(g)$ algorithm. This is not the case for the GDCV and GICV methods that perform even worse than the classical leave-one-out CV.

To get the further qualitative insight into the accuracy of the examined data-driven bandwidth selectors let

$$\Delta(h,h^*) = MISE(h) - MISE(h^*)$$

be the distance between some h and h^* being the minimizer of $MISE(h)$ in (30). Clearly, $\Delta(h,h^*) \geq 0$. Then, we define the following relative accuracy index for the bandwidth selector $\widehat{h}$

$$\mathbb{S}(\widehat{h}) = \frac{\Delta(h_{CV},h^*) - \Delta(\widehat{h},h^*)}{\Delta(h_{CV},h^*)}, \qquad (31)$$

where h_{CV} is the classical leave-one-out bandwidth minimizing (15). Hence, the index $\mathbb{S}(\widehat{h})$ defines the relative accuracy measure of $\widehat{h}$ with respect to to h_{CV}. Clearly, $\mathbb{S}(h_{CV}) = 0$, and the large value of $\mathbb{S}(\widehat{h})$ indicates that the selector $\widehat{h}$ outperforms the classical CV method. Furthermore, if $\mathbb{S}(\widehat{h}) < 0$ then the selector $\widehat{h}$ works worst than the universal CV method.

In Figures 3-7 the $\mathbb{S}$-values for the $MCV(l)$ and $PCV(g)$ techniques are depicted for various combinations of the nonlinear and linear characteristics of the Hammerstein system. Specifically, $PCV(g)$ for $g = 2, 5, 10, 15$ are displayed. In Figure 6, however, only the case $g = 2$ is shown. The reason is that other values of g do not lead to useful results as it can be drawn from Table 4. Also we do not show the $\mathbb{S}$-value for the GDCV or GICV methods because they perform poorly as it can be learned from Tables 1-5.

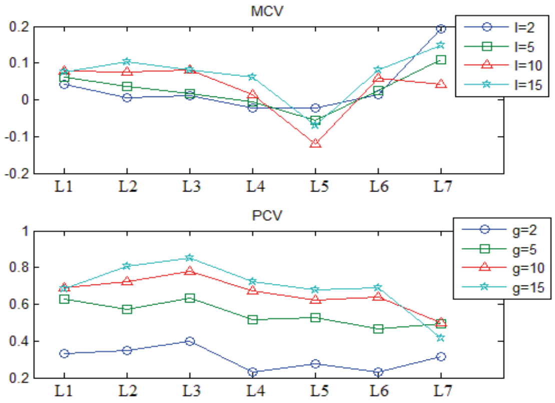

Figure 3. The index $\mathbb{S}$ values for the $MCV(l)$ and $PCV(g)$ methods versus l and g. The Hammerstein system with the N1-nonlinearity and linear subsystems $L1$-$L7$.

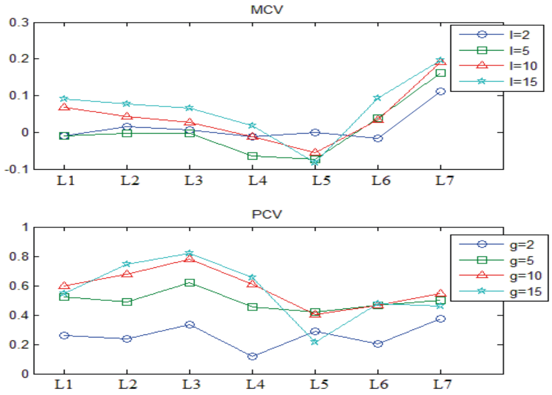

Figure 4. The index $\mathbb{S}$ values for the $MCV(l)$ and $PCV(g)$ methods versus l and g. The Hammerstein system with the N2-nonlinearity and linear subsystems $L1$-$L7$.

Table 1. The Hammerstein system with the N1-nonlinearity and various linear subsystems. The average value of $\widehat{h}$ ($\times 10^{-1}$) and the corresponding $MISE(\widehat{h})$ ($\times 10^{-1}$).

	L1	L2	L3	L4	L5	L6	L7
Optimal	10.5[1.15]	9.5[0.90]	9.5[0.89]	10.0[1.03]	11.0[2.12]	10.0[1.10]	10.5[1.11]
leave-1-out CV	8.5[1.38]	6.1[1.15]	5.6[1.23]	6.2[1.52]	8.7[2.40]	7.2[1.45]	10.1[1.47]
MCV (l=2)	8.6[1.37]	6.2[1.15]	5.6[1.23]	6.2[1.53]	8.5[2.41]	7.2[1.45]	10.3[1.40]
MCV (l=5)	8.7[1.37]	6.3[1.15]	5.7[1.22]	6.2[1.52]	8.5[2.42]	7.3[1.44]	10.3[1.44]
MCV (l=10)	8.7[1.37]	6.4[1.14]	5.8[1.20]	6.3[1.51]	8.4[2.44]	7.3[1.43]	10.2[1.46]
MCV (l=15)	8.8[1.37]	6.4[1.13]	5.9[1.20]	6.4[1.49]	8.5[2.42]	7.4[1.42]	10.4[1.42]
PCV (g=2)	9.1[1.31]	6.9[1.07]	6.6[1.09]	6.9[1.41]	9.1[2.33]	7.7[1.37]	10.9[1.36]
PCV (g=5)	9.9[1.24]	7.8[1.01]	7.5[1.01]	7.9[1.27]	9.9[2.25]	8.4[1.29]	11.8[1.29]
PCV (g=10)	10.7[1.22]	8.2[0.97]	8.0[0.96]	8.3[1.19]	10.7[2.23]	9.0[1.23]	12.6[1.29]
PCV (g=15)	11.2[1.22]	8.4[0.95]	8.2[0.94]	8.6[1.17]	11.1[2.21]	9.3[1.21]	13.3[1.32]
GDCV	8.2[1.45]	4.2[1.34]	2.1[1.91]	4.2[1.81]	8.1[2.54]	6.1[1.59]	10.2[1.91]
GICV	8.7[1.55]	4.2[1.36]	2.2[1.91]	4.2[1.81]	9.0[2.52]	6.2[1.61]	10.0[2.06]

Table 2. The Hammerstein system with the N2-nonlinearity and various linear subsystems. The average value of $\widehat{h}$ ($\times 10^{-1}$) and the corresponding $MISE(\widehat{h})$ ($\times 10^{-1}$).

	L1	L2	L3	L4	L5	L6	L7
Optimal	9.5[0.62]	7.5[0.28]	7.0[0.26]	7.0[0.26]	8.5[1.45]	7.5[0.33]	10.5[0.99]
leave-1-out CV	8.6[0.80]	5.9[0.42]	5.3[0.47]	5.9[0.39]	8.7[1.61]	7.1[0.43]	10.0[1.37]
MCV (l=2)	8.6[0.80]	5.9[0.42]	5.3[0.47]	5.9[0.40]	8.5[1.61]	7.0[0.43]	10.2[1.33]
MCV (l=5)	8.6[0.80]	6.0[0.42]	5.4[0.47]	6.0[0.40]	8.4[1.62]	7.1[0.43]	10.3[1.31]
MCV (l=10)	8.8[0.78]	6.1[0.42]	5.5[0.46]	6.1[0.40]	8.4[1.62]	7.2[0.43]	10.4[1.30]
MCV (l=15)	8.9[0.78]	6.2[0.41]	5.5[0.45]	6.2[0.39]	8.5[1.62]	7.4[0.42]	10.5[1.29]
PCV (g=2)	8.9[0.75]	6.5[0.39]	6.1[0.40]	6.6[0.38]	9.1[1.56]	7.4[0.41]	10.8[1.22]
PCV (g=5)	9.5[0.71]	7.1[0.35]	6.9[0.34]	7.2[0.33]	9.6[1.54]	7.9[0.38]	11.5[1.18]
PCV (g=10)	10.1[0.69]	7.5[0.32]	7.2[0.30]	7.5[0.31]	10.2[1.55]	8.2[0.38]	12.4[1.16]
PCV (g=15)	10.6[0.70]	7.8[0.31]	7.5[0.29]	7.8[0.31]	10.8[1.58]	8.6[0.38]	13.0[1.19]
GDCV	8.3[0.81]	4.0[0.50]	2.1[0.84]	4.1[0.47]	8.0[1.66]	6.0[0.46]	10.4[1.41]
GICV	8.8[0.87]	4.0[0.51]	2.1[0.84]	4.2[0.48]	9.1[1.70]	6.0[0.46]	10.1[1.53]

Table 3. The Hammerstein system with the N3-nonlinearity and various linear subsystems. The average value of $\widehat{h}$ ($\times 10^{-1}$) and the corresponding $MISE(\widehat{h})$ ($\times 10^{-1}$).

	L1	L2	L3	L4	L5	L6	L7
Optimal	9.0[0.49]	6.0[0.12]	5.5[0.07]	6.0[0.11]	8.5[1.32]	7.0[0.27]	11.5[0.93]
leave-1-out CV	8.3[0.70]	5.3[0.14]	4.5[0.12]	5.5[0.13]	8.3[1.46]	6.7[0.32]	10.1[1.38]
MCV (l=2)	8.3[0.71]	5.4[0.14]	4.5[0.11]	5.4[0.14]	8.2[1.45]	6.6[0.32]	10.3[1.38]
MCV (l=5)	8.4[0.71]	5.4[0.14]	4.6[0.11]	5.4[0.14]	8.1[1.47]	6.7[0.32]	10.4[1.36]
MCV (l=10)	8.5[0.69]	5.5[0.14]	4.7[0.11]	5.5[0.14]	7.9[1.49]	6.7[0.32]	10.4[1.38]
MCV (l=15)	8.7[0.68]	5.6[0.14]	4.7[0.11]	5.6[0.14]	7.9[1.48]	6.8[0.32]	10.6[1.35]
PCV (g=2)	8.6[0.64]	5.5[0.14]	5.0[0.10]	5.7[0.13]	8.7[1.41]	6.9[0.31]	10.6[1.26]
PCV (g=5)	8.9[0.57]	5.8[0.13]	5.5[0.08]	5.9[0.13]	8.7[1.39]	6.9[0.29]	11.1[1.14]
PCV (g=10)	9.1[0.54]	6.0[0.13]	5.7[0.08]	6.1[0.12]	9.0[1.38]	7.0[0.29]	11.6[1.08]
PCV (g=15)	9.4[0.54]	6.2[0.13]	6.0[0.08]	6.2[0.12]	9.4[1.37]	7.3[0.29]	12.0[1.11]
GDCV	8.5[0.65]	4.5[0.15]	2.5[0.19]	4.5[0.15]	8.3[1.43]	6.5[0.31]	10.8[1.27]
GICV	9.0[0.73]	4.5[0.15]	2.6[0.18]	4.5[0.15]	9.1[1.49]	6.5[0.32]	10.4[1.34]

Table 4. The Hammerstein system with the N4-nonlinearity and various linear subsystems. The average value of $\widehat{h}$ ($\times 10^{-1}$) and the corresponding $MISE(\widehat{h})$ ($\times 10^{-1}$).

	L1	L2	L3	L4	L5	L6	L7
Optimal	6.0[1.02]	2.5[0.25]	2.5[0.17]	2.5[0.23]	5.5[1.72]	4.0[0.48]	8.5[1.61]
leave-1-out CV	5.9[1.22]	2.7[0.27]	2.2[0.19]	2.7[0.25]	5.8[1.89]	3.9[0.54]	8.1[2.09]
MCV (l=2)	6.0[1.24]	2.7[0.27]	2.2[0.19]	2.7[0.25]	5.7[1.90]	3.9[0.54]	8.2[2.07]
MCV (l=5)	6.1[1.23]	2.7[0.27]	2.2[0.19]	2.7[0.25]	5.7[1.90]	3.9[0.54]	8.3[2.07]
MCV (l=10)	6.2[1.21]	2.8[0.27]	2.2[0.19]	2.7[0.25]	5.7[1.90]	4.0[0.54]	8.4[2.08]
MCV (l=15)	6.4[1.19]	2.8[0.27]	2.3[0.19]	2.8[0.25]	5.7[1.91]	4.0[0.54]	8.6[2.10]
PCV (g=2)	6.5[1.20]	3.0[0.27]	2.5[0.19]	2.9[0.25]	6.3[1.86]	4.2[0.54]	8.9[2.04]
PCV (g=5)	7.4[1.22]	3.5[0.28]	3.0[0.20]	3.5[0.26]	7.2[1.87]	4.9[0.55]	10.3[1.90]
PCV (g=10)	8.0[1.19]	4.0[0.31]	3.6[0.22]	4.0[0.29]	8.0[1.91]	5.3[0.57]	11.3[1.93]
PCV (g=15)	8.8[1.26]	4.4[0.33]	4.1[0.26]	4.5[0.32]	8.5[1.97]	5.8[0.61]	11.6[2.13]
GDCV	6.1[1.17]	2.1[0.28]	1.1[0.28]	2.2[0.26]	5.9[1.89]	3.7[0.54]	8.6[1.98]
GICV	6.2[1.23]	2.1[0.28]	1.1[0.28]	2.2[0.25]	6.3[1.91]	3.7[0.54]	8.5[2.04]

Table 5. The Hammerstein system with the N5-nonlinearity and various linear subsystems. The average value of $\widehat{h}$ ($\times 10^{-1}$) and the corresponding $MISE(\widehat{h})$ ($\times 10^{-1}$).

	L1	L2	L3	L4	L5	L6	L7
Optimal	7.5[0.78]	5.0[0.35]	2.0[0.29]	5.5[0.34]	7.5[1.64]	6.0[0.49]	10.0[1.30]
leave-1-out CV	7.6[0.94]	3.6[0.38]	2.7[0.32]	3.7[0.37]	7.7[1.82]	5.7[0.54]	9.4[1.74]
MCV (l=2)	7.6[0.94]	3.7[0.38]	2.7[0.32]	3.7[0.37]	7.4[1.83]	5.6[0.55]	9.5[1.75]
MCV (l=5)	7.7[0.93]	3.7[0.37]	2.8[0.32]	3.7[0.37]	7.3[1.83]	5.7[0.55]	9.5[1.78]
MCV (l=10)	7.8[0.93]	3.9[0.37]	2.9[0.32]	3.9[0.37]	7.3[1.83]	5.7[0.54]	9.6[1.77]
MCV (l=15)	7.9[0.93]	4.0[0.37]	3.0[0.32]	4.0[0.37]	7.3[1.83]	5.8[0.55]	9.8[1.74]
PCV (g=2)	7.7[0.90]	4.5[0.36]	3.8[0.31]	4.5[0.36]	7.8[1.76]	6.0[0.53]	9.9[1.66]
PCV (g=5)	8.1[0.86]	5.0[0.36]	4.5[0.32]	4.9[0.35]	8.3[1.74]	6.1[0.51]	10.9[1.52]
PCV (g=10)	8.5[0.86]	5.1[0.36]	4.7[0.32]	5.1[0.35]	8.6[1.74]	6.3[0.51]	11.3[1.53]
PCV (g=15)	8.6[0.88]	5.3[0.36]	5.0[0.32]	5.3[0.35]	8.9[1.76]	6.4[0.52]	11.9[1.61]
GDCV	7.8[0.88]	2.4[0.41]	0.7[0.62]	2.5[0.42]	7.6[1.80]	5.4[0.55]	9.9[1.65]
GICV	8.4[0.97]	2.4[0.42]	0.7[0.62]	2.5[0.41]	8.0[1.88]	5.5[0.55]	9.8[1.71]

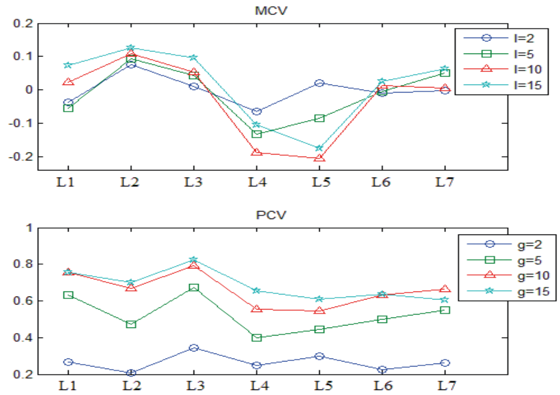

Figure 5. The index $\mathbb{S}$ values for the $MCV(l)$ and $PCV(g)$ methods versus l and g. The Hammerstein system with the N3-nonlinearity and linear subsystems $L1$-$L7$.

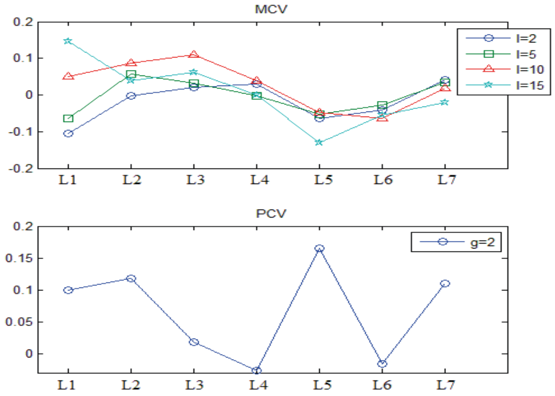

Figure 6. The index $\mathbb{S}$ values for the $MCV(l)$ and $PCV(g)$ methods versus l and g. The Hammerstein system with the N4-nonlinearity and linear subsystems $L1$-$L7$.

5 Discussion and Conclusions

Based on the simulations presented in Section 3, we can conclude that the $PCV(g)$ bandwidth selection method leads to the considerably smaller estimation error (MISE) comparing to the leave-one-out CV in the most examined cases. We have shown that for the Hammerstein system with the N1, N2, N3, N5 input nonlinearities, the optimal choice of g is of order $g = 10$ or $g = 15$. The corresponding $\mathbb{S}$-value is commonly close to 0.8 indicating the degree of the improvement of the $PCV(g)$ choice over the CV selection. Furthermore, the $PCV(g)$ method often leads to the estimation error being near to the optimal value $MISE(h^*)$.

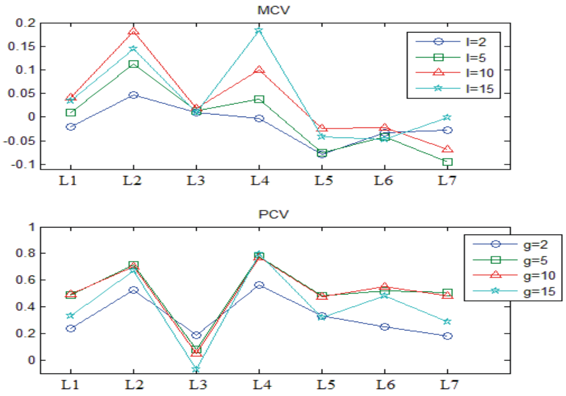

Figure 7. The index $\mathbb{S}$ values for the $MCV(l)$ and $PCV(g)$ methods versus l and g. The Hammerstein system with the N5-nonlinearity and linear subsystems $L1$-$L7$.

The fact that $g = 10$ (or $g = 15$) reveals the need of taking into account the correlation structure of the residual noise $\{\varepsilon_t\}$ appearing in the Hammerstein system as it is defined in (22). The special case is the nonlinearity N4, where the value $g = 2$ leads to the smallest estimation error. This takes place since the nonlinearity N4 varies quickly around the point $u = 0$, where the input density has the maximum value. As a result, splitting the data set into too many subgroups will make it more difficult to estimate the nonlinearity at points being far from $u = 0$. Nevertheless, the $PCV(g)$ method with $g = 2$ improves the estimation accuracy for the most linear subsystems connected with the N4 nonlinearity. Finally, note that for the combinations (L4, N4) and (L6, N4) the $PCV(g)$ method works comparable to the leave-one-out CV method.

Concerning the nonlinearity N5 being the piecewise constant function, the $PCV(g)$ method performs also very well. This is due to the fact that the jump points are not so significant as the rapid change in the nonlinearity N4 yielding the possibility to utilize all data points within the estimation interval.

The accuracy of the $PCV(g)$ algorithm depends on the choice of g. This can be selected in practice by using first the pilot kernel GRNN estimate where the bandwidth is specified by the leave-one-out CV method. Then, the obtained estimate can be visually inspected to detect the segments of rapid and constant intervals of the estimated function variability. The number of the low variability intervals can be used as a value for the parameter g that is related to the memory size of the dynamical part of the Hammerstein system.

In conclusion, the $PCV(g)$ method can greatly improve the accuracy of the kernel GRNN estimate in the context of the Hammerstein system identification and as such it is the method of choice in practical applications.

Regarding the $MCV(l)$ procedure it has been observed that it works poorly for the case of the positively correlated residual noise $\{\varepsilon_t\}$ in (22). Also the corresponding $\mathbb{S}$-value for $MCV(l)$ is significantly smaller than the one of the $PCV(g)$ method. The algorithms GDCV and GICV do not work well in almost all examined cases.

Finally, it is worth mentioning that aforementioned data-driven bandwidth selection methods have been developed for the fixed design regression case where the distribution of the input data does not play any significant role. In the Hammerstein system context we have stochastic input process with unknown distribution along with the complex residual noise $\{\varepsilon_t\}$ defined in (22).
For the related studies concerning the classical regression analysis we refer to [29, 34].

6 Future Work

This paper examines several cross-validation data-driven algorithms for selecting the bandwidth of the kernel GRNN estimate applied for nonparametric identification of the Hammerstein system. The conducted experimental studies reveal that the partitioned cross-validation (PCV) method can be recommended in practical applications of the Hammerstein system. Our paper is focusing on the choice of the global bandwidth. It would be a logical extension to consider similar studies for the local and semi-local bandwidth specifications. This would include the $k-$nearest neighbor methods and their extensions such as random forest [12, 13].

The examined bandwidth selection procedures can also be directly extended to the multiple-input Hammerstein system [35, 20], where one wishes to estimate the $d-$dimensional system nonlinearity $m(u)$, $u \in \mathbb{R}^d$. In this case the kernel GRNN estimate in (6) takes the form

$$\widehat{m}_h(u) = \frac{\sum_{i=1}^{n} Y_i K_h(||u - U_i||)}{\sum_{j=1}^{n} K_h(||u - U_j||)},$$

where $K_h(\cdot) = h^{-d} K(\cdot/h)$ is the scaled univariate kernel function. This is the single-bandwidth counterpart of the estimate in (6). The multiple bandwidth generalizations of $\widehat{m}_h(u)$ would be worth further studies.

Yet another extension of interest would be to consider the time-varying version of (11), i.e., when

$$Y_t = m_t(U_t) + \varepsilon_t,$$

where $m_t(\cdot)$ are functions that smoothly vary with time. In this case one should design kernel GRNN estimates that combine smoothing in both the input signal domain as well as the time domain, see [23] for some studies into this direction.

In addition, it is worthwhile to explore the bandwidth selection problem for other types of important block-oriented systems such as Wiener, sandwich and parallel models [6].

References

[1] K. Hunt, M. Munih, N. Donaldson, F. Barr, Investigation of the Hammerstein hypothesis in the modeling of electrically stimulated muscle, IEEE Transactions on Biomedical Engineering 45 (1998) 998–1009.

[2] T. Kara, I. Eker, Nonlinear modeling and identification of a DC motor for bidirectional operation with real time experiments, Energy Conversion and Management 45 (2004) 1087–1106.

[3] T. Quatieri, D. Reynolds, G. O'Leary, Estimation of handset nonlinearity with application to speaker recognition, IEEE Transactions on Speech and Audio Processing 8 (2000) 567–584.

[4] J. Turunen, J. Tanttu, P. Loula, Hammerstein model for speech coding, EURASIP Journal on Applied Signal Processing (2003) 1238–1249.

[5] E. Capobianco, Hammerstein system representation of financial volatility processes, The European Physical Journal B 27 (2002) 201–211.

[6] W. Greblicki, M. Pawlak, Nonparametric System Identification, Cambridge University Press, 2008.

[7] L. Ljung, System Identification: Theory for the User, Prentice Hall, New Jersey, 1987.

[8] F. Giri, E. Bai, Block-Oriented Nonlinear System Identification, Springer-Verlag, 2010.

[9] H.-F. Chen, W. Zhao, Recursive Identification and Parameter Estimation, CRC Press, 2014.

[10] R. Pintelon, J. Schoukens, System Identification: A Frequency Domain Approach, John Wiley & Sons, 2012.

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

[12] W. Greblicki, M. Pawlak, The weighted nearest neighbor estimate for Hammerstein system identification, IEEE Transactions on Automatic Control 64 (2019) 1550–1565.

[13] W. Greblicki, M. Pawlak, Hammerstein system identification with the nearest neighbor algorithm, IEEE Transactions on Information Theory 63 (2017) 4746–4757.

[14] J. G. Smith, S. Kamat, K. Madhavan, Modeling of ph process using wavenet based Hammerstein model, Journal of Process Control 17 (6) (2007) 551–561.

[15] K. Fruzzetti, A. Palazoğlu, K. McDonald, Nolinear model predictive control using Hammerstein models, Journal of Process Control 7 (1) (1997) 31–41.

[16] L. Jia, X. Li, M.-S. Chiu, Correlation analysis based mimo neuro-fuzzy Hammerstein model with noises, Journal of Process Control 41 (2016) 76–91.

[17] J. Wang, Q. Zhang, Detection of asymmetric control valve stiction from oscillatory data using an extended Hammerstein system identification method, Journal of Process Control 24 (1) (2014) 1–12.

[18] W. Wu, D.-W. Jhao, Control of a direct internal reforming molten carbonate fuel cell system using wavelet network-based Hammerstein models, Journal of Process Control 22 (3) (2012) 653–658.

[19] C. Qi, H.-T. Zhang, H.-X. Li, A multi-channel spatio-temporal Hammerstein modeling approach for nonlinear distributed parameter processes, Journal of Process Control 19 (1) (2009) 85–99.

[20] G. Harnischmacher, W. Marquardt, A multi-variate Hammerstein model for processes with input directionality, Journal of Process Control 17 (2007) 539–550.

[21] M. Pawlak, On the series expansion approach to the identification of Hammerstein systems, IEEE Transactions on Automatic Control 36 (6) (1991) 763–767.

[22] D. Specht, A general regression neural network, IEEE Transactions on Neural Networks 2 (1991) 568–576.

[23] P. Duda, M. Jaworski, L. Rutkowski, Convergent time-varying regression models for data streams: Tracking concept drift by the recursive Parzen-based generalized regression neural networks, International Journal of Neural Systems 28 (2018) 1750048.

[24] J. S. Marron, Automatic smoothing parameter selection: a survey, Empirical Economics 13 (3-4) (1988) 187–208.

[25] W. Härdle, P. Hall, J. S. Marron, How far are automatically chosen regression smoothing parameters from their optimum?, Journal of the American Statistical Association 83 (401) (1988) 86–95.

[26] J. Rice, Bandwidth choice for nonparametric regression, The Annals of Statistics (1984) 1215–1230.

[27] T. Gasser, A. Kneip, W. Köhler, A flexible and fast method for automatic smoothing, Journal of the American Statistical Association 86 (415) (1991) 643–652.

[28] J. S. Marron, Partitioned cross-validation, Econometric Reviews 6 (2) (1987) 271–283.

[29] C.-K. Chu, J. S. Marron, Comparison of two bandwidth selectors with dependent errors, The Annals of Statistics (1991) 1906–1918.

[30] N. Altman, Kernel smoothing of data with correlated errors, Journal of the American Statistical Association 85 (411) (1990) 749–759.

[31] J. Hart, P. Vieu, Data-driven bandwidth choice for density estimation based on dependent data, The Annals of Statistics 18 (2) (1990) 873–890.

[32] K. D. Brabanter, J. D. Brabanter, J. Suykens, Kernel regression in the presence of correlated errors, The Journal of Machine Learning Research 12 (2011) 1955–1976.

[33] Q. Yao, H. Tong, Cross-validatory bandwidth selections for regression estimation based on dependent data, Journal of Statistical Planning and Inference 68 (2) (1998) 387–415.

[34] A. Quintela del Río, Comparison of bandwidth selectors in nonparametric regression under dependence, Computational Statistics & Data Analysis 21 (5) (1996) 563–580.

[35] I. Goethals, K. Pelckmans, J. Suykens, B. De Moor, Identification of MIMO Hammerstein models using least squares support vector machines, Automatica 41 (2005) 1263–1272.

Jiaqing Lv received the Ph.D. degree in electrical and computer engineering from the University of Manitoba, Canada in 2019 under supervision of Prof. Pawlak. He is currently a researcher at the University of Manitoba and also holds a visiting position at the AGH University of Science and Technology, Poland. His research interests include machine learning and nonparametric system modeling with applications to power engineering systems. He has published his research results in the field of machine learning, large scale power systems as well as automation and control. He also holds a patent registered in the USA in the field of computer science.

Miroslaw Pawlak received the Ph.D. and D.Sc. degrees in computer engineering from Wrocław University of Technology, Wrocław, Poland. He is currently a Professor at the Department of Electrical and Computer Engineering, University of Manitoba, Canada. He has held a number of visiting positions in North American, Australian, and European Universities. He was at the University of Ulm, University in Goettingen and Marburg University as an Alexander von Humboldt Foundation Fellow. His research interests include statistical signal processing, machine learning, and nonparametric modeling. Among his publications in these areas are the books Image Analysis by Moments (Wrocław Univ. Technol. Press, 2006), and Nonparametric System Identification (Cambridge Univ. Press, 2008), coauthored with Prof. Włodzimierz Greblicki. Dr. Pawlak has been an Associate Editor of the Journal of Pattern Recognition and Applications, Pattern Recognition, International Journal on Sampling Theory in Signal and Image Processing, Journal of Artificial Intelligence and Soft Computing Research, Opuscula Mathematica and Statistics in Transition-New Series.