Sparse Data Modelling Using Combined Locally Regularized Orthogonal Least Squares and D-Optimality Design

S. Chen[†], X. Hong[‡] and C.J. Harris[†]

[†] Department of Electronics and Computer Science
University of Southampton, Southampton SO17 1BJ, U.K.

[‡] Department of Cybernetics
University of Reading, Reading RG6 6AY, U.K.

Abstract

The paper proposes to combine a locally regularized orthogonal least squares (LROLS) model selection with a D-optimality experimental design for efficient and robust sparse kernel data modelling. The LROLS algorithm alone is capable of producing a very parsimonious model with excellent generalization performance. The D-optimality design criterion further enhances the model efficiency and robustness. An added advantage is that the user only needs to specify a weighting for the D-optimality cost in the combined model selecting criterion and the entire model construction procedure becomes automatic. The value of this weighting does not influence the model selection procedure critically and it can be chosen with ease from a wide range of values.

Keywords: Sparse modelling, orthogonal least squares, regularization, Bayesian learning, experimental design, D-optimality.

I. Introduction

A basic principle in practical data modelling is the parsimonious principle. The orthogonal least squares (OLS) algorithm [1],[2] is an efficient learning procedure for constructing sparse regression models. If data are highly noisy, the parsimonious principle alone may not be entirely immune to overfitting, and small models constructed may still fit into noise. A useful technique for overcoming overfitting is regularization [3]–[5]. From the Bayesian viewpoint, a regularization parameter is equivalent to the ratio of the related hyperparameter to the noise parameter and an effective Bayesian learning method is the evidence procedure which iteratively optimizes model parameters and associated hyperparameters [6]. Adopting this Bayesian learning method to regression models, the LROLS algorithm [7]-[9] has recently been proposed, which introduces individual regularizer for each weight. This LROLS algorithm provides an efficient procedure for constructing sparse models that generalize well.

Optimal experimental designs [10] have been used to construct smooth model response surfaces based on the setting of the experimental variables under well controlled experimental conditions. In optimal design, model adequacy is evaluated by design criteria that are statistical measures of goodness of experimental designs by virtue of design efficiency and experimental effort. For kernel regression models, quantitatively, model adequacy is measured as function of the eigenvalues of the design matrix. There are a variety of optimal design criteria based on different aspects of experimental design [10]. The D-optimality criterion is most effective in optimizing the parameter efficiency and model robustness via the maximization of the determinant of the design matrix. In a recent work [11],[12], an effective model construction algorithm has been proposed based on the OLS algorithm coupled with the D-optimality design. This paper shows that further advantages can be gained by combining the LROLS algorithm with the D-optimality design

II. The kernel regression model

Consider the general kernel regression model of the form:

$$y(k) = \hat{y}(k) + e(k) = \sum_{i=1}^{n_M} \theta_i \phi_i(k) + e(k), \ \ 1 \le k \le N, \tag{1}$$

where $y(k)$ is the target, $e(k)$ is the error between $y(k)$ and the model output $\hat{y}(k)$, θ_i are the model weights, $\phi_i(k)$ are the regressors, n_M is the total number of candidate regressors, and N the number of training samples. By letting $\boldsymbol{\phi}_i = [\phi_i(1) \cdots \phi_i(N)]^T$, for $1 \le i \le n_M$, and defining

$$\mathbf{y} = \left[\begin{array}{c} y(1) \\ \vdots \\ y(N) \end{array} \right], \ \ \boldsymbol{\Phi} = [\boldsymbol{\phi}_1 \cdots \boldsymbol{\phi}_{n_M}], \ \ \boldsymbol{\theta} = \left[\begin{array}{c} \theta_1 \\ \vdots \\ \theta_{n_M} \end{array} \right], \ \ \mathbf{e} = \left[\begin{array}{c} e(1) \\ \vdots \\ e(N) \end{array} \right], \tag{2}$$

the regression model (1) can be written in the matrix form

$$\mathbf{y} = \mathbf{\Phi}\boldsymbol{\theta} + \mathbf{e}. \tag{3}$$

Let an orthogonal decomposition of the matrix $\mathbf{\Phi}$ be

$$\mathbf{\Phi} = \mathbf{W}\mathbf{A} \tag{4}$$

where

$$\mathbf{A} = \begin{bmatrix} 1 & a_{1,2} & \cdots & a_{1,n_M} \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & a_{n_M-1,n_M} \\ 0 & \cdots & 0 & 1 \end{bmatrix} \tag{5}$$

and

$$\mathbf{W} = [\mathbf{w}_1 \cdots \mathbf{w}_{n_M}] \tag{6}$$

with columns satisfying $\mathbf{w}_i^T\mathbf{w}_j = 0$, if $i \neq j$. The model (3) can alternatively be expressed as

$$\mathbf{y} = \mathbf{W}\mathbf{g} + \mathbf{e} \tag{7}$$

where the orthogonal weight vector $\mathbf{g} = [g_1 \cdots g_{n_M}]^T$ satisfy the triangular system $\mathbf{A}\boldsymbol{\theta} = \mathbf{g}$.

III. THE LOCALLY REGULARIZED OLS ALGORITHM WITH D-OPTIMALITY DESIGN

The LROLS algorithm adopts the following error criterion:

$$J_R(\mathbf{g}, \boldsymbol{\lambda}) = \mathbf{e}^T\mathbf{e} + \sum_{i=1}^{n_M} \lambda_i g_i^2 = \mathbf{e}^T\mathbf{e} + \mathbf{g}^T\mathbf{\Lambda}\mathbf{g} \tag{8}$$

where $\boldsymbol{\lambda} = [\lambda_1 \cdots \lambda_{n_M}]^T$ is the regularization parameter vector, and $\mathbf{\Lambda} = \text{diag}\{\lambda_1, \cdots, \lambda_{n_M}\}$. It can readily be shown that the criterion (8) can be expressed as

$$\mathbf{e}^T\mathbf{e} + \mathbf{g}^T\mathbf{\Lambda}\mathbf{g} = \mathbf{y}^T\mathbf{y} - \sum_{i=1}^{n_M} \left(\mathbf{w}_i^T\mathbf{w}_i + \lambda_i\right) g_i^2. \tag{9}$$

Normalizing (9) by $\mathbf{y}^T\mathbf{y}$ yields

$$\left(\mathbf{e}^T\mathbf{e} + \mathbf{g}^T\mathbf{\Lambda}\mathbf{g}\right) / \mathbf{y}^T\mathbf{y} = 1 - \sum_{i=1}^{n_M} \left(\mathbf{w}_i^T\mathbf{w}_i + \lambda_i\right) g_i^2 / \mathbf{y}^T\mathbf{y}. \tag{10}$$

As in the case of the OLS algorithm [1], the regularized error reduction ratio due to $\mathbf{w}_i$ is defined by

$$[\text{rerr}]_i = \left(\mathbf{w}_i^T\mathbf{w}_i + \lambda_i\right) g_i^2 / \mathbf{y}^T\mathbf{y}. \tag{11}$$

At each stage, a term is selected based on the selection criterion $l^* = \arg\max\{[\text{rerr}]_i, l \leq i \leq n_M\}$, and the selection process is terminated at the n_s-th stage when

$$1 - \sum_{l=1}^{n_s} [\text{rerr}]_l < \xi \tag{12}$$

is satisfied, where $0 < \xi < 1$ is a chosen tolerance. This produces a sparse model containing n_s $(\ll n_M)$ significant regressors. The Bayesian evidence procedure [6] can readily be used to optimize the regularization parameters. Applying this evidence procedure leads to the updating formulas:

$$\lambda_i^{\text{new}} = \frac{\gamma_i^{\text{old}}}{N - \gamma^{\text{old}}} \frac{\mathbf{e}^T\mathbf{e}}{g_i^2}, \ \ 1 \leq i \leq n_M, \tag{13}$$

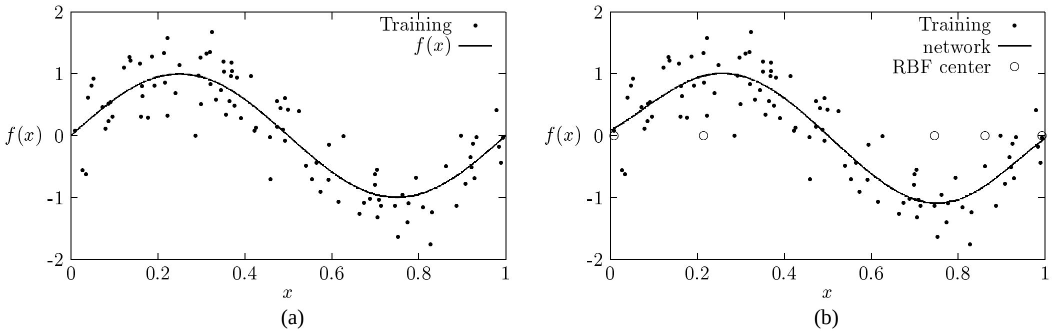


Fig. 1. Simple scalar function modelling problem. (a): Noisy training data y (dots) and underlying function $f(x)$ (curve); and (b): model mapping (curve) produced by the LROLS + D-optimality algorithm with $\beta = 10^{-5}$, circles indicate the RBF centers.

where

$$\gamma_i = \frac{\mathbf{w}_i^T \mathbf{w}_i}{\lambda_i + \mathbf{w}_i^T \mathbf{w}_i} \tag{14}$$

and

$$\gamma = \sum_{i=1}^{n_M} \gamma_i . \tag{15}$$

Usually a few iterations are sufficient to find an optimal $\boldsymbol{\lambda}$. The details of the algorithm is given in [9].

In experimental design, the data covariance matrix $\boldsymbol{\Phi}^T \boldsymbol{\Phi}$ is called the design matrix. The D-optimality design criterion maximizes the determinant of the design matrix for the constructed model. It is straightforward to verify that maximizing $\det(\boldsymbol{\Phi}^T \boldsymbol{\Phi})$ is identical to maximizing $\det(\mathbf{W}^T \mathbf{W})$ or, equivalently, minimizing $-\log \det(\mathbf{W}^T \mathbf{W})$ [11],[12]. Thus the combined LROLS and D-optimality algorithm can be viewed as based on the combined criterion

$$J_C(\mathbf{g}, \boldsymbol{\lambda}, \beta) = J_R(\mathbf{g}, \boldsymbol{\lambda}) + \beta \sum_{i=1}^{n_M} -\log(\mathbf{w}_i^T \mathbf{w}_i) \tag{16}$$

where β is a fixed small positive weighting for the D-optimality cost. In this combined algorithm, the updating of the model weights and regularization parameters is exactly as in the LROLS algorithm, but the selection is according to the combined error reduction ratio defined as

$$[\text{cerr}]_i = \left((\mathbf{w}_i^T \mathbf{w}_i + \lambda_i) g_i^2 + \beta \log(\mathbf{w}_i^T \mathbf{w}_i) \right) / \mathbf{y}^T \mathbf{y} \tag{17}$$

and the selection is terminated with an n_s-term model when

$$[\text{cerr}]_i \le 0 \ \text{ for } \ n_s + 1 \le i \le n_M . \tag{18}$$

The introduction of the D-optimality cost into the algorithm further enhances the efficiency and robustness of the selected subset model and, as a consequence, the combined algorithm can often produce sparser models with equally good generalization properties, compared with the LROLS algorithm. An additional advantage is that it simplifies the selection procedure. Notice that it is no longer needed to specify the tolerance ξ and the algorithm will terminate automatically when the condition (18) is reached. Unlike the combined OLS and D-optimality algorithm [11],[12], the value of weighting β does not critically influence the performance of this combined LROLS and D-optimality algorithm. The weighting β can (almost) be chosen arbitrarily from a large range of values.

IV. Modelling examples

Example 1. This example used a radial basis function (RBF) network to model the scalar function

$$f(x) = \sin(2\pi x), \ \ 0 \le x \le 1. \tag{19}$$

The RBF model employed Gaussian kernel function with a variance of 0.04. One hundred training data were generated from $y = f(x) + \epsilon$, where the input x was uniformly distributed in $(0,\ 1)$ and the noise ϵ

TABLE I

COMPARISON OF MODELLING ACCURACY FOR SIMPLE SCALAR FUNCTION MODELLING.

D-optimality	variance over noise training data		variance over noise-free testing data		number of terms	
weighting β	LROLS + D-opt	OLS + D-opt	LROLS + D-opt	OLS + D-opt	LROLS + D-opt	OLS + D-opt
1e-8	0.15766	0.14743	0.00168	0.02138	6	15
1e-7	0.15766	0.14743	0.00168	0.02138	6	15
1e-6	0.15823	0.14743	0.00202	0.02138	6	15
1e-5	0.15705	0.14743	0.00194	0.02138	5	15
1e-4	0.15826	0.14761	0.00246	0.02068	5	15
1e-3	0.15705	0.14933	0.00194	0.01585	5	12
1e-2	0.15705	0.15560	0.00194	0.00423	5	6
1e-1	0.15911	0.15544	0.00223	0.00427	5	6

was Gaussian with zero mean and variance 0.16. The noisy training points y and the underlying function $f(x)$ are plotted in Fig. 1 (a). As each training data x was considered as a candidate RBF center, there were $n_M = 100$ regressors in the model (1). The training data were very noisy. One hundred noise-free data $f(x)$ with equally spaced x were also generated as the testing data set for model validation.

Table I compares the mean square error values over the training and testing sets for the models constructed by the combined LROLS and D-optimality algorithm with those of the combined OLS and D-optimality algorithm, given a wide range of β values. It can be seen that using the D-optimality alone without regularization the constructed models can still fit into the noise unless the weighting β is set to some appropriate value. Combining regularization with D-optimality design, the results obtained are consistent over a wide range of β values. In the previous works [7]-[9], the LROLS algorithm alone produced a 6-term model with similar generalization performance as those produced by the combined LROLS and D-optimality algorithm. It is seen that the latter is capable of producing even sparser models. The model map of the 5-term model produced by the combined LROLS and D-optimality algorithm with $\beta = 10^{-5}$ is shown in Fig. 1 (b).

Example 2. This was a two-dimensional simulated nonlinear time series given by

$$\begin{aligned} y(k) &= \big(0.8 - 0.5\exp(-y^2(k-1))\big)\, y(k-1) - \big(0.3 + 0.9\exp(-y^2(k-1))\big)\, y(k-2) \\ &+ \; 0.1\sin(\pi y(k-1)) + \epsilon(k) \end{aligned} \tag{20}$$

where the noise $\epsilon(k)$ was Gaussian with zero mean and variance 0.09. One thousand noisy samples were generated given $y(0) = y(-1) = 0.0$. The first 500 data points were used for training, and the other 500 samples were used for model validation. The underlying noise-free system

$$\begin{aligned} y_d(k) &= \big(0.8 - 0.5\exp(-y_d^2(k-1))\big)\, y_d(k-1) - \big(0.3 + 0.9\exp(-y_d^2(k-1))\big)\, y_d(k-2) \\ &+ \; 0.1\sin(\pi y_d(k-1)) \end{aligned} \tag{21}$$

was specified by a limit circle, as shown in Fig. 2 (a). A Gaussian RBF model of the form

$$\hat{y}(k) = f_{RBF}(y(k-1), y(k-2)) \tag{22}$$

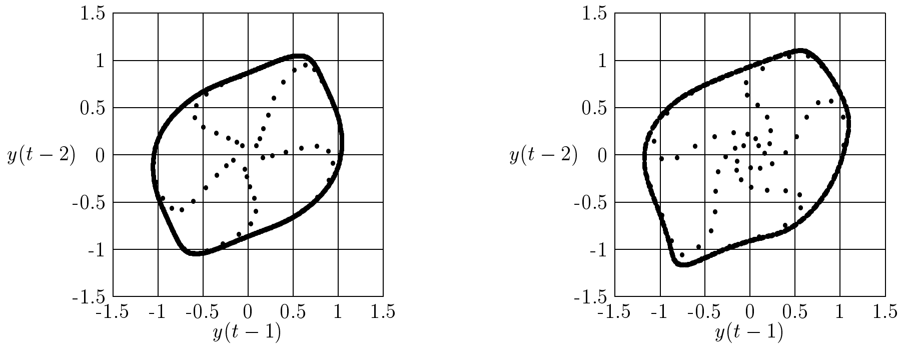
was constructed using the noisy training data. The Gaussian kernel function had a variance of 0.81. As each data point $[y(k-1)\; y(k-2)]^T$ was considered as a candidate RBF center, $n_M = 500$.

The modelling accuracies over both the training and testing sets are compared in Table II for the two algorithms with a range of β values. For this example, a 18-term model was produced using the LROLS algorithm alone in [9], and the resulting mean square errors over the training and testing set were 0.09264 and 0.09678, respectively. Again it is seen that the combined with the D-optimality design, the LROLS is able to produce sparser models with equally good generalization performance and the model construction

TABLE II

COMPARISON OF MODELLING ACCURACY FOR 2-DIMENSIONAL SIMULATED TIME SERIES MODELLING.

D-optimality	variance over training data		variance over testing data		number of terms	
weighting β	LROLS + D-opt	OLS + D-opt	LROLS + D-opt	OLS + D-opt	LROLS + D-opt	OLS + D-opt
1e-6	0.09275	0.07764	0.09635	2.53132	19	94
1e-4	0.09311	0.07762	0.09607	0.41540	13	93
1e-2	0.09338	0.08966	0.09750	0.09379	13	25
1e+0	0.09395	0.09360	0.09667	0.09627	13	14



(a)

(b)

Fig. 2. Two-dimensional time series modelling problem. (a): Phase plot of the noise-free time series ($y_d(0) = y_d(-1) = 0.1$); (b): Phase plot of the iterative RBF model output ($\hat{y}_d(0) = \hat{y}_d(-1) = 0.1$), the model was constructed by the LROLS + D-optimality algorithm with $\beta = 10^{-4}$.

process is insensitive to the value of β. The model produced by the combined LROLS and D-optimality algorithm with $\beta = 10^{-4}$ was used to iteratively generate the time series according to

$$\hat{y}_d(k) = f_{RBF}(\hat{y}_d(k-1), \hat{y}_d(k-2)) \tag{23}$$

given $\hat{y}_d(0) = \hat{y}_d(-1) = 0.1$. The resulting phase plot is shown in Fig. 2 (b).

Example 3. This example constructed a model representing the relationship between the fuel rack position (input) and the engine speed (output) for a Leyland TL11 turbocharged, direct injection diesel engine operated at low engine speed. It is known that at low engine speed, the relationship between the input and output is nonlinear [13]. Detailed system description and experimental setup can be found in [13]. The data set contained 410 samples. The first 210 data points were used in modelling and the last 200 points in model validation. A RBF model of the form:

$$\hat{y}(k) = f_{RBF}(y(k-1), u(k-1), u(k-2)) \tag{24}$$

was used to model the data. As each data vector $[y(k-1)\ u(k-1)\ u(k-2)]^T$ was considered as a candidate RBF center, there were $n_M = 210$ regressors in the regression model (1). The variance of the RBF kernel function was chosen to be 1.69.

The mean square errors of the models produced by the LROLS + D-optimality algorithm and the OLS + D-optimality one are compared in Table III, given a range of β values. For this real data set, a 34-term model was produced using the LROLS algorithm alone in [7]– [9], and the resulting mean square errors over the training and testing set were 0.000435 and 0.000487, respectively. The constructed RBF model by the combined LROLS and D-optimality algorithm with $\beta = 10^{-5}$ was used to generate the one-step prediction $\hat{y}(k)$ of the system output according to (24). The iterative model output $\hat{y}_d(k)$ was also produced using

$$\hat{y}_d(k) = f_{RBF}(\hat{y}_d(k-1), u(k-1), u(k-2)). \tag{25}$$

The one-step model prediction and iterative model output for this 22-term model selected by the LROLS + D-optimality algorithm are shown in Fig. 3, in comparison with the system output.

TABLE III

COMPARISON OF MODELLING ACCURACY FOR ENGINE DATA SET.

D-optimality	variance over training data		variance over testing data		number of terms	
weighting β	LROLS + D-opt	OLS + D-opt	LROLS + D-opt	OLS + D-opt	LROLS + D-opt	OLS + D-opt
1e-8	0.000459	0.000336	0.000488	0.000872	22	60
1e-7	0.000442	0.000345	0.000484	0.000831	27	58
1e-6	0.000441	0.000345	0.000479	0.000838	25	57
1e-5	0.000452	0.000429	0.000499	0.000517	22	24
1e-4	0.000586	0.000445	0.000606	0.000497	20	22
1e-3	0.000478	0.000503	0.000501	0.000536	20	19
1e-2	0.000884	0.000883	0.000982	0.000987	16	16
1e-1	0.004951	0.004951	0.005050	0.005052	12	12

V. Conclusions

A locally regularized OLS algorithm with the D-optimality design has been proposed for data modelling. It has been demonstrated that combining regularization with D-optimality experimental design provides a state-of-art procedure for constructing very sparse models with excellent generalization performance. It has been shown that the performance of the algorithm is insensitive to the D-optimality cost weighting, and the model construction process is fully automated. The computational requirements of this iterative model selection procedure are very simple and its implementation straightforward.

References

[1] S. Chen, S.A. Billings and W. Luo, "Orthogonal least squares methods and their application to non-linear system identification," *Int. J. Control*, Vol.50, No.5, pp.1873–1896, 1989.

[2] S. Chen, C.F.N. Cowan and P.M. Grant, "Orthogonal least squares learning algorithm for radial basis function networks," *IEEE Trans. Neural Networks*, Vol.2, No.2, pp.302–309, 1991.

[3] A.E. Hoerl and R.W. Kennard, "Ridge regression: biased estimation for non-orthogonal problems," *Technometrics*, Vol.12, pp.55–67, 1970.

[4] C.M. Bishop, "Improving the generalisation properties of radial basis function neural networks," *Neural Computation*, Vol.3, No.4, pp.579–588, 1991.

[5] S. Chen, E.S. Chng and K. Alkadhimi, "Regularised orthogonal least squares algorithm for constructing radial basis function networks," *Int. J. Control*, Vol.64, No.5, pp.829–837 1996.

[6] D.J.C. MacKay, "Bayesian interpolation," *Neural Computation*, Vol.4, No.3, pp.415–447, 1992.

[7] S. Chen, "Kernel-based data modelling using orthogonal least squares selection with local regularisation," in *Proc. 7th Annual Chinese Automation and Computer Science Conf. in U.K.* (Nottingham, U.K.), Sept.22, 2001, pp.27–30.

[8] S. Chen, "Locally regularised orthogonal least squares algorithm for the construction of sparse kernel regression models," to be presented at *6th Int. Conf. Signal Processing* (Beijing, China), Aug.26-30, 2002.

[9] S. Chen, "Local regularization assisted orthogonal least squares regression," submitted to *IEEE Trans. Neural Networks*, 2001.

[10] A.C. Atkinson and A.N. Donev, *Optimum Experimental Designs*. Oxford: Clarendon Press, 1992.

[11] X. Hong and C.J. Harris, "Nonlinear model structure design and construction using orthogonal least squares and D-optimality design," *IEEE Trans. Neural Networks*, to appear, 2002.

[12] X. Hong and C.J. Harris, "Experimental design and model construction algorithms for radial basis function networks," submitted to *IEEE Trans. Neural Networks*, 2002

[13] S.A. Billings, S. Chen and R.J. Backhouse, "The identification of linear and non-linear models of a turbocharged automotive diesel engine," *Mechanical Systems and Signal Processing*, Vol.3, No.2, pp.123–142, 1989.

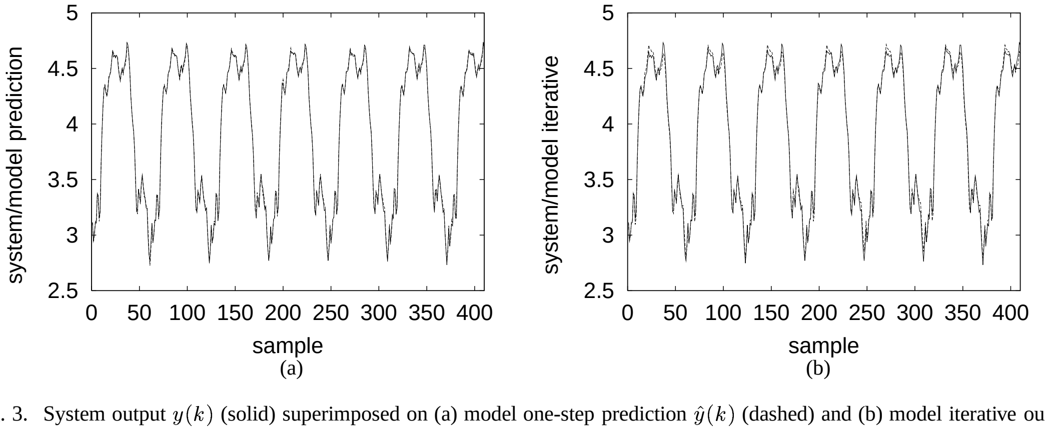


Fig. 3. System output $y(k)$ (solid) superimposed on (a) model one-step prediction $\hat{y}(k)$ (dashed) and (b) model iterative output $\hat{y}_d(k)$ (dashed). The model was selected by the LROLS + D-optimality algorithm with $\beta = 10^{-5}$.