Correspondence

Robust Nonlinear Model Identification Methods Using Forward Regression

X. Hong, C. J. Harris, S. Chen, and P. M. Sharkey

***Abstract*—In this correspondence new robust nonlinear model construction algorithms for a large class of linear-in-the-parameters models are introduced to enhance model robustness via combined parameter regularization and new robust structural selective criteria. In parallel to parameter regularization, we use two classes of robust model selection criteria based on either experimental design criteria that optimizes model adequacy, or the predicted residual sums of squares (PRESS) statistic that optimizes model generalization capability, respectively. Three robust identification algorithms are introduced, i.e., combined A- and D-optimality with regularized orthogonal least squares algorithm, respectively; and combined PRESS statistic with regularized orthogonal least squares algorithm. A common characteristic of these algorithms is that the inherent computation efficiency associated with the orthogonalization scheme in orthogonal least squares or regularized orthogonal least squares has been extended such that the new algorithms are computationally efficient. Numerical examples are included to demonstrate effectiveness of the algorithms.**

***Index Terms*—Cross validation, experimental design, forward regression, generalization, structure identification.**

I. Introduction

A large class of nonlinear models and neural networks can be classified as a kernel regression model [1]–[3]. The forward regression approach is an efficient model construction method [4] which selects regressors in a forward manner by virtue of their contribution to the maximization of the model error reduction ratio. Regularization techniques have been incorporated into the orthogonal least squares (OLS) algorithm to produce a regularized orthogonal least squares (ROLS) algorithm that reduces the variance of parameter estimates [5], [6]. To produce a model with good generalization capabilities, model selection criteria such as the Akaike information criterion (AIC) [7] are usually incorporated into the procedure to determinate the model construction process. Yet the use of AIC or other information based criteria, if used in forward regression, only affects the stopping point of the model selection, but does not penalize regressors that might cause poor model performance, e.g. too large parameter variance or ill-posedness of the regression matrix, if this is selected. This is due to the fact that AIC or other information based criteria are usually simplified measures derived as approximation formulas that is particularly sensitive to model complexity.

In order to achieve a model structure with improved model generalization, it is natural that a model generalization capability cost function should be used in the overall model searching process, rather than only being applied as a measure of model complexity. Optimum experimental designs have been used [8] to construct smooth network response surfaces based on the setting of the experimental variables under well controlled experimental conditions. In optimum 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. Quantitatively, model adequacy is measured as function of the eigenvalues of the design matrix. In a recent study [9], the authors have outlined an efficient learning algorithm, in which a composite cost function was introduced to optimize the model approximation ability using the forward orthogonal least squares (OLS) algorithm [10], and simultaneously determined model adequacy using an A-optimality design criterion (i.e., minimizes the variance of the parameter estimates). This algorithm has been further studied [11] as part of the B-spline based neurofuzzy model (NeuDec) and applied to model real dynamical processes. It was shown that the resultant models can be improved based on the reduction of parameter variance. There are a variety of optimality design criteria such as A- and D-optimality [8] based on different aspects of experimental design. The D-optimality criterion is most effective in optimising the parameter efficiency and model robustness via the maximization of the determinant of the design matrix. A new model construction and design algorithm using a new cost function based on the D-optimality criterion has been introduced [12]. This algorithm leads automatically to an unbiased model parameter estimate with an overall robust and parsimonious model structure. The efficiency of the algorithm lies in the construction of the new D-optimality based cost function based on the orthogonalization process to gain computational advantages, and hence to maintain the inherent advantage of computational efficiency associated with the conventional forward OLS approach.

Because the evaluation of model generalization capability is based directly on a concept of cross validation [13], it is desirable to develop new model selective criteria, based on the fundamental concept of cross validation, that can distinguish model generalization capability during the model construction process. A fundamental concept in cross validation is that of delete-1 cross validation in statistics, and the associated concept of the PRESS statistic [14] and [15]. These are usually computational expensive, however, recently an automatic nonlinear regression model construction algorithm has been introduced based on forward regression and the PRESS statistic that minimizes computational expense [16].

Because parameter regularization and robust model structure selection are effective and complementary approaches for robust modeling, it is also highly desirable to develop algorithms combining parameter regularization and model structure selection via direct model generalization capability for maximized model robustness. Recently a locally regularized orthogonal least squares (LROLS) algorithm has been introduced [17] which uses an effective Bayesian evidence method [18] to optimize local regularization parameters. The LROLS algorithm [17] alone is capable of producing a very parsimonious model with excellent generalization performance. Following [17], parameter regularization and robust model structural selection criteria have been combined for enhanced model robustness, and these are formed by combining a locally regularized orthogonal least squares (LROLS) model selection with a D-optimality experimental design [19], and by combining regularized orthogonal least squares and the PRESS statistic [20].

In [19], 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 specific value of this weighting factor does not

Manuscript received August 20, 2002; revised November 14, 2002. This paper was recommended by Associate Editor B. J. Oommen.

X. Hong and P. M. Sharkey are with the Department of Cybernetics, University of Reading, Reading RG6 6AY, U.K. (e-mail: x.hong@ reading.ac.uk; p.m.sharkey@reading.ac.uk).

C. J. Harris and S. Chen are with the Department of Electronics and Computer Science University of Southampton, Southampton SO17 1BJ, U.K. (e-mail: cjh@ecs.soton.ac.uk; sgc@ecs.soton.ac.uk).

Digital Object Identifier 10.1109/TSMCA.2003.809217

influence the model selection procedure critically and it can be chosen with ease from a wide range of values. In order to combine parameter regularization with regularized model structure selection based on the PRESS statistic [20], we initially derived the PRESS error in the orthogonal weight regularized model (see Appendix C). Due to the inherent computation efficiency associated with forward regularized orthogonal least squares, the effort involved in the computation of the PRESS statistic is minimized. The PRESS statistic is applied directly in the forward regression model structure construction process as a cost function in order to optimize the model generalization capability. The basic idea of improving computational efficiency is to reduce the computational expense to minimize the computation of PRESS errors, in which an inherent orthogonalization is used to avoid a matrix inversion. Further significant reduction in computation arises owing to the derivation of a forward recursive formula to compute PRESS errors. Based on the properties of the PRESS statistic the proposed algorithm can achieve a fully automatic procedure without resort to another validation data set for model assessment.

This paper systematically reviews recent advances on robust modeling techniques based on forward regression developed by the authors. The remainder of the paper is organized as follows, Section II provides a general background for the proposed algorithms, including optimal experimental criteria and the PRESS statistic. Section III, supplemented by Appendices, presents three robust algorithms and some appropriate analysis. Numerical examples in Section IV demonstrate the effectiveness of algorithms introduced in Section III-B and III-C. Section V is devoted to conclusions.

II. Preliminaries

A linear-in-the-parameters model [radial basis function (RBF) neural network, B-spline neurofuzzy network] can be formulated as [1] and [2]

$$y(t) = \sum_{k=1}^{M} p_k(\mathbf{x}(t))\theta_k + \xi(t) \tag{1}$$

where $t = 1, 2, \cdots, N$, and N is the size of the estimation data set. $y(t)$ is system output variable, $\mathbf{x}(t) = [y(t-1), \cdots, y(t-n_y), u(t-1), \cdots, u(t-n_u)]^T$ is system input vector of observables with assumed known dimension of $(n_y + n_u)$. $u(t)$ is system input variable. $p_k(\cdot)$ is a known nonlinear basis function, such as RBF, or B-spline fuzzy membership functions. $\xi(t)$ is an uncorrelated model residual sequence with zero mean and variance of σ^2. Equation (1) can be written in the matrix form as

$$\mathbf{y} = \mathbf{P}\Theta + \Xi \tag{2}$$

where $\mathbf{y} = [y(1), \cdots, y(N)]^T$ is the output vector, $\mathbf{\Theta} = [\theta_1, \cdots, \theta_M]^T$ is parameter vector, $\Xi = [\xi(1), \cdots, \xi(N)]^T$ is the residual vector, and $\mathbf{P}$ is the regression matrix

$$\mathbf{P} = \begin{bmatrix} p_1(\mathbf{x}(1)), & p_2(\mathbf{x}(1)) & \cdots & p_M(\mathbf{x}(1)) \\ p_1(\mathbf{x}(2)), & p_2(\mathbf{x}(2)) & \cdots & p_M(\mathbf{x}(2)) \\ \cdots\cdots & \cdots\cdots & \cdots & \cdots\cdots \\ p_1(\mathbf{x}(N)), & p_2(\mathbf{x}(N)) & \cdots & p_M(\mathbf{x}(N)) \end{bmatrix}.$$

By setting a cost function of $J_1 = \sum_{t=1}^{N}(y(t) - \sum_{k=1}^{M} p_k(\mathbf{x}(t))\theta_k)^2$, the least squares estimates of $\mathbf{\Theta}$ is given by [21]

$$\hat{\mathbf{\Theta}} = (\mathbf{P}^T\mathbf{P})^{-1}\mathbf{P}^T\mathbf{y}. \tag{3}$$

Assume that (2) represents the data generating process. If $\mathbf{P}^T\mathbf{P}$ is nonsingular, then

$$\begin{aligned} &(i)\ E\hat{\mathbf{\Theta}} = \Theta \\ &(ii)\ \text{cov}\hat{\mathbf{\Theta}} = \sigma^2(\mathbf{P}^T\mathbf{P})^{-1} \end{aligned} \tag{4}$$

where the matrix $(\mathbf{P}^T\mathbf{P})$ is called the design matrix. It is well known that a model based on least squares estimates tends to be unsatisfactory for a near ill conditioned regression matrix (or design matrix). The condition number of the design matrix is given by $C = (\max\lambda_k)/(\min\lambda_k)$, where $\lambda_k, (k = 1, \cdots, M)$ are the eigenvalues of the design matrix. Too large a condition number of the design matrix will result in unstable parameter estimates if a least squares algorithm is used [1] and [2], whilst a small condition number of the design matrix leads to model robustness. Optimum experimental designs have been used [8] to construct smooth system response surfaces based on the design or setting of the experimental variables under well controlled experimental conditions. Design criteria are statistical measures of goodness of experimental designs by virtue of design efficiency and experimental effort, and are aimed at avoiding model poor designs with unnecessarily large model parameter estimates' variance or extravagant model sizes that waste resources. Experimental design criteria of A-optimality and D-optimality are explained in Section II-A in order to provides background for the two model identification algorithms introduced in Sections III-A and Section III-B.

Alternatively, parameter estimates can be derived based on a regularized cost function of $J_r = \sum_{t=1}^{N}(y(t) - \sum_{k=1}^{M} p_k(\mathbf{x}(t))\theta_k)^2 + \sum_{k=1}^{M}\gamma_k\theta_k^2$, where $\gamma_k > 0$, $k = 1, 2, \cdots, M$ are regularization parameters. The regularized least squares estimates of $\hat{\mathbf{\Theta}}_r$, are given by [22]

$$\hat{\mathbf{\Theta}}_r = (\mathbf{P}^T\mathbf{P} + \Gamma)^{-1}\mathbf{P}^T\mathbf{y} \tag{5}$$

where $\Gamma = \text{diag}\{\gamma_1, \gamma_2, \cdots, \gamma_M\}$. The concept of parameter regulasation may be incorporated into a forward orthogonal least squares algorithm as a locally regularized orthogonal least square estimator (see Appendix A for details), which forms the foundation for all the robust identification algorithms introduced in this paper (see Section III).

A. Optimal Experimental Design Criteria

It is natural to consider model subset selection from an initial model base with M regressors in the framework of the optimal experiment design. In doing so, not just the model size has been taken into account, but different models of the same size can be evaluated. Consider the application of experimental design criteria in the context of model subset selection. In this section, we initially introduce the concepts of experimental design criteria including A-optimality and D-optimality based on using a fixed sized subset. The subset model is constructed from the full model with regression matrix $\mathbf{P}$ by using n_θ regressors selected from M regressors in $\mathbf{P}$, $n_\theta \ll M$. The resultant regression matrix is denoted $\mathbf{P}_k \in \Re^{N\times n_\theta}$, the resultant design matrix by $\mathbf{P}_k^T\mathbf{P}_k$, and λ_k, $k = 1, \ldots, n_\theta$ are the eigenvalues of $\mathbf{P}_k^T\mathbf{P}_k$.

Definition 1: A-optimality criterion minimizes the sum of the variance of a parameter estimate vector $\hat{\mathbf{\Theta}} = [\theta_1, \cdots, \theta_{n_\theta}]^T$

$$\min\left\{J_2 = \text{tr}\left[\text{cov}\hat{\mathbf{\Theta}}\right] = \sigma^2\sum_{k=1}^{n_\theta}\frac{1}{\lambda_k}\right\}. \tag{6}$$

Definition 2: The D-optimality criterion maximizes the determinant of the design matrix of $\mathbf{P}_k^T\mathbf{P}_k$

$$\max\left\{J_3 = \det\left(\mathbf{P}_k^T\mathbf{P}_k\right) = \prod_{k=1}^{n_\theta}\lambda_k\right\}. \tag{7}$$

This criterion favors models with smaller condition numbers to ensure model efficiency and robustness. It is well known that a model based on least squares estimates tend to be unsatisfactory for a near ill-conditioned regression matrix (or design matrix). The D-optimality criterion [8] inherently improves model robustness. Robust identification algorithms using the combined experimental design optimality criteria with regularized orthogonal least squares are introduced in Section III-A and III-B.

B. PRESS Statistic

Cross validation criteria are metrics that measures a model's generalization capability, which can alternatively be used as a model selection criterion for robustness. One commonly used version of cross-validation is the so called leave-one-out cross-validation. The idea is that, for any model, each data point in the estimation data set $D_N = \{\mathbf{x}(t), y(t)\}_{t=1}^{N}$ is sequentially set aside. In turn, a model is estimated using the remaining $(N-1)$ data, and the prediction error is derived using the data point that was removed. To select a model by using the leave-one-out cross-validation as the model selective criterion, the model with a minimal mean squares of the prediction errors is selected. There is an elegant way to generate the prediction error known as the PRESS statistic [15] for linear-in-the-parameters models (without actually sequentially splitting the estimation data set) by using the Sherman–Morrison–Woodbury Theorem [15].

Consider a predictor that is identified based on (1), the PRESS errors $\xi^{(-t)}(t|t-1)$ can be calculated using [15] as

$$\begin{aligned}\xi^{(-t)}(t|t-1) &= y(t) - \hat{y}^{(-t)}(t|t-1) \\ &= \frac{\xi(t)}{1 - \mathbf{p}(t)^T[\mathbf{P}^T\mathbf{P}]^{-1}\mathbf{p}(t)}\end{aligned} \tag{8}$$

where $\hat{y}^{(-t)}(t|t-1)$ is model prediction using leave-one-out data (without sample at t), $\mathbf{p}(t)$ is used for $\mathbf{p}(\mathbf{x}(t))$ for notational simplicity,and the PRESS statistic is computed by

$$J_p = E\left[[\xi^{(-1)}(t|t-1)]^2\right]. \tag{9}$$

A robust identification algorithm using the PRESS statistic and regularized orthogonal least squares is introduced in Section III-C.

III. Robust Nonlinear Model Identification Methods Using Forward Regression

New robust nonlinear model construction algorithms for a large class of linear-in-the-parameters models are introduced to enhance model robustness via combined parameter regularization and new robust structural selective criteria. Three algorithms are introduced: 1) two robust identification algorithms using combined A or D-optimality and regularized orthogonal least squares algorithm respectively, and 2) a robust identification algorithm using combined PRESS statistic and a regularized orthogonal least squares algorithm.

Unfortunately, the experimental design criteria of (6) and (7) or the PRESS statistic of (9) are inherently inefficient and computationally prohibitive if applied for model subset selection, due to the derivation of eigenvalues, exponential growth of possible subsets, matrix inversion, incuring the associated curse of dimensionality. Subset selection based on computationally efficient algorithms, such as forward regression is preferable. Locally parameter regularization may be combined with forward orthogonal least squares, and this forms a basis in the development of the proposed robust model construction algorithms introduced in Section III.

A key to understanding the development of these algorithms and their properties is to see that all of the algorithms are based on the orthogonalized space. Model selective cost functions or algorithm derivations are based on orthgonal basis $\mathbf{w}_k$ rather than original regressor $\mathbf{p}_k$, or auxiliary parameter vector $\mathbf{g}_k$ (associated with orthogonal space), rather than the original parameters θ_k. Due to this aspect, a common characteristic of these algorithms is that the inherent computation efficiency associated with the orthogonalization has been extended such that the new algorithms are computationally efficient. For simplicity of notation, as a function of forward regression step k, the resultant model selection criteria for all the proposed algorithms are denoted as $J^{(k)}$.

A. Robust Identification Using Combined A-Optimality and Regularized Orthogonal Least Squares

This algorithm is provided in Appendix B. Consider the A-optimality design criterion given in Definition 1, but based on model (37) (Appendix A) with orthogonal basis $\mathbf{w}_k$. The A-optimality cost function that minimizes the sum of the variance of the auxiliary parameter estimate vector $\mathbf{g} = [g_1, \cdots, g_{n_\theta}]^T$ for a subset model with n_θ regressors is given by

$$\min\left\{J_A = \mathrm{tr}[\mathrm{cov}\hat{\mathbf{g}}] = \sigma^2 \sum_{k=1}^{n_\theta} \frac{1}{\kappa_k}\right\}. \tag{10}$$

Since $\mathbf{A}\boldsymbol{\Theta} = \mathbf{g}$, it can be assumed that to penalize the large variance of the auxiliary parameter vector $\mathbf{g}$ will also consequently penalize large variance of parameter vector $\boldsymbol{\Theta}$.

A composite cost function is defined as

$$\begin{aligned}J &= J_1 + \alpha_1 J_A \\ &= \frac{1}{N}\left(\mathbf{y}^T\mathbf{y} - \sum_{k=1}^{n_\theta} g_k^2 \kappa_k\right) + \alpha \sum_{k=1}^{n_\theta} \frac{1}{\kappa_k}\end{aligned} \tag{11}$$

where, for the sake of simplicity, $\alpha = \sigma^2\alpha_1$, is a positive small number. Equation (11) can be directly incorporated into the conventional forward OLS algorithm to select the most relevant kth regressor at the kth forward regression stage, via

$$J^{(k)} = J^{(k-1)} - \frac{1}{N} g_k^2 \kappa_k + \frac{\alpha}{\kappa_k}. \tag{12}$$

At the kth forward regression stage, a candidate regressor is selected as the kth regressor if it produces the smallest $J^{(k)}$ and further reduction on $J^{(k-1)}$. The selection procedure will terminate if $J^{(k)} \geq J^{(k-1)}$ at the derived model size n_θ. This is significant because it means that the proposed approach can automatically detect a parsimonious model size.

The above A-optimality based design model construction algorithm was firstly introduced [9] and applied as part of the B-spline based neurofuzzy model (NeuDec) [11]. It was shown in [9] and [11] that the resultant models can be improved based on the reduction of model parameter variance.

B. Robust Identification Using Combined D-Optimality and Regularized Orthogonal Least Squares

This algorithm is provided in Appendix B. Consider the orthogonal decomposition of the subset model with regression matrix $\mathbf{P}_k$. Denote $\mathbf{P}_k = \mathbf{W}_k\mathbf{A}_k$, where $\mathbf{A}_k$ is a unit upper triangular matrix generated based on the orthogonal triangularization of $\mathbf{P}_k$. Consider the D-optimality design criterion given in Definition 2, but based on model (37) with orthogonal basis $\mathbf{w}_k$. The D-optimality design criterion that maximizes the determinant of the design matrix of $\mathbf{W}_k^T\mathbf{W}_k$ is given by

$$\max\{\mathrm{J}_{\mathrm{D}_0} = \det\left(\mathbf{W}_k^T\mathbf{W}_k\right) = \prod_{k=1}^{n_\theta} \kappa_k\}. \tag{13}$$

Because

$$\begin{aligned}\det(\mathbf{P}_k^T\mathbf{P}_k) &= \det(\mathbf{A}_k^T)\det(\mathbf{W}_k^T\mathbf{W}_k)\det(\mathbf{A}_k) \\ &= \det(\mathbf{W}_k^T\mathbf{W}_k)\end{aligned} \tag{14}$$

due to $\det(\mathbf{A}_k) = 1$; this establishes the equivalence of (7) and (13).

This implies that the selection of a subset of $\mathbf{P}_k$ from $\mathbf{P}$ is equivalent to the selection of a subset of $\mathbf{W}_k$ from $\mathbf{W}$, or that a better conditioned $\mathbf{P}_k$ can be achieved via a better conditioned $\mathbf{W}_k$ due to the equivalence of (7) and (13).

Unfortunately a composite cost function based on the linear combination of J_1 and J_{D_0} is unusable, because 1) J_1 is to be minimized and J_{D_0} is to be maximized; and 2) the algorithm efficiency cannot be achieved due to the product term in (13).

Construct instead the following cost function:

$$J_D = \psi(J_{D_0}) = -\log(J_{D_0}) = \sum_{k=1}^{n_\theta} \log\left[\frac{1}{\kappa_k}\right]. \tag{15}$$

Clearly the maximization of J_{D_0} is equivalent to the minimization of $\psi(J_{D_0})$, due to the fact that the solution of

$$\partial\psi(J_{D_0}) = -\frac{1}{J_{D_0}}\partial J_{D_0} = 0 \tag{16}$$

is equivalent to that of

$$\partial J_{D_0} = 0 \tag{17}$$

for $J_{D_0} > 0$ (as the design matrix is nonnegative).

The new augumented cost function is defined as

$$\begin{aligned} J &= J_1 + \beta J_D \\ &= \frac{1}{N}\left(\mathbf{y}^T\mathbf{y} - \sum_{k=1}^{n_\theta} g_k^2\kappa_k\right) + \beta\sum_{k=1}^{n_\theta}\log\left[\frac{1}{\kappa_k}\right] \end{aligned} \tag{18}$$

where β is a small positive number. Equation (18) can be incorporated into the forward OLS algorithm to select the most relevant kth regressor at the kth forward regression stage, via

$$J^{(k)} = J^{(k-1)} - \frac{1}{N}g_k^2\kappa_k + \beta\log\left[\frac{1}{\kappa_k}\right]. \tag{19}$$

At the kth forward regression stage, a candidate regressor is selected as the kth regressor if it produces the smallest $J^{(k)}$ and further reduction on $J^{(k-1)}$. Because J_D is an increasing function if $\kappa_k < 1$, which is true for some $k > K$, the selection procedure will terminate if $J^{(k)} \geq J^{(k-1)}$ at the derived model size n_θ if a proper β is set.

The robust identification using combined D-optimality and regularized orthogonal least squares based on the forward Gram-Schmidt procedure is shown in Appendix B. For the complete procedure, including optimization of regularization parameters, see [19], where an effective Bayesian evidence method [18] has been introduced to optimize local regularization parameters. Note that a very small positive regularization parameter such as ($\gamma_i = 10^{-6}, \forall i$) can be used to simplify the modeling procedure yet improve the resultant parameter estimate variance.

Clearly if $\beta = 0$, the proposed algorithm reduces to the conventional forward regularized OLS algorithm. By using the error reduction ratio (ERR), the conventional forward OLS algorithm equivalently uses $J_1^{(k)} = 1/N(\mathbf{y}^T\mathbf{y} - \sum_{k=1}^{n_\theta} g_k^2\kappa_k)$ as a selection criteria, which is a monotonically decreasing function of the iteration step k. (An error tolerance ρ is needed for the termination of the selection procedure.) This suggests that the selection procedure can be alternatively terminated in the usual way as in the conventional OLS procedure, but with any arbitrarily small β to improve model robustness. However in the proposed algorithm, the new cost function $J^{(k)}$ will increase after some selection stage; this follows by analysing the effect of the term $\beta\log[1/\kappa_k]$ in (19). An arbitrarily small β will detect the near singularity of the regression matrix during the selection procedure to prevent the selection of an oversized ill-posed model. Moreover in this algorithm the error tolerance ρ is not necessary for the termination of the selection procedure if an appropriate β is applied. Compared to the forward orthogonal least squares (OLS) algorithm [4] that combines the AIC criterion [7] in order to stop model construction before a high parameter estimate variance becomes a problem, the proposed method aims to prevent the high parameter estimate variance problem at the earliest selection stage. This should enhance the power of the forward orthogonal least squares (OLS) algorithm. Also note that the parameter regularization in the orthogonal parameter space is incorporated by using (39) for parameter estimation. (This can be clearly seen in Appendix B.) It has also been shown that using the D-optimality criterion can also enhance regularized orthogonal least squares [19].

Note that at any stage k, the decremental cost function (19) contains two terms: 1) $g_k^2\kappa_k/N$, which represents the model error reduction due to the inclusion of a new regressor; and 2) $\beta\log[1/\kappa_k]$, which represents overall model deterioration (the increase of condition number due to the inclusion of a new small eigenvalue from this additional regressor.) This latter term helps to distinguish bad model terms that have large parametric values g_k^2, but significantly small κ_k values (norm of an orthogonal basis). β should be sufficiently small so that new regressors are added to the model if and only if its contribution to model error reduction, $1/Ng_k^2\kappa_k$, is significant, here we require $\beta < g_k^2\kappa_k/[N\log 1/\kappa_k]$. If this inequality is violated, the model selection automatically stops. In practice, it is suggested that cross validation by using two data sets, an estimation set and a validation set, should be used to find an appropriate β. Cross validation based statistical information criterion such as AIC (or other statistical model selective criteria, e.g., GVC, FPE or MDL) should be directly applicable in the selection of β. It has also been shown [19] that value of β does not critically influence the model selection procedure and can be chosen from a wide range of values.

C. Robust Identification Using Combined PRESS Statistic and Regularized Orthogonal Least Squares

This algorithm is provided in Appendix B. Alternatively, the PRESS statistic of (9) that optimizes model generalation capability can be used as a robust model selective criterion. Note that (8) does not incorporate parameter regularization. In order to combine the PRESS statistic into a model with regularization and the forward regression learning algorithm, initially it is necessary to derive the PRESS error in an orthogonal weight regularized model. This is given in Appendix C. Consider the PRESS errors $\xi^{(-t)}(t|t-1)$ in (64) of Appendix C, which are based on the system in the orthogonalized form of (37), yielding

$$\begin{aligned} \xi^{(-t)}(t|t-1) &= y(t) - \hat{y}^{(-t)}(t|t-1) \\ &= \frac{\xi(t)}{1 - \mathbf{w}(t)^T[\mathbf{W}^T\mathbf{W} + \Gamma]^{-1}\mathbf{w}(t)} \\ &= \frac{\xi(t)}{\eta_M(t)} \end{aligned} \tag{20}$$

where

$$\eta_M(t) = 1 - \sum_{i=1}^{M}\frac{w_i^2(t)}{k_i + \gamma_i}. \tag{21}$$

The PRESS error, if not being computed on orthogonal regression space (associated with a diagnonal Hessian matrix), generally involves extensive computation due to matrix inversion [15]. However the amount of computation is significantly reduced by using [20], in which no matrix inversion is involved. This is due to 1) the calculation of the PRESS error being based on an orthogonalised model with a diagnonal Hessian matrix; and 2) the regularization in the orthogonal parameter vector $\mathbf{g}$ rather than the original parameter vector $\mathbf{\Theta}$.

In the following, it is shown that computational expense can be further significantly reduced by utilizing the forward regression process via a recursive formula. In the forward regression process, the model size is configured as a growing variable k. Consider the model construction by using a subset of k regressors ($k \ll M$); that is, a subset selected from the full model set consisting of M initial regressors [given by (2)] to approximate the system. The PRESS errors in (20) can be written, by replacing M with a variable model size k, as

$$\xi^{(-t)}(t|t-1) = \frac{\xi_k(t)}{\eta_k(t)} \tag{22}$$

where

$$\eta_k(t) = 1 - \sum_{i=1}^{k}\frac{w_i^2(t)}{\kappa_i + \gamma_i} \tag{23}$$

and $\xi_k(t)$ is the model residual associated with a subset model structure with k regressors. $\eta_k(t)$ can be written as a recursive formula, given by

$$\eta_k(t) = \eta_{k-1}(t) = \frac{w_k^2(t)}{\kappa_k + \gamma_\kappa}. \tag{24}$$

This is advantageous in that, for a new model with size increased from $(k-1)$ to k, the PRESS error coefficient $\eta_k(t)$ needs only to be adjusted for model size with a minimal computational effort.

As in conventional forward regression [4], a Gram–Schmidt procedure is used to construct the orthogonal basis $\mathbf{w}_k$ in a forward regression manner. At each regression step, the PRESS statistic can be computed using the algorithm and then used as a regressor selective criteria for model construction that minimizes the mean square PRESS errors

$$J^{(k)} = E\left[\left[\xi^{(-t)}(t|t-1)\right]^2\right] = E\left[\frac{[\xi_k(t)]^2}{\eta_k^2(t)}\right] = \frac{1}{N}\sum_{t=1}^{N}\frac{[\xi_k(t)]^2}{\eta_k^2(t)}. \quad (25)$$

Due to the properties associated with the minimization of the PRESS statistic, a fully automatic nonlinear predictive model contruction algorithm can be achieved. This can be initially explained intuitively. Analysis of the function $J^{(k)}$ shows that it is concave with respect to k, because $J^{(k)}$ is decremental for small k, while $E[\xi_k^2(t)]$ decreases significantly. As the model structure grows, and the decrease in $E[\xi_k^2(t)]$ (the contribution toward better model approximation due to the kth regressor) becomes negligible (as the model has achieved sufficient approximation capability), $J^{(k)}$ will become incremental due to PRESS error inflation, for some $k > n_\theta$. n_θ would clearly be the optimal model size based on the minimization of the PRESS statistic.

This point is clarified via the following analysis. Denote the model residual $\xi(t)$ for a model with size k as $\xi_k(t) = y(t) - \sum_{i=1}^{k} w_i(t)g_i$. Clearly

$$\xi_k(t) = \xi_{k+1}(t) + w_{k+1}(t)g_{k+1}. \quad (26)$$

From (25) and (26), the PRESS statistic for a model of size k is given by

$$J^{(k)} = E\left[\frac{[\xi_k(t)]^2}{\eta_k^2(t)}\right] = E\left[\frac{[\xi_{k+1}(t) + w_{k+1}(t)g_{k+1}]^2}{\eta_k^2(t)}\right] = E\left[\frac{[\xi_{k+1}(t)]^2}{\eta_k^2(t)}\right] + E\left[\frac{w_{k+1}^2(t)g_{k+1}^2}{\eta_k^2(t)}\right] \quad (27)$$

by assuming that the model residual sequence is uncorrelated with model regressors.

The change of $J^{(k)}$ over the kth regression step, by increasing k to $(k+1)$, can be written by

$$\Delta J = J^{(k+1)} - J^{(k)} = E\left[\frac{[\xi_{k+1}(t)]^2}{\eta_{k+1}^2(t)}\right] - E\left[\frac{[\xi_{k+1}(t)]^2}{\eta_k^2(t)}\right] - E\left[\frac{w_{k+1}^2(t)g_{k+1}^2}{\eta_k^2(t)}\right]. \quad (28)$$

The difference between the first two terms in (27), $E[([\xi_{k+1}(t)]^2)/(\eta_{k+1}^2(t))] - E[([\xi_{k+1}(t)]^2)/(\eta_k^2(t))]$, represents the effects of the PRESS errors inflation (from a model with k regressors to that of $(k+1)$ regressors). Clearly $E[([\xi_{k+1}(t)]^2)/(\eta_{k+1}^2(t))] - E[([\xi_{k+1}(t)]^2)/(\eta_k^2(t))] > 0$ for $\|w_{k+1}^2(t)\| \neq 0$, due to $\eta_{k+1}^2(t) < \eta_k^2(t)$. The effect of this property tends to increase $J^{(k)}$. Alternatively, the last term in (27), $E[(w_{k+1}^2(t)g_{k+1}^2)/(\eta_k^2(t))]$, representing the contribution of the kth regressor in model approximation, tends to decrease $J^{(k)}$. As the model achieves sufficient approximation capability at a certain model size $k = n_\theta + 1$, the last term in (27) becomes insignificant, such that this term is smaller than the effects of the PRESS errors inflation of $J^{(k)}$ (at a level of $\sim O(1/N)$ per regression step), resulting in $\Delta J > 0$. That is, $\Delta J > 0$ requires that $E[(w_{k+1}^2(t)g_{k+1}^2)/(\eta_k^2(t))] < E[([\xi_{k+1}(t)]^2)/(\eta_{k+1}^2(t))] - E[([\xi_{k+1}(t)]^2)/(\eta_k^2(t))]$. The forward regression model construction algorithm selects significant regressor $(\|w_{k+1}^2(t)\| \neq 0)$ that minimizes the PRESS statistic, with a growing model structure until $\Delta J > 0$ at a derived model size n_θ, when the contribution of the kth regressor becomes insignificant. The proposed algorithm terminates at $J^{(n_\theta+1)} > J^{(n_\theta)}$, where the model is optimized based on the minimization of the PRESS statistic at $J^{(n_\theta)}$.

This property (the sign change of ΔJ as k grows) can be applied to construct the automatic identification algorithm, introduced in Appendix B. This is based on the forward regression model construction with an incremental k, by simultaneously monitoring $J^{(k)}$. The procedure can be automatically terminated at a derived model with size n_θ, when $J^{(k)} > J^{(k-1)}$. Neither a separate criterion to terminate the selection procedure, nor any iteration of the procedure is needed (as the procedure does not use any predetermined controlling parameter to be adjusted via iteration).

The proposed algorithm is based on the standard Gram–Schmidt [4] Procedure in which the orthogonal basis $\mathbf{w}_k$ is constructed in a forward regression manner. At each regression step, the PRESS statistic can be

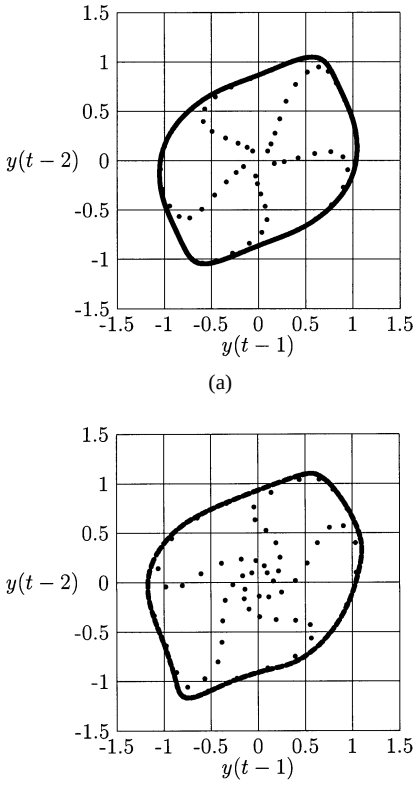

Fig. 1. Nonlinear time series modeling problem; (a) phase plot of the noise free time series; and (b) phase plot of the iterative RBF model output, using the combined D-optimality and ROLS algorithm with $\beta = 10^{-4}$.

TABLE I
COMPARISON OF MODELLING ACCURACY FOR NONLINEAR TIME SERIES MODELLING

D-optimality weighting β	MSE over training data		MSE over testing data		number of terms	
	LROLS and D-opt	OLS and D-opt	LROLS and D-opt	OLS and D-opt	LROLS and D-opt	OLS and 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

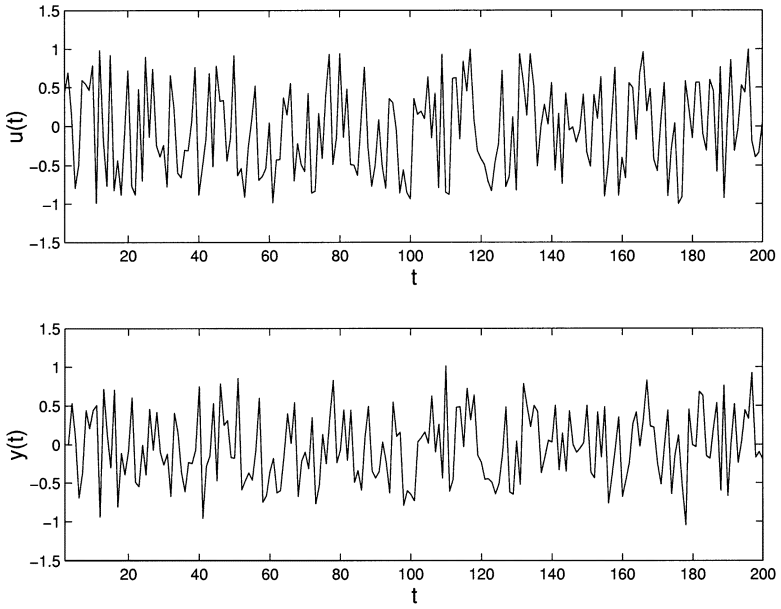

Fig. 2. System input-output observations in example 2.

formed and applied as a regressor selective criterion for model construction, as shown in Appendix B. Note that in this algorithm a very small positive regularization parameter such as ($\gamma_i = 10^{-6}, \forall i$) can be used to reduce parameter variance. Alternatively a similar optimization procedure of the regularization parameters can be implemented [17], [19], and this is under investigation by the authors.

IV. Numerical Examples

This section simply illustrates the operation of the robust algorithms introduced in Section III.B and III.C, respectively. Additional examples on simulated data and practical implementation of these algorithms can be found in [9], [11], [16], [17], [19], and [20].

Example 1: A simulated nonlinear time series given by

$$\begin{aligned} y(t) = &(0.8 - 0.5\exp(-y^2(t-1)))y(t-1) \\ &- (0.3 + 0.9\exp(-y^2(t-1)))y(t-2) \\ &+ 0.1\sin(\pi y(t-1)) + \xi(t) \end{aligned} \tag{29}$$

where the noise $\xi(t)$ is given as $\xi \sim N(0, 0.3^2)$. One thousand noisy samples were generated. The first 500 data points were used for training, and another 500 samples were used for model validation. The underlying noise-free system

$$\begin{aligned} y_d(t) = &(0.8 - 0.5\exp(-y_d^2(t-1)))y_d(t-1) \\ &- (0.3 + 0.9\exp(-y_d^2(t-1)))y_d(t-2) \\ &+ 0.1\sin(\pi y(t-1)) \end{aligned} \tag{30}$$

is specified by a limit cycle, as shown in Fig. 1(a). A Gaussian RBF model taking the form of

$$\hat{y}(t) = f_{RBF}(y(t-1),\, y(t-2)) \tag{31}$$

is used to construct the process underlying the noisy data. The Gaussian kernel function has a variance of 0.81.

The modeling accuracy over both the training and validation data sets are compared in Table 1 by using combined D-optimality with OLS or with ROLS, respectively. For this example, an 18–term model was produced using th LROLS algorithm alone (not using D-optimality) [17] and the resultant mean squares errors over the training and testing sets were 0.092 64 and 0.096 78, respectively. From the results shown in Table 1, it is seen that the combined D-optimality and regularized orthogonal least squares subset selection algorithm of Section III-B is able to produce sparser models with an equally good generalization performance. In addition, the model construction 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 interatively generate the time series according to

$$\hat{y}_d(t) = f_{RBF}(y_d(t-1),\, y_d(t-2)) \tag{32}$$

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

Example 2: Consider the following benchmark nonlinear dynamic system given by [23] and [24]

$$z(t) = \frac{z(t-1)z(t-2)z(t-3)u(t-2)[z(t-3)-1] + u(t-1)}{1 + z^2(t-2) + z^2(t-3)} \tag{33}$$

where the system input $u(t)$ is given as a uniformly distributed random signal in the range $[-1, 1]$, and $y(x) = z(x) + \xi$, in which the noise is $\xi \sim N(0, 0.05^2)$. As shown in Fig. 2, 200 data points were generated. The input vector is predetermined as a 5-input vector as $\mathbf{x}(t) = [y(t-1), y(t-2), y(t-3), u(t-1), u(t-2)]^T$. The Gaussian function $\phi(x, c_i) = \exp\{-\|x - c_i\|^2/\tau^2\}$ is used as basis functions to construct an RBF model, with a width $\tau = 1$. All 200 training data points are used as the candidate centre set. The proposed combined PRESS statistic and regularized orthogonal least squares subset selection algorithm of Section III-C was applied for automatic model structure detection, in which the regularization parameter was set as $\gamma_i = 10^{-6}, \forall i$. A parsimonious model structure can be detected at a derived model size when the PRESS statistic is minimized. During the forward regression model construction process, the PRESS statistic gradually decreases until $n_\theta = 37$, with an increment of $\Delta J = 1.97 \times 10^{-7} > 0$, such that the model with 37 centers is automatically derived as the final model. The results of the derived RBF model with 37 centers, are shown in

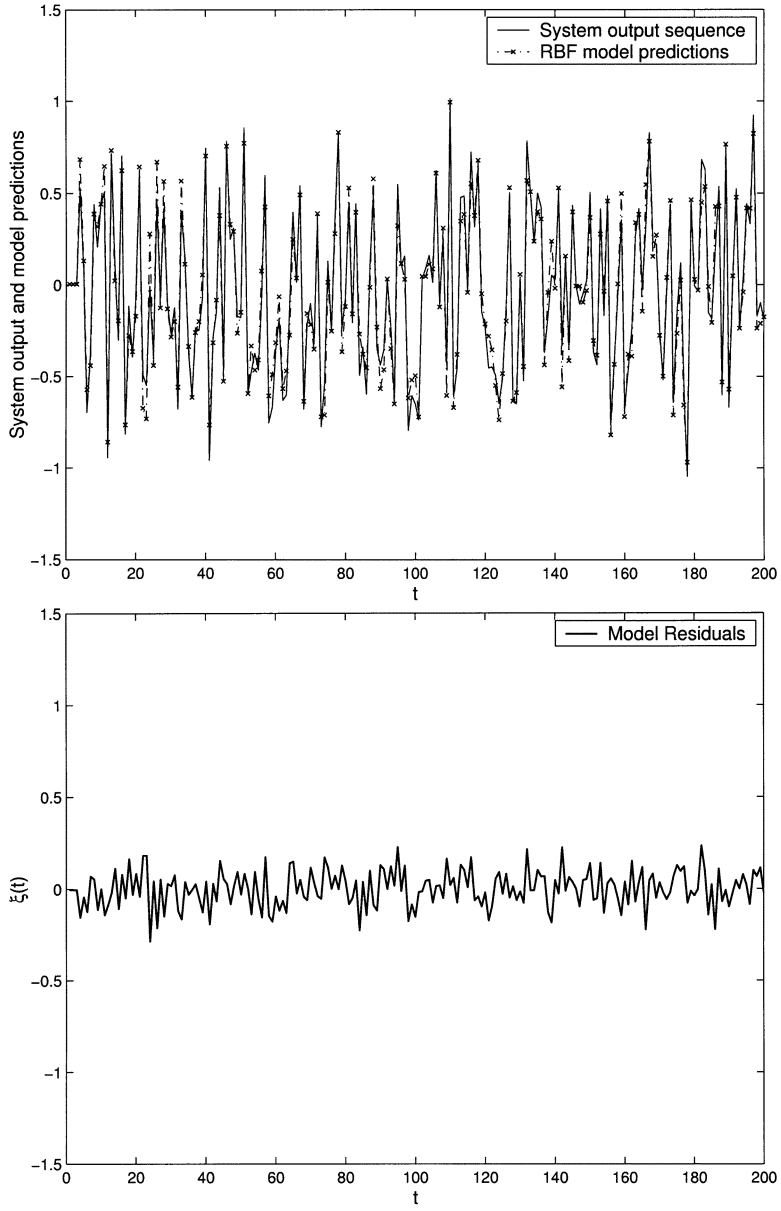

Fig. 3. Modeling results using RBF network with 37 centers (example 2).

Fig. 3. The model MSE and PRESS at $n_\theta = 37$, is 0.0995^2, and 0.11^2 respectively, demonstrating that the model is appropriate.

V. Conclusions

In this paper, we have introduced some robust nonlinear modeling techniques by developing algorithms for model construction in the framework of forward regression. In order to enhance model robustness, we use combined parameter regularization and new robust structural selective criteria, including two classes of robust model selective criteria based on either experimental design criteria that optimizes model adequacy, or the PRESS statistic that optimizes model generalization capability, respectively. In particular, a common feature of the algorithms is that computational efficiency has been achieved through the orthogonalization scheme in an orthogonal least squares or regularized orthogonal least squares algorithm. Significantly, the power of the well known forward orthogonal least squares (OLS) algorithm which was originally introduced based on model selection by maximizing output energy has been greatly exhanced for model selection based on various robustness objectives.

Appendix A
Locally Regularized Forward Orthogonal Least Squares Estimator

The LROLS procedure can automatically select a subset of n_θ regressors to construct a parsimonious model with parameter regularization. The forward orthogonal least squares estimator involves selecting a set of n_θ variables $\mathbf{p}_k = [p_k(1), \cdots, p_k(N)]^T, k = 1, \cdots, n_\theta$, from M regressors to form a set of orthogonal basis $\mathbf{w}_k, k = 1, \cdots, n_\theta$, in a forward regression manner. The principle of RLOLS algorithm for the structure determination is as follows.

An orthogonal decomposition of $\mathbf{P}$ is

$$\mathbf{P} = \mathbf{WA} \tag{34}$$

where $\mathbf{A} = \{a_{ij}\}$ is an $M \times M$ unit upper triangular matrix and $\mathbf{W}$ is an $N \times M$ matrix with orthogonal columns that satisfy

$$\mathbf{W}^T\mathbf{W} = \text{diag}\{\kappa_1, \ldots, \kappa_M\} \tag{35}$$

with

$$\kappa_k = \mathbf{w}_k^T\mathbf{w}_k, \quad k = 1, \ldots, M \tag{36}$$

so that (2) can be expressed as

$$\mathbf{y} = (\mathbf{P}\mathbf{A}^{-1})(\mathbf{A}\Theta) + \Xi = \mathbf{W}\mathbf{g} + \Xi \tag{37}$$

where $\mathbf{g} = [g_1, \cdots, g_M]^T$ is an auxiliary vector. The LROLS algorithm uses the following error criterion for parameter estimation.

$$J_r = \Xi^T\Xi + \mathbf{g}^T\Gamma\mathbf{g}. \tag{38}$$

Because $\xi(t)$ is uncorrelated with past output signals, it may be shown [4]and [5] that

$$g_k = \frac{\mathbf{w}_k^T\mathbf{y}}{\mathbf{w}_k^T\mathbf{w}_k + \gamma_k}, \quad k = 1, \cdots, M. \tag{39}$$

The original model coefficient vector $\mathbf{\Theta} = [\theta_1, \cdots, \theta_{n_\theta}]^T$ can then be calculated from $\mathbf{A}\mathbf{\Theta} = \mathbf{g}$ through backsubstitution. If $\gamma_i = 0, \forall i$, the procedure reduces to conventional forward OLS procedure [4], and if $\gamma_1 = \cdots = \gamma_M = \gamma$, this is then the regularized OLS algorithm with a global regularization parameter [5].

The ROLS procedure can use the conventional OLS procedure for model term selection which maximizes model approximation capability in a forward regression manner. The principle of the method is shown below. The number of all possible regressors M can be much larger than n_θ, but n_θ significant regressors can be identified using the forward OLS procedure. As the orthogonality property $\mathbf{w}_i^T\mathbf{w}_j = 0$ for $i \neq j$ holds, (37) is multiplied by itself and the time average is then taken, and the following equation is easily derived

$$\frac{1}{N}\mathbf{y}^T\mathbf{y} = \frac{1}{N}\sum_{k=1}^{M} g_k^2\mathbf{w}_k^T\mathbf{w}_k + \frac{1}{N}\Xi^T\Xi. \tag{40}$$

The output variance $E[y^2(t)] = (1/N)\mathbf{y}^T\mathbf{y}$ consists of two parts, $(1/N)\sum_{k=1}^{M} g_k^2\mathbf{w}_k^T\mathbf{w}_k$, the output variance explained by the regressors and $(1/N)\Xi^T\Xi$, the part of unexplained variance. The Error Reduction Ratio $[ERR]_k$, which is defined as the increment toward the overall output variance $E[y^2(t)]$ due to each regressor or input variable $p_k(t)$ divided by the overall output variance is computed as

$$[ERR]_k = \frac{g_k^2\mathbf{w}_k^T\mathbf{w}_k}{\mathbf{y}^T\mathbf{y}}, \quad k = 1, \cdots, M. \tag{41}$$

The most relevant n_θ regressors can be forward selected according to the value of the error reduction ratio $[ERR]_k$. At the kth selection, a candidate regressor is selected as the kth basis of the subset if it produces the largest value of $[ERR]_k$ from the remaining $(M - k + 1)$ candidates. By setting an appropriate tolerance ρ, which can be found by trial and error or via some statistical information criterion such as Akaike's information criterion(AIC) [7] that forms a compromise between the model performance and model complexity, the variable selection is terminated when

$$1 - \sum_{k=1}^{n_\theta}[ERR]_k < \rho. \tag{42}$$

This procedure can automatically select a subset of n_θ regressors to construct a parsimonious model. Equivalently, this procedure can be expressed as

$$J^{(k)} = J^{(k-1)} - \frac{1}{N}g_k^2\kappa_k \tag{43}$$

where $J^{(0)} = \mathbf{y}^T\mathbf{y}$. At the kth forward regression stage, a candidate regressor is selected as the kth regressor if it produces the smallest $J^{(k)}$. Equation (43) is used in the derivation of experimental design criteria based algorithms in Sections III.A and Section III.B.

APPENDIX B
THREE ALGORITHMS (COMBINED A- AND D-OPTIMALITY OR PRESS STATISTIC WITH ROLS SUBSET SELECTION ALGORITHMS)

1) At the first step, initialize $J^{(0)} = \mathbf{y}^T\mathbf{y}$, for $1 \leq i \leq M$

   For combined A-optimality and ROLS, set $\alpha > 0$

   For combined A-optimality and ROLS, set $\beta > 0$

   For combined PRESS statistics and ROLS,

   $\xi_0(t) = y(t), \eta_o(t) = 1, \forall t,$ i.e., $(t = 1, \cdots, N)$

   and compute

$$\begin{aligned} \mathbf{w}_1^{(i)} &= \mathbf{p}_i \\ \kappa_1^{(i)} &= (\mathbf{w}_1^{(i)})^T\mathbf{w}_1^{(i)} \\ g_1^{(i)} &= \frac{(\mathbf{w}_1^{(i)})^T\mathbf{y}}{\kappa_1^{(i)} + \gamma_1}. \end{aligned} \tag{44}$$

2) Use one of the optional block representing one of three different algorithms. See (45)–(47) as shown at the bottom of the page.
3) Find

$$J_{(i1)}^{(1)} = \min\left\{J_{(i)}^{(1)},\ 1 \leq i \leq M\right\} \tag{48}$$

Combined A-optimality and ROLS

$$J_{(i)}^{(1)} = J^{(0)} - \frac{1}{N}\left(g_1^{(i)}\right)^2\kappa_1^{(i)} + \frac{\alpha}{\kappa_1^{(i)}}. \tag{45}$$

or

Combined D-optimality and ROLS

$$J_{(i)}^{(1)} = J^{(0)} - \frac{1}{N}\left(g_1^{(i)}\right)^2\kappa_1^{(i)} + \beta\log\left[\frac{1}{\kappa_1^{(i)}}\right]. \tag{46}$$

or

Combined Press statistic and ROLS

$$\begin{aligned} \xi_1^{(i)}(t) &= \xi_0(t) - w_1^{(i)}(t)g_1^{(i)}, \quad (t = 1, \cdots, N) \\ \eta_1^{(i)}(t) &= \eta_0^{(i)}(t) - \frac{[w_1^{(i)}(t)]^2}{\kappa_1^{(i)} + \gamma_1}, \quad (t = 1, \cdots, N) \\ J_{(i)}^{(1)} &= \frac{1}{N}\sum_{t=1}^{N}\frac{[\xi_1^{(i)}(t)]^2}{[\eta_1^{(i)}(t)]^2}. \end{aligned} \tag{47}$$

Combined A-optimality and ROLS

$$J_{(i)}^{(k)} = J^{(k-1)} - \frac{1}{N}\left(g_k^{(i)}\right)^2 \kappa_k^{(i)} + \frac{\alpha}{\kappa_k^{(i)}}. \quad (51)$$

or

Combined D-optimality and ROLS

$$J_{(i)}^{(k)} = J^{(k-1)} - \frac{1}{N}\left(g_k^{(i)}\right)^2 \kappa_k^{(i)} + \beta \log\left[\frac{1}{\kappa_k^{(i)}}\right]. \quad (52)$$

or

Combined PRESS statistic and ROLS

$$\xi_k(t) = \xi_{k-1}(t) - w_k^{(i)}(t) g_k^{(i)}, \quad (t = 1, \cdots, N)$$

$$\eta_k^{(i)}(t) = \eta_{k-1}^{(i)}(t) - \frac{\left[w_k^{(i)}(t)\right]^2}{\kappa_k^{(i)} + \gamma_k}, \quad (t = 1, \cdots, N)$$

$$J_{(i)}^{(k)} = \frac{1}{N}\sum_{t=1}^{N} \frac{[\xi_k(t)]^2}{\left[\eta_k^{(i)}(t)\right]^2}. \quad (53)$$

and select

$$\mathbf{w}_1 = \mathbf{w}_1^{(i_1)} = \mathbf{p}_{i_1}$$
$$J^{(1)} = J_{(i_1)}^{(1)}. \quad (49)$$

4) At the kth step where $k \geq 2$ for $1 \leq i \leq M, i \neq i_1 \ldots i \neq i_{k-1}$, compute

$$a_{jk}^{(i)} = \frac{\mathbf{w}_j^T \mathbf{p}_i}{\mathbf{w}_j^T \mathbf{w}_j}, \quad 1 \leq j < k$$
$$\mathbf{w}_k^{(i)} = \mathbf{p}_i - \sum_{j=1}^{k-1} a_{jk}^{(i)} \mathbf{w}_j$$
$$\kappa_k^{(i)} = (\mathbf{w}_k^{(i)})^T \mathbf{w}_k^{(i)}$$
$$g_k^{(i)} = \frac{(\mathbf{w}_k^{(i)})^T \mathbf{y}}{\kappa_k^{(i)} + \gamma_k} \quad (50)$$

5) Use one of the optional block representing one of three different algorithms. See (51)–(53) as shown at the top of the page.

6) Find

$$J_{(i_k)}^{(k)} = \min\{J_{(i)}^{(k)}, 1 \leq i \leq M, i \neq i_1, \ldots i \neq i_{k-1}\} \quad (54)$$

and select

$$a_{jk} = a_{jk}^{(i_k)}$$
$$\mathbf{w}_k = \mathbf{w}_k^{(i_k)} = \mathbf{p}_{i_k} - \sum_{j=1}^{k-1} a_{jk} \mathbf{w}_j$$
$$J^{(k)} = J_{(i_k)}^{(k)}. \quad (55)$$

7) The procedure is monitored and terminated at the derived $k = n_\theta$ step, when $J^{(k)} \geq J^{(k-1)}$. Otherwise, set $k = k + 1$, go to step 2.

Appendix C
PRESS Error in an Orthogonal Weight Regularized Model

Following (39), the parameter vector in an orthogonal weight regularized model is

$$\mathbf{g} = [\mathbf{W}^T \mathbf{W} + \Gamma]^{-1} \mathbf{W}^T \mathbf{y} = \mathbf{H}^{-1} \mathbf{W}^T \mathbf{y} \quad (56)$$

where $\Gamma = \text{diag}\{\gamma_1, \cdots, \gamma_M\} \in \Re^{M \times M}$. The model residual based on the orthogonal weight regularized model is

$$\xi(t) = y(t) - \mathbf{g}^T \mathbf{w}(t) = y(t) - \mathbf{y}^T \mathbf{W} \mathbf{H}^{-1} \mathbf{w}(t). \quad (57)$$

If the data sample indexed at t is deleted from estimation data set, the leave-one-out model parameter vector from the orthogonal weight regularized model is given by

$$\mathbf{g}^{(-t)} = \left\{[\mathbf{W}^{(-t)}]^T \mathbf{W}^{(-t)} + \Gamma\right\}^{-1} [\mathbf{W}^{(-t)}]^T \mathbf{y}^{(-t)}$$
$$= [\mathbf{H}^{(-t)}]^{-1} [\mathbf{W}^{(-t)}]^T \mathbf{y}^{(-t)} \quad (58)$$

where $\mathbf{W}^{(-t)}$ and $\mathbf{y}^{(-t)}$ denote the resultant regression matrix, and output vector respectively. By derivation, it can be shown that

$$\mathbf{H}^{(-t)} = \mathbf{H} - \mathbf{w}(t)\mathbf{w}^T(t). \quad (59)$$
$$[\mathbf{y}^{(-t)}]^T \mathbf{W}^{(-t)} = \mathbf{y}^T \mathbf{W} - y(t)\mathbf{w}^T(t) \quad (60)$$

The PRESS error evaluated at t in an orthogonal weight regularized model is given by

$$\xi^{(-1)}(t|t-1) = y(t) - [\mathbf{g}^{(-t)}]^T \mathbf{w}(t)$$
$$= y(t) - [\mathbf{y}^{(-t)}]^T \mathbf{W}^{(-t)} [\mathbf{H}^{(-t)}]^{-1} \mathbf{w}(t). \quad (61)$$

From (59), and by using the matrix inversion lemma, yields

$$[\mathbf{H}^{(-t)}]^{-1} = [\mathbf{H} - \mathbf{w}(t)\mathbf{w}^T(t)]^{-1}$$
$$= \mathbf{H}^{-1} + \frac{\mathbf{H}^{-1}\mathbf{w}(t)\mathbf{w}^T(t)\mathbf{H}^{-1}}{1 - \mathbf{w}^T(t)\mathbf{H}^{-1}\mathbf{w}(t)} \quad (62)$$

and

$$[\mathbf{H}^{(-t)}]^{-1} \mathbf{w}(t) = \frac{\mathbf{H}^{-1}\mathbf{w}(t)}{1 - \mathbf{w}^T(t)\mathbf{H}^{-1}\mathbf{w}(t)}. \quad (63)$$

Substituting (60) and (63) into (61), yields

$$\xi^{(-1)}(t|t-1) = y(t) - [\mathbf{y}^T \mathbf{W} - y(t)\mathbf{w}^T(t)] \times \frac{\mathbf{H}^{-1}\mathbf{w}(t)}{1 - \mathbf{w}^T(t)\mathbf{H}^{-1}\mathbf{w}(t)}$$
$$= \frac{y(t) - \mathbf{y}^T \mathbf{W} \mathbf{H}^{-1} \mathbf{w}(t)}{1 - \mathbf{w}^T(t)\mathbf{H}^{-1}\mathbf{w}(t)}$$
$$= \frac{\xi(t)}{1 - \mathbf{w}^T(t)\mathbf{H}^{-1}\mathbf{w}(t)}$$
$$= \frac{\xi(t)}{1 - \mathbf{w}^T(t)[\mathbf{W}^T\mathbf{W} + \Gamma]^{-1}\mathbf{w}(t)}. \quad (64)$$

Acknowledgment

The authors would like to thank the referees for their constructive comments.

REFERENCES

[1] C. J. Harris, X. Hong, and Q. Gan, *Adaptive Modeling, Estimation and Fusion From Data: A Neurofuzzy Approach*. New York: Springer-Verlag, 2002.

[2] M. Brown and C. J. Harris, *Neurofuzzy Adaptive Modeling and Control*. Englewood Cliffs, NJ: Prentice-Hall, 1994.

[3] L. X. Wang and J. M. Mendel, "Fuzzy basis functions, universal approximation, and orthogonal least squares learning," *IEEE Trans. Neural Networks*, vol. 3, pp. 807–814, Sept. 1992.

[4] S. Chen, S. A. Billings, and W. Luo, "Orthogonal least squares methods and their applications to nonlinear system identification," *Int. J. Control*, vol. 50, pp. 1873–1896, 1989.

[5] S. Chen, Y. Wu, and B. L. Luk, "Combined genetic algorithm optimization and regularized orthogonal least squares learning for radial basis function networks," *IEEE Trans. Neural Networks*, vol. 10, pp. 1239–1243, Sept. 1999.

[6] M. J. L. Orr, "Regularization in the selection of radial basis function centers," *Neural Comput.*, vol. 7, no. 3, pp. 954–975, 1993.

[7] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Automat. Contr.*, vol. AC-19, pp. 716–723, Dec. 1974.

[8] A. C. Atkinson and A. N. Donev, *Optimum Experimental Designs*. Oxford, U.K.: Clarendon, 1992.

[9] X. Hong and C. J. Harris, "Nonlinear model structure detection using optimum design and orthogonal least squares," *IEEE Trans. Neural Networks*, vol. 12, pp. 435–439, Mar. 2001.

[10] S. Chen, C. F. N. Cowan, and P. M. Grant, "Orthogonal least squares learning algorithm for radial basis function neural networks," *IEEE Trans. Neural Networks*, vol. 2, pp. 302–309, Mar. 1991.

[11] X. Hong and C. J. Harris, "Neurofuzzy design and model construction of nonlinear dynamical processes from data," *IEE Proc D. Control Theory Applications*, vol. 148, no. 6, pp. 530–538, 2001.

[12] ——, "Nonlinear model structure design and construction using orthogonal least squares and D-optimality design," *IEEE Trans. Neural Networks*, vol. 13, pp. 1245–1250, Sept. 2002.

[13] M. Stone, "Cross validatory choice and assessment of statistical predictions," *J. R. Statist. Soc. B.*, vol. 36, pp. 117–147, 1974.

[14] L. Breiman, "Stacked regression," *Mach. Learn.*, vol. 5, pp. 49–64, 1996.

[15] R. H. Myers, *Classical and Modern Regression With Applications*, 2nd ed. Boston, MA: PWS-KENT, 1990.

[16] X. Hong, P. M. Sharkey, and K. Warwick, "Automatic nonlinear predictive model construction algorithm using forward regression and the PRESS statistic," , 2002, submitted for publication.

[17] S. Chen, "Locally regularization assisted orthogonal least squares regression," , 2001, submitted for publication.

[18] D. J. C. MacKay, "Bayesian interpolation," *Neural Comput.*, vol. 4, no. 3, pp. 415–447, 1992.

[19] S. Chen, X. Hong, and C. J. Harris, "Sparse kernel regression modeling using combined locally regularized orthogonal least squares and D-optimality experimental design," , 2002, submitted for publication.

[20] X. Hong, S. Chen, and P. M. Sharkey, "Automatic kernel regression modeling using combined PRESS statistic and regularized orthogonal least squares," , 2002, submitted for publication.

[21] T. Soderström and P. Stoica, *System Identification*. Englewood Cliffs, NJ: Prentice-Hall, 1989.

[22] D. W. Marquardt, "Generalized inverse, ridge regression, biased linear estimation and nonlinear estimation," *Technometrics*, vol. 12, no. 3, pp. 591–612, 1970.

[23] K. S. Narendra and K. Parthasarathy, "Identification and control of dynamic systems using neural networks," *IEEE Trans. Neural Networks*, vol. 1, pp. 4–27, Mar. 1990.

[24] J. H. Nie and T. H. Lee, "Rule-based modeling: fast construction and optimal manipulation," *IEEE Trans. Syst., Man, Cybern. A*, vol. 26, pp. 728–738, Nov. 1996.

Dynamical Neural Networks for Planning and Low-Level Robot Control

Mathias Quoy, Sorin Moga, and Philippe Gaussier

***Abstract*—We use dynamical neural networks based on the neural field formalism for the control of a mobile robot. The robot navigates in an open environment and is able to plan a path for reaching a particular goal. We will describe how this dynamical approach may be used by a high level system (planning) for controlling a low level behavior (speed of the robot). We give also results about the control of the orientation of a camera and a robot body.**

***Index Terms*—Dynamical systems, neural networks, robot control and planning.**

I. INTRODUCTION

Our research team develops architectures for the control of mobile robots. These robots are able to navigate in an open environment and to find a path toward a particular goal. There are several different approaches for solving this problem ([1], and [2] and references therein]). The recent ones which relate more directly to our work rely on potential functions [3], [4] or neural network (NN) implementations of dynamical systems [5], [6]. See also the more classical NN approach to control given in [7]–[9]. The potential functions (or potential fields) approach is not a new one [10]. The main drawbacks are the existence of local minima and a difficult use in changing environments. It also relies on a Cartesian map of the environment. This map is not always available or accurate enough. Dynamical systems in a NN formalism try to avoid the shortcomings of this approach.

Following Polderman and Willems [11], we will define a *behavior* as the solution of a given dynamical system. From an ethological point of view, a behavior will correspond to a particular trajectory for going from one location to another, for instance. Note that the behavior being the solution of a dynamical system does not mean that the whole trajectory may be computed knowing the initial conditions. The solution may be defined implicitly. Thus the differential equation has to be solved numerically. This definition of a behavior is much used in an animat approach [12]. Indeed, one often defines the behavior as the *transient* dynamics leading from an initial condition to the attractor of the system. It is the case in the potential field case for instance. The nature of the attractor in this case is far less important than the transient dynamics. If the attractor is a fixed point, then the system stays on it and does not evolve anymore. We would like here to extend the notion of behavior to the dynamics *on the attractor*. In this case, a stable behavior can be reached when the system is on an attractor (among others: stable fixed point, limit cycle or even chaotic strange attractors [13], [14]). In this article, we will only consider the case of fixed points with asymptotic stability as defined in [15] for instance. In that case the robot keeps

Manuscript received March 8, 2002; revised November 20, 2002. This work was supported by two French GIS contracts on Cognitive Sciences "comparison of control architectures for the problem of action selection" in collaboration with Animat lab (J.Y. Donnart, A. Guillot, J.A. Meyer), LISC (G. Deffuant) and RFIA (F. Alexandre, H. Frezza), and "Mobile robots as simulation tools for the dynamics of neurobiological models: adaptation and learning for visual navigation tasks" in collaboration with CRNC (G. Schöner) and LAAS (R. Chatila). This paper was recommended by Associate Editor R. A. Hess.

M. Quoy and P. Gaussier are with the Neurocybernetics team, ETIS, Université de Cergy-Pontoise—ENSEA, BP 44, 95014, Cergy-Pontoise, France (e-mail: mathias.quoy@dept-info.u-cergy.fr).

S. Moga is with the Department IASC, ENST Bretagne, 29238 Brest, France.

Digital Object Identifier 10.1109/TSMCA.2003.809224