Orthogonal Least Squares Based on Singular Value Decomposition for Spare Basis Selection

Min Han and De-cai Li

School of Electronic and Information Engineering, Dalian University of Technology, Dalian 116023, China
minhan@dlut.edu.cn

Abstract. This paper proposes an improved orthogonal least square algorithm based on Singular Value Decomposition for spare basis selection of the linear-in-the-weights regression models. The improved algorithm is based on the idea of reducing meaningless calculation of the selection process through the improvement of orthogonal least square by using the Singular Value Decomposition. This is achieved by dividing the original candidate bases into several parts to avoid comparing among poor candidate regressors. The computation is further simplified by utilizing the Singular Value Decomposition to each sub-block and replacing every sub-candidate bases with the obtained left singular matrix, which is a unitary matrix with lower dimension. It can avoid the computation burden of the repeated orthogonalisation process before each optimal regressor is determined. This algorithm is applied to the linear-in-the-weights regression models with the predicted residual sums of squares (PRESS) statistic and minimizes it in an incremental manner. For several real and benchmark examples, the present results indicate that the proposed algorithm can relieve the load of the heave calculation and achieve a spare model with good performance.

Keywords: singular value decomposition, orthogonal least square, predicted residual sums of squares (PRESS) statistic, spare bases selection.

1 Introduction

In many situations, multivariate time series are required to model the complex dynamic of chaotic systems [1, 2]. It has been shown that predictions using multivariate time series may be significantly better than those using univariate time series [3]. Although the multivariate inputs can provide more information for modeling, the increment of the inputs also means more complex model structure, which would produce very poor generalization and heavy time consuming. In order to achieve accurate predictions with multivariate inputs, the effective complexity of the model has to be controlled based on the principle that ensures the smallest possible model that fits the training data well.

Several methods have been developed for simplifying the model complexity, which are typically divided into constructive approaches [4, 5], and pruning methods [6, 7]. In the constructive approach, the structure of the network is incrementally built through adding nodes to the hidden layer one by one or group by group, while the pruning

L. Zhang, J. Kwok, and B.-L. Lu (Eds.): ISNN 2010, Part I, LNCS 6063, pp. 413–421, 2010.

approach starts with an initial selection of a large number of hidden units which is reduced as the algorithm proceeds.

In a recent publication, Billings [8] introduced the forward orthogonal least squares (OLS) algorithm for model construction. For a large class of linear-in-the-weights model, the orthogonal least square method has been known as an useful method for model complexity control and the ill-conditioning problems can also be solved effectively. To achieve a spare model with directly optimizing model generalization capability, a full automated procedure is proposed by using the PRESS statistic as a cost function for iterative model evaluation [9]. In [10], the OLS and a D-optimality criterion are used to determine the structure as for optimizing the model approximation ability, spare, and robust simultaneously. However, In the OLS algorithm, to select a candidate regressor, the vectors formed by the candidate regressors must be processed by using orthogonal methods, which is time consuming.

In the present paper, an orthogonal least squares based on Singular Value Decomposition for Spare Basis Selection is proposed (OLS-SVD). The new algorithm divides the candidate regressors into some sub-blocks to avoid comparing among the poor neurons and uses the forward orthogonal least squares algorithm based on the SVD approach to select the candidate regressors. The paper is organized as follows. Section 2 briefly reviews some primarily acknowledge on Linear-in-the-weights regression model. The OLS-SVD algorithm based on SVD and Press statistic is described in section 3. In section 4, two examples are simulated to illustrate the performance of the new algorithm. Finally, the conclusions of this paper are given in section 5.

2 Linear-in-the-Weights Regression Model

Consider a discrete nonlinear dynamical system of the form

$$y(k) = f\left(y(k-1),\ldots,y\left(k-n_y\right),u(k-1),\ldots,u\left(k-n_u\right)\right)+e(k) \qquad (1)$$

where $f(\cdot)$ is an unknown nonlinear mapping, $u(k)$ and $y(k)$ are the input and output variables of the system at discrete time step k, n_u and n_y represent the maximal orders in $u(k)$ and $y(k)$, respectively, while e(k) is assumed to be Gaussian noise with zero mean and unit variance. A *linear-in-weights* regression model of the form (2) can approximate $f(\cdot)$ with zero error.

$$\begin{aligned} y(k) &= \hat{y}(k)+e(k)=\sum_{i=1}^{n_M}\theta_i\Phi_i\left(x(k)\right)+e(k) \\ &= \boldsymbol{\Phi}^T(k)\boldsymbol{\theta}+e(k) \end{aligned},\ k=1,\ldots,N \qquad (2)$$

where N is the size of the observation data set, $\hat{y}(k)$ is the model output, θ_i are the model weights, $\boldsymbol{\theta} = [\theta_1,..., \theta_{nM}]^T$, $\Phi_i(x(k))$ are the regressors and $\boldsymbol{\Phi}(k) = [\Phi_1(x(k)),..., \Phi_{nM}(x(k))]^T$, $x(k)=[y(k-1),\ldots,y(k-n_y),u(k-1),...,u(k-n_u)]^T$ denotes the system input vector, and n_M is the total number candidate regressors.

The above N equations can be written compactly as

$$\mathbf{y} = \boldsymbol{\Phi}\boldsymbol{\theta}+\mathbf{e} \qquad (3)$$

where $\mathbf{\Phi} = [\boldsymbol{\Phi}(1),..., \boldsymbol{\Phi}(N)]^T$, where $\boldsymbol{\Phi}_i = [\Phi_i(1),..., \Phi_i(N)]^T$, $1 \le i \le n_M$, and defining $\mathbf{y} = [y(1),..., y(N)]^T$, $\mathbf{e} = [e(1),..., e(N)]^T$.

The forward selection algorithm is often used to construct a parsimonious model with a subset of $n_0 \ll n_M$ regressors by some model-selective criterion, among which the leave-one-out cross validation are metrics that measures a model's generalization capability.

Let $\{x(k), y(k)\}_{k=1:N,-k}$, be the resulting data set by removing the kth data point from the training data set $\{x(k), y(k)\}_{k=1:N}$, and denote estimated model output with j regressors as $\hat{y}_{j,-k}(k)$, and the related predicted residual at k as $\varepsilon_{j,-k}(k)$. For a linear-in-the-weights model with n_M candidate regressors, the PRESS errors are calculates as

$$\begin{aligned} \varepsilon_{n_M,-k}(k) &= y(k) - \hat{y}_{n_M,-k}(k) \\ &= \frac{\varepsilon_{n_M}(k)}{1 - \Phi^T(k)(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\Phi(k)} \end{aligned} \tag{4}$$

where $\varepsilon_{n_M}(k) = y(k) - \hat{y}_{n_M}(k)$

3 OLS_SVD Algorithm Based on SVD and Press Statistic

It appears that the computation burden of choosing the best subset model, which minimizes the mean square PRESS error $E[\varepsilon_{n_M,-k}(k)]$, will be expensive, as the matrix inversion involving. However, if employing an orthogonal forward regression to incrementally minimize PRESS error as presented in [11, 12], the model selection procedure would become computationally affordable.

Consider the linear-in-the-weights regression model (3), several OLS algorithms have been developed for selecting the candidate regressiors, such as Classical Gram-Schmidt (CGS) algorithm, Modified Gram-Schmidt (MGS) algorithm and Householder algorithm. The three methods have the same drawbacks to select candidate regressors with a forward selection procedure, as almost all the unjustified regressors need an orthogonal process with the predetermined ones before each optimal column is selected, and the computational complexity of the orthogonalisation procedure increases with the number of columns has been selected. Moreover, at each selection step, the PRESS statistic or other selective criterion with each unjustified column has to be formed for comparison. It appears that the computation burden increases as the growing of model size, and repeated comparing among the poor candidate regressors would also make the computation effort of the selection procedure extensive.

To overcome these drawbacks, an improved orthogonal least square algorithm based on SVD is proposed for spare model construction. The new algorithm divides the design matrix into several sub-blocks to avoid comparing among poor candidate regressors. And then SVD is applied to each sub-block, which would avoid the repeated orthogonalisation process by replacing the sub-design matrix with the obtained orthogonal singular matrix. Based on the model form of (3), the algorithm is showed as follows.

Firstly, dividing the design matrix $\mathbf{\Phi}_{all}$ into d parts equally by column, $\mathbf{\Phi}$ is a sub-block derived from $\mathbf{\Phi}_{all}$ with n_M (n_M<n_{Mall}) columns. Assuming the rank of $\mathbf{\Phi}$ is p in columns, and $\mathbf{\Phi}$ can be decomposed according to the Singular Vector Decomposition theorem as

$$\mathbf{\Phi} = \mathbf{U}\begin{bmatrix} \mathbf{\Sigma} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} \end{bmatrix}\mathbf{V}^H \tag{5}$$

where $\mathbf{\Sigma} = diag\left(\sigma_1, \sigma_2, \cdots, \sigma_p\right)$, and the diagonal elements of Σ are in the order $\sigma_1 \geq \sigma_2 \geq, \ldots, \geq \sigma_p$, $\mathbf{U}$ is an $N \times N$ orthogonal matrix consisting of all the orthogonalized eigenvectors associate with the eigenvalues of $\mathbf{\Phi}\mathbf{\Phi}^T$, and $\mathbf{V}$ is and $n_M \times n_M$ orthogonal matrix. Then only the first r eigenvectors (associated with the larger eigenvalues) are retained, while (p-r) smaller components are discarded, assuming that the latter describe mostly noise, and the r selected eigenvalues satisfy

$$\frac{\sum_{i=1}^{r}\sigma_i^2}{\sum_{j=1}^{p}\sigma_j^2} > \eta_0 \tag{6}$$

where η_0 is a user defined parameter, $0 < \eta_0 < 1$. According to the r eigenvalues, blocking the matrix $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{\Sigma}$ as

$$\mathbf{V} = [\mathbf{V}_1 \mid \mathbf{V}_2], \quad \mathbf{U} = [\mathbf{U}_1 \mid \mathbf{U}_2], \quad \mathbf{\Sigma} = [\mathbf{\Sigma}_1 \mid \mathbf{\Sigma}_2] \tag{7}$$

where $\mathbf{V}_1$ is a n_M by r Matrix, $\mathbf{V}_2$ is a M by (n_M-r) Matrix, $\mathbf{U}_1$ is a N by r Matrix, $\mathbf{U}_2$ is a N by (N-r) Matrix, and $\mathbf{\Sigma}$ is a r by r Matrix, $\mathbf{V}_2$ is a (p-r) by (p-r) Matrix.

Therefore, by neglecting the small singular values, which can be shown that mainly represent noise, Eq. (5) can be simplified by $\mathbf{\Phi}_r$, whose rank equals the number of remaining singular values.

$$\mathbf{\Phi}_r = \mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}_1^H \tag{8}$$

where $\mathbf{U}_1 = \left[\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_r\right]$.

Then, based on the approximation matrix $\mathbf{\Phi}_r$, the linear-in-the-weights regression model (3) can be expressed as

$$\mathbf{Y} = \mathbf{\Phi}\boldsymbol{\theta} + \mathbf{e} \approx \mathbf{\Phi}_r\boldsymbol{\theta} + \mathbf{e} = \mathbf{U}_1 \cdot \mathbf{\Sigma}_1\mathbf{V}_1^H \cdot \boldsymbol{\theta} + \mathbf{e} \tag{9}$$

Define $\mathbf{g} = \mathbf{\Sigma}\mathbf{V}_1^H \cdot \boldsymbol{\theta} = \left[g_1, g_2, \cdots, g_M\right]^T$, and Eq. (9) can be rewritten as

$$\mathbf{Y} = \mathbf{U}_1\mathbf{g} + \mathbf{e} \tag{10}$$

Eq. (8) is similar to the linear-in-the-weights regression model (3), but there are several significant differences between the two candidate regression matrix, $\mathbf{U}_1$ and $\mathbf{\Phi}$. First, compared to the matrix $\mathbf{\Phi}$, $\mathbf{U}_1$ is an orthogonal matrix, which means all the candidate regressors are already orthogonal with each other, and the repeated orthogonalisation decomposition of $\mathbf{\Phi}$ is avoided in the forward selection process. Second, by neglecting the small singular values in the matrix $\mathbf{\Sigma}$, $\mathbf{U}_1$ is an N by r orthogonal matrix consisting of only r orthogonalized eigenvectors associate with the r most significant eigenvalues of $\mathbf{\Phi}\mathbf{\Phi}^T$. Substituting $\mathbf{\Phi}$ with $\mathbf{U}_1$ would reduce the number of the candidate regressors,

and avoid the comparing among the poor regressors at each selection step, which is time consuming and complicated.

Then, the PRESS statistic can be embodied as a selective criterion for model construction, and the formulation is similar to the case given in [9]. For the improved orthogonal model (8) with r candidate regressors, the PRESS errors are calculates as

$$\begin{aligned}\varepsilon_{n,-k}(k) &= y(k) - \hat{y}_{n,-k}(k) \\ &= \frac{\varepsilon_n(k)}{1 - \mathbf{U}_1^T(k)(\mathbf{U}_1^T\mathbf{U}_1)^{-1}\mathbf{U}_1(k)} = \frac{\varepsilon_n(k)}{\eta_n(k)}\end{aligned} \tag{11}$$

where $\mathbf{U}_1$ is an orthogonal matrix.

To measure the generalization capability of a sub-model with n candidate regressors, the mean square PRESS error is the given by averaging all these PRESS errors.

$$J_n = E[\varepsilon_{n,-k}^2(k)] = E\left[\left(\frac{\varepsilon_n(k)}{\eta_n(k)}\right)^2\right] = \frac{1}{N}\frac{\varepsilon_n^2(k)}{\eta_n^2(k)} \tag{12}$$

Note that the model residual $\varepsilon_n(k)$ for the n-term model can be computed recursively as

$$\varepsilon_n(k) = y(k) - \sum_{i=1}^{n} u_i(k) g_i = \varepsilon_{n-1}(k) - u_n(k) g_n \tag{13}$$

where

$$g_i = \frac{\mathbf{u}_i^T \mathbf{y}}{\mathbf{u}_i^T \mathbf{u}_i} \tag{14}$$

And similarly, the PRESS error weighting $\eta_n(k)$ can be written in a recursive formula by

$$\eta_n(k) = 1 - \sum_{i=1}^{n} \frac{u_i^2(k)}{\mathbf{u}_i^T \mathbf{u}_i} = \eta_{n-1}(k) - \frac{u_n^2(k)}{\mathbf{u}_n^T \mathbf{u}_n} \tag{15}$$

Assume $\mathbf{\Phi}_1, \mathbf{\Phi}_2, \ldots, \mathbf{\Phi}_d$ are the d sub-blocks of the design matrix $\mathbf{\Phi}_{all}$ and $\tilde{\mathbf{U}}_i$ (i=1,2,…,d) is the ith subsets obtained from the selection process described above with each $\mathbf{\Phi}_i$, then the final basis function for model (3) is reconstructed as $\tilde{\mathbf{U}} = [\tilde{\mathbf{U}}_1, \tilde{\mathbf{U}}_2, ..., \tilde{\mathbf{U}}_d]$. If the size of $\tilde{\mathbf{U}}$ is still a large number, we can repeat this strategy till an appropriate model structure is achieved. It is worth to notice that d is a user defined parameter. The computation of selection procedure is simpler as the increase of d, for the candidate regressors in each sub-blocks is small. However, the computation burden of SVD needed is also increase as more subsets have to be considered. Given no a priori knowledge, there is no automated optimal way of choosing d, so in a practical scenario, we depend on the user's experiment and familiarity with size of data available to guide the choice of this parameter.

4 Simulations

The sparse modeling procedure described in the previous sections is applied to the simulated and benchmark data set, respectively.

4.1 Approximation of SinC Function with Noise

In this example, a RBF network with the Gaussian basis function is employed to approximate the SinC function, which is a popular choice to illustrate Support Vector Machine for regression (SVR) in the literature

$$y(x) = \begin{cases} \dfrac{\sin(x)}{x} & x \neq 0 \\ 1 & x = 0 \end{cases} \tag{16}$$

The Gaussian basis function employed is given by

$$\Phi_i(\mathbf{x}) = \exp\left(-\frac{\| \mathbf{x} - \mathbf{c}_i \|^2}{2\sigma^2}\right) \tag{17}$$

and $\mathbf{c}_i$ and σ are the centers and the widths of the basis functions, respectively.

The proposed algorithm is firstly compared with an OLS method based on and PRESS statistic to illustrate its efficiency in computation cost, in which the orthogonalisation process is achieved by employing a modified Gram-Schmidt (MGS) method. Moreover, for comparing the influence of the size of training set, 5 experiments are provided for each algorithm. In these experiments, the training set $\{x(k), y(k)\}$ is created from $y(x)$ where the input $x(k)$ is uniformly distribution on the interval (-10, 10). The number of the training sample is given in Table 1, which changes from 200 to 1000. In order to make the regression problem real, the Gaussian noise with zero mean and standard deviation 0.2 has been added to all the training samples. A testing set $\{x(k), y(k)\}$ is also created from $y(x)$ with two hundred noise free data, in which $x(k)$ is equally spaced in (-10, 10). The optimal kernel width is found to be σ^2=10 empirically, and the candidate center is taken as each training data point $x(k)$. Therefore, the number of regressors in each experiment is the same as that of the training set $\{x(k), y(k)\}$.

Table 1. Performance comparison of computation cost and accuracy for the simulated data set

Algorithm	Training time	Training MSE	Testing MSE	Training size	Number of sub-blocks
OLS with PRESS	2.9312	0.0429	0.0424	200	1
OLS-SVD *with* PRESS	0.2320	0.0408	0.0421	200	1
OLS with PRESS	12.0031	0.0287	0.0294	400	2
OLS-SVD *with* PRESS	0.8125	0.0290	0.0292	400	2
OLS with PRESS	26.5750	0.0331	0.0330	600	3
OLS-SVD *with* PRESS	2.1125	0.0305	0.0310	600	3
OLS with PRESS	47.7125	0.0235	0.0244	800	4
OLS-SVD *with* PRESS	3.9688	0.0237	0.0238	800	4
OLS with PRESS	73.5094	0.0172	0.0179	1000	5
OLS-SVD *with* PRESS	6.8024	0.0179	0.0185	1000	5

Table 1 compares the performance of the two OLS based method in the term of the computation cost and the modeling accuracy. As observed from Table 1, both the algorithms are comparable in accuracy. However, in the term of computation time, the OLS algorithm based on PRESS statistic is more time consuming for the repeated

orthogonalisation process in each selection step, and the training time increases dramatically with the number of the sample size. For the proposed algorithm, the model selection procedure is computation efficient by dividing the original design matrix into sub blocks and employing the SVD to simplify the orthogonalisation process. Compared with the OLS algorithm based on PRESS statistic, training time of the proposed algorithm increases slowly with the sample size.

4.2 Approximation of Nonlinear Dynamic Control System

Consider the following nonlinear dynamic control system

$$z(k)=\frac{z(k-1)z(k-2)z(k-3)u(k-2)(z(k-3)-1)+u(k-1)}{1+z^2(k-2)+z^2(k-3)} \tag{18}$$

A training set $\{x(k), y(k)\}$ and testing set $\{x(k), y(k)\}$ are created form (18) with 200 and 100 samples, respectively, where the system input u(k) is uniformly distribution on the interval [-1, 1]. In order to make the problem real, the Gaussian noise with zero mean and standard deviation 0.05 has been added to all the training and testing samples.

A RBF network with the thin-plate-spline basis function is employed to follow the dynamic process

$$\Phi_i\left(\mathbf{x}(k)\right)=\left\|\mathbf{x}(k)-\mathbf{c}_i\right\|^2\log\left(\left\|\mathbf{x}(k)-\mathbf{c}_i\right\|\right) \tag{19}$$

And the system input vector is denoted as

$$\mathbf{x}(k)=\left[\,y(k-1)\,y(k-2)\,y(k-3)u(k-1)u(k-2)\right]^T \tag{20}$$

As each training data point $\mathbf{x}(k)$ is considered as a candidate center of the network, there are 200 candidate regressors. In addition, for the small sample size, no dividing process is provided in this example.

To illustrate the generalization ability of the proposed algorithm, it is also compared with 5 other model construction methods, such as the LROLS algorithm with PRESS statistic, the OLS algorithm with PRESS statistic, the LROLS algorithm with MSE, the RVM algorithm, and the enhanced k-means clustering and least squares (CLS). All these algorithms have been employed in [9].

Table 2. Performance comparison of model size and accuracy for the nonlinear system

Algorithm	Validation set used	Model size	Training MSE	PRESS statistic	Testing MSE
OLS with PRESS [9]	No	51	0.002280	0.003864	0.005187
LROLS with PRESS [9]]	No	31	0.003192	0.003706	0.005892
LROLS with MSE[[9]	No	42	0.001883	0.003067	0.004872
RVM [9]	No	42	0.001598	0.002577	0.004935
CLS [9]	Yes	49	0.003940	0.007607	0.005580
OLS-SVD with PRESS	No	30	0.002527	0.003517	0.005210

Table 2 illustrate that, the six algorithms are comparable with each other in the term of model size and accuracy. In this example, the 200 candidate regressors are reduced to a size of 66 first, and a sparse model with 30 terms is achieved after the forward selection process. The model size of the proposed algorithm is comparable with the LROLS algorithm with PRESS statistic, while much smaller than the other four algorithms. However, although the LROLS with PRESS can construct a parsimonious model, its modeling accuracy is the worse, compared with the other five algorithms. For the proposed algorithm, as only the small singular values, which can be shown that mainly represent noise, are neglected in the process of SVD, the reconstructed candidate regressors in model (10) kept most of the useful information. Hence, the proposed algorithm can be acceptable in modeling accuracy. In addition, as same as the OLS with PRESS, the LROLS with PRESS and the RVM, the proposed algorithm is also an automated procedure without any requirement of a validation set.

5 Conclusions

In this paper, a new OLS algorithm based on SVD for linear-in-the-weights regression models is proposed. This is achieved by dividing the original candidate bases into several parts to avoid comparing among poor candidate regressors. The computation is further simplified by utilizing the Singular Value Decomposition to each sub-block. It can avoid the computation burden of the repeated orthogonalisation process before each optimal regressor is determined and further reduce the number of the candidate regressors. The results obtained from the examples which include the SinC function and a nonlinear dynamic control system demonstrate its effectiveness and accuracy.

Acknowledgements

This research is supported by the project (60674073) of the National Nature Science Foundation of China and the project (2007AA04Z158) of the National High Technology Research and Development Program of China (863 Program).

References

[1] Chen, S.M., Hwang, J.R.: Temperature prediction using fuzzy time series. IEEE Transactions on Systems, Man and Cybernetics -Part B 30(2), 263–275 (2000)
[2] Coulibaly, P., Anctil, F., Bobee, B.: Multivariate reservoir inflow forecasting using temporal neural networks. Journal of Hydrologic Engineering 6(5), 367–376 (2001)
[3] Han, M., Wang, Y.J.: Analysis and modeling of multivariate chaotic time series based on neural network. Expert System with Applications 36(2), 1280–1290 (2009)
[4] Feng, G., Huang, G.B., Lin, Q., Gay, R.: Error Minimized Extreme Learning Machine with Growth of Hidden Nodes and Incremental Learning. IEEE Transactions on Neural Networks 20(8), 1352–1357 (2009)
[5] Huang, G.B., Chen, L., Siew, C.K.: Universal Approximation Using Incremental Constructive Feedforward Networks with Random Hidden Nodes. IEEE Transactions on Neural Networks 17(4), 879–892 (2006)

[6] Salmerón, M., Ortega, J., Puntonet, C.G., Prieto, A.: Improved RAN sequential prediction using orthogonal techniques. Neurocomputing 41, 153–172 (2001)
[7] Rojas, I., Pomares, H., Bernier, J.L., Ortega, J., Pino, B., Pelayo, F.J., Prieto, A.: Time series analysis using normalized PG-RBF network with regression weights. Neurocomputing 42, 267–285 (2002)
[8] Billings, S.A., Wei, H.L.: A New Class of Wavelet Networks for Nonlinear System Identification. IEEE Transactions on Neural Networks 16(4), 862–874 (2005)
[9] Hong, X., Chen, S.: M-estimator and D-optimality model construction using orthogonal forward regression. IEEE Transactions on Systems, Man and Cybernetics, Part: B 35(1), 155–162 (2005)
[10] Chen, S., Hong, X., Harris, C.J., Sharkey, P.M.: Sparse Modeling Using Orthogonal Forward Regression with PRESS Statistic and Regularization. IEEE Transactions on Systems, Man and Cybernetics, Part: B 34(2), 898–911 (2004)