# Hybrid Wavelet Model Construction Using Orthogonal Forward Selection with Boosting Search

Meng Zhang[1], Jiaogen Zhou[2], Lihua Fu[3] and Tingting He[1]
*1 Department of Computer Science, Central China Normal University, 430079 Wuhan, China*
*2 Digital Engineering Center, Wuhan University, 430072 Wuhan, China.*
*3 School of Mathematics and Physics, Chinese University of Geosciences, 430079 Wuhan China*
*Chinazhouyg1997@yahoo.com.cn*

## Abstract

*This paper considers sparse regression modeling using a generalized kernel model in which each kernel regressor has its individually tuned center vector and diagonal covariance matrix. An orthogonal least squares forward selection procedure is employed to select the regressors one by one using a guided random search algorithm. In order to prevent the possible over-fitting, a practical method to select termination threshold is used. A novel hybrid wavelet is constructed to make the model sparser. The experimental results show that this generalized model outperforms traditional methods in terms of precision and sparseness. And the models with wavelet and hybrid kernel have a much faster convergence rate as compared to that with conventional RBF kernel.*

## 1. Introduction

Objective of modeling from data is not that a model should fit well to the training data. Rather, the goodness of a model is characterized by its generalization capability, and the model should be easy to interpret and to extract knowledge from. All these vital properties depend on crucially the ability of a modeling process to obtain appropriately sparse representations.

Some sparse kernel modeling techniques, such as the support vector machine (SVM) and linear programming (LP) [1-3], have become popular in data modeling applications.

---

This work was supported by the National Natural Science Foundation of China No.60442005, 60673040 and SRF for OYT, CUG-Wuhan under grant CUGQNL0520

However, when tackling some problems which are often encountered in science and engineering areas, it is unsuitable to use the conventional kernel methods.

For example, for non-flat function estimation problem, those methods adopt a single common variance for all kernel regressors and estimate both the steep and smooth variations using an unchanged scale.

Recently, a revised version of SVR, namely multi-scale support vector regression (MSSVR) [4, 5], is proposed by combining several feature spaces rather than a single feature space in standard SVR. The constructed multi-feature space is induced by a set of kernels with different scales. MSSVR outperforms traditional methods in terms of precision and sparseness, which will also be illuminated in our experiments.

Kernel basis pursuit (KBP) algorithm [6] is another possible solution which enables us to build a $l_1$ -regularized multiple-kernel estimator for regression. However, KBP is prone to over-fit the noisy data. We will compare its performance with our new algorithm.

Forward selection using the orthogonal least squares (OLS) algorithm [7–10] is a simple and efficient method that is capable of producing parsimonious linear-in-the-weights nonlinear models with excellent generalization performance. And orthogonal least squares regression (OLSR) is an efficient learning procedure for constructing sparse regression models [7-9]. A key feature of OLSR is its ability to select candidate model regressors with different scales and centers, which allows the produced model to fit different parts of original function with different scales. Some global searching algorithms, such as the genetic algorithm, adaptive simulated annealing and repeating weighted boosting search (RWBS), can be used to determine the parameters of regressor [9-11].

When applying OLSR, many researchers usually regard Gaussian function as the first choice for kernel function, for its good generalized ability. But sometimes real applications require the kernel function holds good local property to describe the local character of original function.

Wavelet techniques have shown promise for non-stationary function estimation [12, 13]. Since the local property of wavelet makes efficient the estimation of the function having local characters, it is valuable for us to study the combination of wavelet and OLSR. In order to obtain an even sparser model, this paper also constructs a novel hybrid wavelet as kernel function. This new kernel function is very flexible besides good local property, which composes left and right part of two mother wavelet with same center and different scale.

In this paper, multi-scale models with wavelet kernel and hybrid wavelet kernel are constructed by use of OLSR. OLSR algorithm used here tunes the dilation parameter and translation parameter of individual wavelet regressors by incrementally minimizing the training mean square error (MSE) using RWBS.

In modeling noisy dataset, OLSR can fit a function by any precision which is prone to cause over-fitting. So when the user should stop selecting regressors is also a problem. By virtue of cross validation, an algorithm to select termination threshold is presented in order to prevent possible over-fitting.

The simulations are performed on the function estimation problem of both artificial dataset and real dataset. The experimental results show that

1 The OLSR model outperforms traditional ones by precious and sparseness.

2 OLSRs with wavelet and hybrid wavelet kernel have much faster convergence than that with Gaussian kernel.

## 2. Theory

Consider the problem of fitting the $N$ pairs of training data $\{\mathbf{x}(l), y(l)\}_{l=1}^{N}$ with the regression model

$$y(l) = \hat{y}(l) + e(l) = \sum_{i=1}^{M} w_i \phi_i(l) + e(l), \quad l = 1, 2, \cdots, N \quad (1)$$

where $\hat{y}(l)$ denotes the "approximated" model output, $w_i$'s the model weights, $e(l)$ the modeling error at $\mathbf{x}(l)$ and $\phi_i(l) = k(\mathbf{c}(i), \mathbf{x}(l))$ are the regressors generated from a given kernel function $k(\cdot,\cdot)$ with center vector $\mathbf{c}(i)$. If we choose $k(\cdot,\cdot)$ as a Gaussian kernel and $\mathbf{c}(i) = \mathbf{x}(i)$, then model (1) describes a RBF network with each data as a RBF center and a fix RBF width. We are to find the best model $\sum_{i=1}^{M} w_i \phi_i(l)$ to describe the mapping $f(\mathbf{x})$ between the input $\mathbf{x}(l)$ and the output $y(l)$.

Let

$$\mathbf{\Phi}_i = [\phi_i(1), \cdots, \phi_i(N)]^T = [k(\mathbf{c}(i), \mathbf{x}(1)), \cdots, k(\mathbf{c}(i), \mathbf{x}(N))]^T, \quad i = 1, 2, \cdots, M,$$

and then the regression matrix $\mathbf{\Phi} = [\mathbf{\Phi}_1, \cdots, \mathbf{\Phi}_M]$, weight vector $\mathbf{w} = [w_1, \cdots, w_M]^T$, output vector $\mathbf{y} = [y(1), \cdots, y(N)]^T$, and error vector $\mathbf{e} = [e(1), \cdots, e(N)]^T$

Then the regression model (1) can be presented as following matrix form

$$\mathbf{y} = \mathbf{\Phi}\mathbf{w} + \mathbf{e} \quad (2)$$

The goal of modeling data is to find the best linear combination of the column of $\mathbf{\Phi}$ (i.e. the best value for $\mathbf{w}$) to explain $\mathbf{y}$ according to some criteria. The popular criteria is to minimize the sum of squared errors $E = \mathbf{e}^T\mathbf{e}$

By OLSR algorithm, the solution is searched in a transformed orthogonal space. In more detail, let an orthogonal decomposition of the regression matrix $\mathbf{\Phi}$ be $\mathbf{\Phi} = \mathbf{H}\mathbf{A}$, where $\mathbf{A}$ is an upper triangular matrix with the unit diagonal element and $\mathbf{H} = [\mathbf{H}_1, \mathbf{H}_2, \cdots, \mathbf{H}_M]$ with the orthogonal columns that satisfy $\mathbf{H}_i^T\mathbf{H}_j = 0$ if $i \neq j$. The regression model (2) can alternatively be expressed as

$$\mathbf{y} = \mathbf{H}\mathbf{\theta} + \mathbf{e} \quad (3)$$

where the new weight vector $\mathbf{\theta} = [\theta_1, \cdots, \theta_M]^T$ satisfies the triangular system $\mathbf{\theta} = \mathbf{A}\mathbf{w}$. Although the problem is converted to find the best solution in the linear space spanned by the column of $\mathbf{H}$ (i.e. the best value for $\mathbf{\theta}$), the resulting model remains equivalent to the solution of (2), which is still an element in the original space. For the orthogonal regression model (3), the training MSE can be expressed as

$$J = \mathbf{e}^T\mathbf{e}/N = \mathbf{y}^T\mathbf{y}/N - \sum_{i=1}^{M} \mathbf{H}_i^T\mathbf{H}_i\theta_i^2/N \quad (4)$$

Thus the training MSE for the $k$-term subset model can be expressed as $J_k = J_{k-1} - \mathbf{H}_i^T\mathbf{H}_i\theta_i^2/N$ with $J_0 = \mathbf{y}^T\mathbf{y}/N$

At the $k$ th stage of regression, the $k$ th regressor is determined by maximizing the error reduction criterion $E_k = \mathbf{H}_k^T\mathbf{H}_k\theta_k^2/N$ with respect to the kernel center $\mathbf{c}_k$ and its scale parameter $d_k$. The selection procedure is determined at $k$ th step if $J_k < \xi$ is satisfied.

Generally, RBF kernel is often the first choice of kernel because of its excellent generalized ability. Since the local property of wavelet makes efficient the estimation of the function having local characters, this

paper will also study the OLSR with wavelet kernel, and compare it with the case with Gaussian kernel.

Wavelet transform turns to be a useful tool in time series analysis and signal processing for its excellent localization property [12, 13]. The idea behind the wavelet analysis is to express or approximate a signal or function by a family of functions generated by dilations and translations of a function $h(x)$ called mother wavelet:

$$h_{c,d}(x) = |d|^{-1/2} h(\frac{x-c}{d}) \tag{5}$$

where $x, d, c \in R$, $d$ is a dilation factor, and $c$ is a translation parameter or center parameter. A multidimensional wavelet function can be written as the product of 1-d wavelet function $h(\mathbf{x}) = \prod_{i=1}^{N} h(x_i)$ with $\{\mathbf{x} = (x_1, \cdots, x_N) \in R^N\}$. In this paper, we use the same mother wavelet as in [14], that is $h(x) = \cos(1.75x)\exp(-x^2/2)$.

In order to obtain an even sparser model, this paper also constructs a novel hybrid wavelet as kernel function.

$$k(x) = \begin{cases} \cos[1.75(x-c)]\exp[\frac{1}{2}(\frac{x-c}{d^{(1)}})^2] & \text{if } x \le c \\ \cos[1.75(x-c)]\exp[\frac{1}{2}(\frac{x-c}{d^{(2)}})^2] & \text{if } x > c \end{cases} \tag{6}$$

Equation (6) shows the hybrid wavelet, which composes left and right part of two mother wavelet with same center and different scales. The center of kernel is denoted as $c$, left and right scales are denoted as $d^{(1)}$ and $d^{(2)}$ respectively.

## 3. Algorithm

Some guided random search methods can be used to determine the parameters of the $k$ th wavelet or hybrid wavelet regressor, that is $d_k$ and $\mathbf{c}_k$, such as the genetic algorithm and adaptive simulated annealing. RWBS is recently proposed global searching algorithm [11]. It is extremely simple and easy to implement, involving a minimum programming effort. So, we perform this optimization by RWBS.

Let the vector $\mathbf{u}_k$ contain both center parameters and scale parameters of $k$ th regressor, that is $\mathbf{u}_k = [d_k, \mathbf{c}_k]^T$. Given the data $\{\mathbf{x}(l), y(l)\}_{l=1}^{N}$, and randomly selecting $Ps$ parameter vectors $\{\mathbf{u}_i \mid i = 1, \cdots, Ps\}$, the basic weighted boosting search algorithm can be implemented as [9,11].

In Fig.1, Condition 1 means that the local minimums obtained at two continuous steps is close enough, that is $\|\tilde{\mathbf{u}}_t - \tilde{\mathbf{u}}_{t+1}\| < \varsigma$. Condition 2 means that the iteration number reaches the threshold $Nb$. The method of searching local minimum of $J$ can refer to [11].

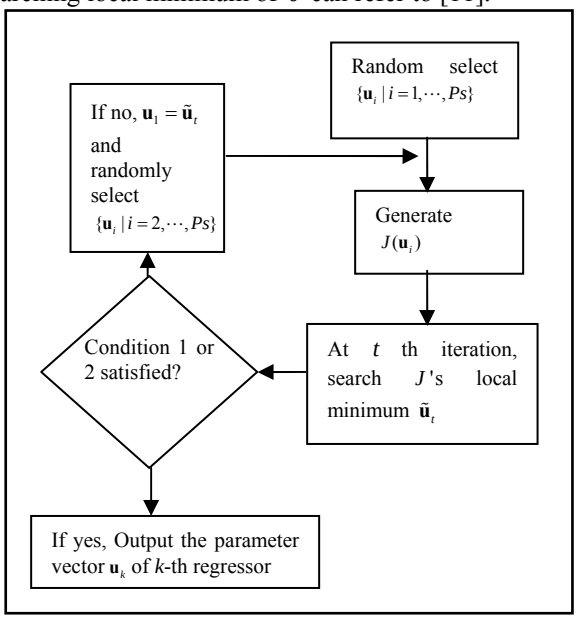



**Fig. 1.** The scheme of basic weighted boosting search algorithm

The cost function $J(\mathbf{u}_i)$ is generated according to the following steps:

*Step 1* for $1 \le i \le Ps$, generate $\mathbf{\Phi}_i$ from $\mathbf{u}_i$, the candidates for the $k$-th model column,

*Step 2* Orthogonalise $\mathbf{\Phi}_i$:

$$\alpha_j^i = \mathbf{H}_j^T \mathbf{\Phi}_i / (\mathbf{H}_j^T \mathbf{H}_j), \quad 1 \le j < k \quad \tilde{\mathbf{H}}_i = \mathbf{\Phi}_i - \sum_{j=1}^{k-1} \alpha_j^i \mathbf{H}_j$$

where $\{\mathbf{H}_j \mid j = 1, \cdots, k-1\}$ denote the already-selected regressors of equation (3) while $\{\tilde{\mathbf{H}}_i \mid i = 1, \cdots, Ps\}$ mean the candidates for the $k$th regressor.

*Step 3* Generate $J(\mathbf{u}_i)$

$$\gamma_i = (\tilde{\mathbf{H}}_i)^T \tilde{\mathbf{H}}_i, \ \theta_i = (\tilde{\mathbf{H}}_i)^T \mathbf{y} / \gamma_i, \text{ and}$$

$$J(\mathbf{u}_i) = J_{k-1} - \gamma_i (\theta_i)^2 / N$$

with $J_{k-1}$ refers to the training MSE for the $k-1$-term subset model.

The above basic weighted boosting search algorithm performs a guided random search and solution obtained may depend on the initial choice of the population. To derive a robust algorithm that ensures a stable and global solution, RWBS algorithm is used by applying the basic weighted boosting search for $NG$ times.

Using RWBS, one can obtain the best dilation and translation factors of the $k$th wavelet regressor.

*Remark 1.* To guarantee a global optimal solution as well as to achieve a fast convergence, the algorithmic

parameters, $NG$, $Nb$ , $Ps$ and $\varsigma$ , need to be set carefully. The appropriate values of these parameters depend on the dimension of $\mathbf{u}$ and how hard the objective functions to be optimized. In this paper, in order to assure a global optimal solution, the thresholds $NG$, $Nb$ and the size of generation size $Ps$ are assigned to a little larger than needed.

In theory, this procedure can generate a model, by any precision, approximating the original mapping $f(\mathbf{x})$ between input $\mathbf{x}(l)$ and output $y(l)$. It will cause over-fitting in noisy setting. So it is necessary to preset a threshold $\xi$ and if the condition $J_k < \xi$ is satisfied, we can stop the regressor selecting procedure before the model is fitted into the noise.

The procedure to generate the whole regression model can be described as:

For n=1:N

Repeated Basic weighted boosting search

If $J_n > J_{n-1}$ or If $J_n \le \xi$

Break

End if

End for

Here, the largest iteration number $N$ can be designed as the size of the training set. Usually the procedure will be ended at $n$-th when any of the two termination conditions satisfied, that is $J_n > J_{n-1}$ and $J_n \le \xi$ with $n << N$ .

## 4. Simulations

To demonstrate the practicability and the good performance of OLSR with wavelet and hybrid wavelet kernel, we applied to some simulated data and real data. Both RBF kernels and wavelet kernel were used in our experiment. In all experiments following, we assign $Ps = 10$ $NG = 50$, $Nb = 9$ and $\varsigma = 0.02$

The example [4] is a highly oscillating function $r(x) = \sin(11\pi/(0.35x+1))$ in the fix domain $x \in [0,10]$. We generated 30 training sets of size $l = 100$ from this function by adding an independent Gaussian noise $n_i \sim N(0, 0.1^2)$ too.

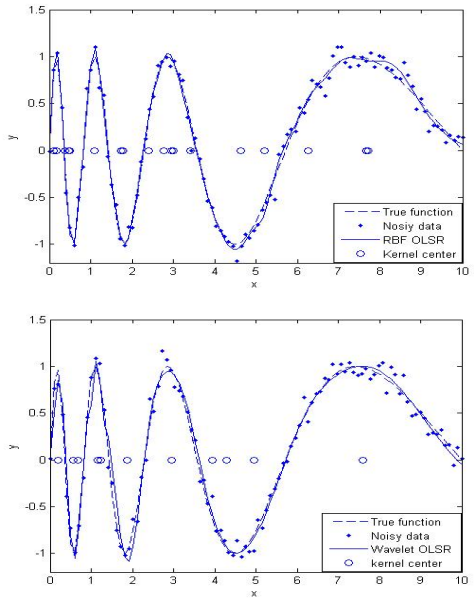
As the method described in Subsection 3.2, one can readily select the threshold $\xi = 2^{-7}$ . Table 1 shows the comparison between OLSR and conventional kernel methods. One can safely draw a conclusion that the OLSR model outperforms traditional ones by precious and sparseness.

**Table 1.** The averaged experimental results for experiment 1. The first 7 results were quoted from [4]

| | Model Size | RMSE |
|---|---|---|
| SVR | 61.5 | $0.0637 \pm 0.0100$ |
| LPR | 33.6 | $0.0634 \pm 0.0098$ |
| LS-SVM | 100.0 | $0.0613 \pm 0.0075$ |
| KBP | 35.5 | $0.0608 \pm 0.0070$ |
| MS-SVR(E) | 27.0 | $0.0581 \pm 0.0087$ |
| MS-SVR(Q) | 25.4 | $0.0515 \pm 0.0072$ |
| MS-SVR(H) | 28.4 | $0.0516 \pm 0.0073$ |
| RBF OLSR | 18.5 | $0.0259 \pm 0.0052$ |
| Wavelet OLSR | 11.4 | $0.0204 \pm 0.0062$ |
| Hybrid wavelet OLSR | 9.0 | $0.0169 \pm 0.0062$ |

Fig.2 shows the performance the OLSR with RBF (above), wavelet (middle) and hybrid wavelet kernel (bottom) respectively. Input noisy data and kernel centers are shown by 'dot' and 'circle' respectively. The true oscillating function $r(x)$ and the regression model are denoted by dashed line and solid line respectively. One can find that OLSR with hybrid wavelet approximates the function by using the smallest model (or by using the fewest kernel centers). On the contrary, OLSR with RBF kernel has the largest model.

Fig.3 shows that OLSR with hybrid wavelet kernel has the fastest convergence rate. As the algorithm mentioned in Subsection 3.2, the procedure was ended at $n$-th when the condition $J_n \le \xi$ is satisfied. Hybrid wavelet, wavelet and RBF OLSRs reach the condition at $8^{th}$, $10^{th}$,and $18^{th}$ respectively.



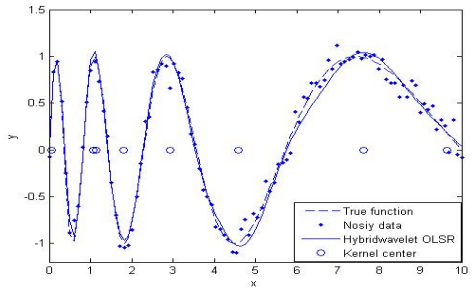



**Fig. 2.** Estimators of the oscillating function, the above is for OLSR with RBF kernel, the middle for wavelet kernel and the bottom is for hybrid wavelet kernel.

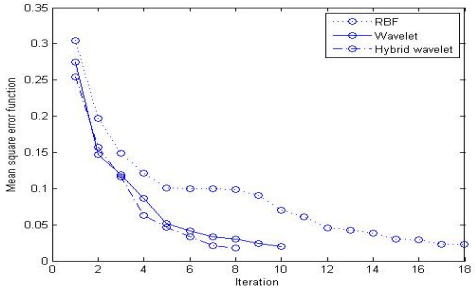



**Fig. 3.** The training error of OLSR with different kernels at each step

## 5. Conclusions

In order to construct a sparse model, this paper proposes a novel multi-scale model by orthogonal least squares regression (OLSR) with wavelet kernel and a new hybrid wavelet kernel. Unlike most of the other kernel method, the wavelet and hybrid wavelet kernel's centers are not restricted to the training input data points. Due to the flexibility of OLSR and good local prosperity of wavelet kernel, this new approach outperforms some traditional methods in our simulation experiments. Both OLSR with hybrid wavelet kernel and with OLSR with hybrid wavelet kernel have a much faster convergence than that with traditional RBF kernel.

## References


1. Smola, A.: Regression Estimation with Support Vector Learning Machines, Master's Thesis, Technische University Műnchen, (1996) Available at (http:// www.kernel-machines.org)
2. Suykens, J.A.K., Vandewalle J.: Least Squares Support Vector Machine Classifiers, Neural Process. Lett. Vol.9 (1999) 293–300.
3. Smola, A., Schőlkopf, B., Rätsch, G.: Linear programs for automatic accuracy control in regression, Proceedings of the Ninth International Conference on Artificial Neural Networks, London, (1999) 575–580.
4. Zheng, D., Wang, J., Zhao Y.: Non-flat Function Estimation with A Multi-scale Support Vector Regression, Neurocomputing, in press
5. Zheng, D., Wang, J., Zhao Y.: Training Sparse MS-SVR with an Expectation-Maximization Algorithm, Neurocomputing, Vol.69 (2006) 1659-1664
6. Guigue, V., Rakotomamonjy, A., Canu, S.: Kernel Basis Pursuit, Proceedings of the 16th European Conference on Machine Learning, Porto, (2005)
7. Chen, S., Billings, S.A., Luo, W.: Orthogonal Least Squares Methods and Their Application to Non-linear System Identification, Int. J. Control Vol.50 (1989) 1873–1896.
8. Chen, S., Cowan, C.F.N., Grant, P.M.: Orthogonal Least Squares Learning Algorithm for Radial Basis Function Networks, IEEE Trans. Neural Networks Vol. 2 (1991) 302–309.
9. Chen, S. Wang, X. X. Brown, D.J.: Orthogonal Least Squares Regression with Tunable Kernels, Electronics Letters Vol 41 No 8 (2005)
10. Chen, S. 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, No.5, (1999) 1239-1243
11. Chen, S., Wang, X.X. Harris, C.J.: Experiments with repeating weighted boosting search for optimization in signal processing applications, IEEE Trans. Syst. Man Cybern. B, Cybern., Vol.35, No.4, (2005) 682-693
12. Mallat, S.: A Wavelet Tour of Signal Processing. Academic Press, (1999).
13. Daubechies,: I. Ten Lectures on Wavelets. CBMS, 61, SIAM, Philadelphia, (1992).
14. Zhang, L., Zhou, W., Jiao, L.: Wavelet Support Vector Machine. IEEE Trans. on System, Man and Cybernetics-Part B: Cybernetics, Vol. 34(2004), 34−39.
15. Dubois, R., Quenet, B., Faisandier, Y. Dreyfus, G. :Building Meaningful Representations for Nonlinear Modeling of 1d and 2d-signals: Applications to Biomedical Signals, Neurocomputing, Vol. 69 (2006) 2180-2192
16. Amitava, C. Amine, N. A. Patrick, S.: An Imput-Delay Neural-Network-Based Approach for Piecewise ECG Signal Compression, IEEE Trans. Biomedical Engineering, Vol.52 No. 5 (2005) 945-947
17. Rodrigo, V. A., Bernadette, D., Jerome, B.: ECG Signal Analysis through Hidden Markov Models, IEEE Trans. Biomedical Engineering, Vol. 53, No. 8 (2006), 1541-1549
18. Shyu, L.Y., Wu, Y. H., Hu, W.: Using Wavelet Transform and Fuzzy Neural Network for VPC Detection from the Hotlter ECG, IEEE Trans. Biomedical Engineering, Vol.51, No. 7, (2004), 1269-1272
19. MIT-BIH Arrhythmia Database [Online]. Available: http://www.physionet.org/physio-bank/database/mitda/