Statistica Sinica **9**(1999), 167-182

SUBSET-SELECTION AND ENSEMBLE METHODS FOR WAVELET DE-NOISING

Andrew G. Bruce, Hong-Ye Gao and Werner Stuetzle

MathSoft Inc., TeraLogic Inc. and University of Washington

Abstract: Many nonparametric regression procedures are based on "subset selection": they choose a subset of carriers from a large or even infinite set, and then determine the coefficients of the chosen carriers by least squares. Procedures which can be cast in this framework include Projection Pursuit, Turbo, Mars, and Matching Pursuit. Recently, considerable attention has been given to "ensemble estimators" which combine least squares estimates obtained from multiple subsets of carriers. In the parametric regression setting, such ensemble estimators have been shown to improve on the accuracy of subset selection procedures in some situations. In this paper we compare subset selection estimators and ensemble estimators in the context of wavelet de-noising. We present simulation results demonstrating that a certain class of ensemble wavelet estimators, based on the concept of "cycle spinning", are significantly more accurate than subset selection methods. These advantages hold even when the subset selection procedures can rely on an oracle to select the optimal number of carriers. We compute ideal thresholds for translation invariant wavelet shrinkage and investigate other approaches to ensemble wavelet estimation.

Key words and phrases: Cycle spinning, model combination, nonparametric regression, stepwise regression, wavelet shrinkage.

1. Introduction

Regression is one of the fundamental problems of statistics. The goal of regression analysis is to estimate the conditional expectation $E(Y \mid \mathbf{X} = \mathbf{x})$ of a response variable Y, given the values of predictor variables $\mathbf{X} = (X_1, \ldots, X_p)$. The estimate is based on a training sample $(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)$ of observations for which the values of response and predictors are known.

In the last 20 years there has been a large amount of research on nonparametric regression procedures that do not assume knowledge of the functional form of $E(Y \mid \mathbf{X} = \mathbf{x})$. A typical example for a nonparametric regression procedure is *Turbo* (Friedman and Silverman (1989)). For our purposes it is sufficient to describe it for the case of a single predictor variable ($p = 1$). Turbo assumes that $E(Y \mid X)$ can be well approximated by a 2nd order (piecewise linear) spline function. It represents the estimate $\hat{f}(x)$ as a linear combination of basis functions

$\phi_j(x) = (x - t_j)_+$ for a fixed set of locations $\{t_j\}$:

$$\hat{f}(x) = b_0 + \sum_{j=1}^{m} b_j \phi_j(x). \tag{1}$$

The key issues are choosing the basis functions to be included in the model, and determining the corresponding coefficients. Turbo addresses these issues in a way typical for many nonparametric regression methods. It constructs a collection of models with different numbers of basis functions using forward/backward stepwise regression. Out of this collection it chooses the model minimizing an estimate of prediction error, such as cross-validated residual sum of squares. The basic characteristic of Turbo is that the coefficients for a given model are estimated by least squares. We refer to nonparametric regression methods that (i) select a subset of basis functions from a large or even infinite pool, and (ii) determine the corresponding coefficients by least squares, as *carrier selection methods*.

Besides Turbo, examples of carrier selection methods include multidimensional additive spline approximation (Friedman, Grosse and Stuetzle (1983)), Mars (Friedman (1991)), radial basis function networks (Chen, Cowan and Grant (1991)), and fuzzy basis function expansions (Wang and Mendel (1992)). These methods differ primarily in the pool of basis functions.

When the model (1) involves a linear combination of more than a few carriers, carrier selection techniques have been shown to be unstable (Breiman (1994b)), and the corresponding estimates have high variance. One way to deal with this instability is to use an *ensemble estimate*. Instead of attempting to find a single best subset of carriers, ensemble estimates combine several carrier selection estimates. The "bagging" procedure of Breiman (1994a) and the "bumping" procedure of Tibshirani and Knight (1995) use the bootstrap to form multiple estimates. They demonstrate that the combined estimator based on the bootstrap samples improves accuracy in certain situations.

In this paper, we compare carrier selection estimates and ensemble estimates in the context of wavelet de-noising. Suppose we observe data $\mathbf{y} = (y_1, \ldots, y_n)^t$ assumed to be generated from the model

$$y_i = f(i) + \sigma z_i \qquad i = 1, \ldots, n, \tag{2}$$

where $f(i)$ is a deterministic function sampled at equally spaced points and $\{z_i\}$ are i.i.d. $N(0,\ 1)$. Our goal is to estimate $\mathbf{f} = [f(1), \ldots, f(n)]^t$ with small mean-square-error

$$R(\hat{\mathbf{f}},\ \mathbf{f}) = \frac{1}{n} \sum_{i=1}^{n} E(\hat{f}(i) - f(i))^2. \tag{3}$$

To avoid unnecessary detail, we restrict our attention to sample sizes $n = 2^L$ for some integer L. The results readily extend to arbitrary sample sizes.

Donoho and Johnstone ((1994) and (1995)) developed a theory and methodology based on "wavelet shrinkage." Their procedure takes the orthogonal wavelet transform of the data, shrinks the coefficients towards zero, and inverts the shrunken coefficients to obtain an estimate $\hat{\mathbf{f}}$. They show that wavelet shrinkage has very broad asymptotic near-optimality properties. For example, wavelet shrinkage achieves the minimax risk over each function class in a variety of smoothness classes and with respect to a variety of losses, including L_2 risk.

Since the introduction of the original Donoho and Johnstone wavelet shrinkage, several alternative methods have been proposed, including "matching pursuit" and "translation invariant wavelet shrinkage". In contrast to the original Donoho and Johnstone wavelet shrinkage procedure, these methods work with a much larger pool of non-orthogonal basis functions. Matching pursuit, developed by Mallat and Zhang (1993) and Qian and Chen (1994), is a carrier selection method similar in spirit to forward stepwise regression. Translation invariant wavelet shrinkage (TIWS), developed by Coifman and Donoho (1995), is an ensemble estimate in that it averages estimates obtained by applying the following operations for some collection of integers k:

1. *Shift*: Shift the original data series by $k\Delta$.
2. *Predict*: Apply orthogonal wavelet shrinkage to the shifted series.
3. *Unshift*: Unshift the result by $-k\Delta$.

This process of shift–predict–unshift is called "cycle spinning." Since the orthogonal wavelet transform is not translation invariant, different shifts $k\Delta$ will yield different estimates. Coifman and Donoho (1995) show the average of the estimates from all the shifts, in many cases, achieves better performance than the estimate from any single shift.

In this paper we present the results of a Monte Carlo simulation study suggesting that TIWS generally has lower mean squared error than matching pursuit and related carrier selection methods. These findings are consistent with the results cited above: carrier selection methods are unstable and ensemble estimators improve on their accuracy. The performance advantage of TIWS holds even when the carrier selection methods rely on an oracle to select the optimal number of carriers. Moreover, TIWS is computationally much faster than the various carrier selection procedures.

We also further develop cycle spinning as an estimation method, extending the work of Coifman and Donoho (1995). We compute ideal shrinkage thresholds for TIWS and investigate optimally weighted ensemble estimates.

In Section 2 we review the methods studied in this paper: TIWS, matching pursuit and other carrier selection methods. The comparison of TIWS to the various carrier selection methods is presented in Section 3. In Section 4 we give ideal thresholds for TIWS. In Section 5 we investigate optimally weighted ensemble estimates. Section 6 summarizes the results and discusses related work. Appendix A gives instructions on how to obtain the software used to conduct the study.

2. Background

2.1. Translation invariant wavelet shrinkage

Let $\mathbf{W}$ be an orthogonal transform matrix corresponding to a set of orthogonal wavelet basis functions $\{\phi_l(t)\}$. Donoho and Johnstone ((1994) and (1995)) propose the wavelet shrinkage estimator $\hat{\mathbf{f}}_{WS} = \mathbf{W}^t \delta(\mathbf{W}\mathbf{y})$ where δ is a shrinkage function. Two common shrinkage functions are hard shrinkage $\delta^H_\lambda(x) = xI_{[|x|>\lambda]}$ and soft shrinkage $\delta^S_\lambda(x) = \text{sgn}(x)(|x| - \lambda)_+$. The threshold λ determines the amount of shrinkage; $\lambda = \infty$ sets the coefficients to zero. The wavelet transform applied to a sample of size $n = 2^L$ has a maximum of L resolution levels. Shrinkage is typically applied to only the detail or high frequency resolution levels. For example, Bruce and Gao (1996) indicate that applying shrinkage to $J = \log_2 n - 4$ resolution levels gives good performance in a range of situations.

The columns of the matrix $\mathbf{W}^t$ correspond to wavelet basis functions, and the wavelet shrinkage estimator $\hat{\mathbf{f}}_{WS}$ can be viewed as a linear combination of basis vectors. In fact, the wavelet shrinkage estimator with the hard shrinkage function is a carrier selection method where the carrier matrix $\mathbf{X} \equiv \mathbf{W}^t$.

A potential drawback with $\hat{\mathbf{f}}_{WS}$ as an estimator is the lack of translation invariance of the wavelet transform. To overcome this translation invariance, we can apply cycle spinning. This leads to the translation invariant wavelet shrinkage estimator (TIWS) defined by

$$\hat{\mathbf{f}}_{TI} = \frac{1}{2^J} \sum_{k=0}^{2^J-1} (\mathbf{W}\mathbf{S}_k)^t \delta(\mathbf{W}\mathbf{S}_k\mathbf{y}), \tag{4}$$

where J is the number of levels to which shrinkage is applied. Here $\mathbf{S}_k$ is the discrete $n \times n$ translation spin (shift) operator

$$\mathbf{S}_k = \begin{pmatrix} \mathbf{0}_{k\times(n-k)} & \mathbf{I}_{k\times k} \\ \mathbf{I}_{(n-k)\times(n-k)} & \mathbf{0}_{(n-k)\times k} \end{pmatrix}, \tag{5}$$

where $\mathbf{I}_{m\times m}$ is the identity matrix with m rows and columns and $\mathbf{0}_{r\times c}$ is the zero matrix with r rows and c columns.

2.2. Matching pursuit and other carrier selection methods

It can be shown that the estimator (4) is equivalent to $\hat{\mathbf{f}}_{TI} = \mathbf{W}_{TI}^{\#}\delta\left(\mathbf{W}_{TI}\mathbf{y}\right)$ where $\mathbf{W}_{TI}$ is the non-decimated or stationary wavelet transform and $\mathbf{W}_{TI}^{\#} \equiv \left(\mathbf{W}_{TI}^{t}\mathbf{W}_{TI}\right)^{-1}\mathbf{W}_{TI}^{t}$ is the generalized inverse of $\mathbf{W}_{TI}$. The non-decimated wavelet transform is an over-sampled wavelet transform corresponding to wavelet basis functions at all integer shifts of the original series: see Shensa (1992) and Nason and Silverman (1995). This suggests the following alternative to the TIWS estimator (4): apply a carrier selection method to the carrier set given by $\mathbf{X}_{TI} \equiv \mathbf{W}_{TI}^{t}$. The columns of $\mathbf{X}_{TI}$ correspond to shifted wavelet basis functions, and include as a subset the columns of the orthogonal wavelet carrier matrix $\mathbf{W}^{t}$. We study three variations of the carrier selection method: the original matching pursuit algorithm proposed by Mallat and Zhang (1993) and Qian and Chen (1994), forward stepwise regression, and forward-backward stepwise regression. All these methods use the same basic algorithm.

(1) Use a greedy selection method to form p nested subsets of carriers $\mathbf{X}_m = (\mathbf{x}_{i_1}, \mathbf{x}_{i_2}, \ldots, \mathbf{x}_{i_m})$, $m = 1, \cdots, p$, where $p \leq n$.
(2) For each subset, determine the regression coefficients by least squares: $\hat{\mathbf{b}}_m = (\mathbf{X}_m^{t}\mathbf{X}_m)^{-1}\mathbf{X}_m^{t}\mathbf{y}$
(3) Choose the model $\mathbf{X}_m$ that minimizes an error estimate such as the cross-validated residual sum-of-squares.

The procedures differ in the technique used to order the carriers in step (1). Matching pursuit orders the carriers by a forward selection procedure as follows:

(1.0) Initialize $j = 1$ and $\mathbf{r}_j = \mathbf{y}$.
(1.1) Find the carrier $\mathbf{x}_{i_j}$ maximally correlated with $\mathbf{r}_j$: $i_j = \arg\max_i \frac{\langle \mathbf{x}_i, \mathbf{r}_j \rangle^2}{\|\mathbf{x}_i\|^2}$.
(1.2) Update the residual vector $\mathbf{r}_{j+1} = \mathbf{r}_j - \frac{\langle \mathbf{x}_{i_j}, \mathbf{r}_j \rangle}{\|\mathbf{x}_{i_j}\|^2}\mathbf{x}_{i_j}$.
(1.3) If $j < p$, then set $j = j + 1$ and go to step (1.1).

Mallat and Zhang (1993) describe fast implementations for a variety of carrier sets (e.g., wavelets, wavelet packets, Gabor functions) based on wavelet pyramid filtering algorithms and the fast Fourier transform.

Forward stepwise regression is the same as matching pursuit except that $\mathbf{x}_{i_j}$ is orthogonalized out of the remaining carriers at each iteration. Pati, Rezaiifar and Krishnaprasad (1993) applied forward stepwise regression in the context of wavelet de-noising under the name "orthogonal matching pursuit", and derived fast algorithms for updating the inner products and vector norms used in forward stepwise regression. Orthogonal matching pursuit, however, is computationally slower than matching pursuit.

Forward-backward stepwise regression is forward stepwise regression followed by backwards elimination. This is the method used in Turbo (Friedman and Silverman (1989)).

2.3. Choice of smoothing parameters

Each of the methods has a parameter controlling the amount of smoothing. In wavelet shrinkage, this is the threshold λ below which wavelets coefficients are set to zero; in the carrier selection methods, it is the number of carriers m retained in step (3) of the basic algorithm. The problem of choosing the smoothing parameters is essentially same for all methods, and not the primary focus of this paper, which is to compare strategies for selecting subsets. Hence, we determine the smoothing parameter by an oracle. In the case of wavelet shrinkage, the optimal threshold λ_{opt} is given by

$$\lambda_{\text{opt}} = \arg\min_{\lambda} \|\hat{\mathbf{f}}_{\lambda} - \mathbf{f}\|_2^2. \tag{6}$$

For carrier selection methods, the optimal number of carriers m_{opt} is

$$m_{\text{opt}} = \arg\min_{m} \|\mathbf{X}_m \hat{\mathbf{b}}_m - \mathbf{f}\|_2^2 \qquad m = 1, \ldots, p. \tag{7}$$

3. Comparision of TIWS with Carrier Selection Methods

We compare translation invariant wavelet shrinkage (TIWS) with the original Donoho and Johnstone wavelet shrinkage and with the three carrier selection methods described in Section 2: matching pursuit, forward stepwise regression, and forward-backward stepwise regression. For all examples presented in this section, we use the least asymmetric "s8" wavelet (Daubechies (1992), Table 6.3, $N = 4$).

The wavelet shrinkage estimates are computed using the hard shrinkage function δ_{λ}^{H} applied to the detail/high frequency coefficients at the finest $J = \log_2 n - 4$ resolution levels in the wavelet decomposition. In other words, shrinkage is not applied to the 32 lowest frequency coefficients. The carrier selection methods are applied to the non-decimated wavelet carrier matrix $\mathbf{X}_{TI} = \mathbf{W}_{TI}^{t}$ with a maximum of $p = n/4 = 64$ carriers. The smoothing parameters are computed by an oracle as described in Section 2.3. In addition, we compute wavelet shrinkage estimates using the Donoho and Johnstone "universal" threshold of $\lambda = \sigma\sqrt{2\log n}$ where σ is assumed known.

We compare the methods on the four functions "blocks", "bumps", "doppler" and "heavisine". These functions were introduced by Donoho and Johnstone (1994) to provide a diverse set of examples for spatially inhomogeneous behavior,

and have since been used to evaluate the performance of nonparametric procedures in several studies (see, for example, Fan and Gijbels (1995) and Bruce and Gao (1996)). We use a sample size of $n = 256$ and root signal-to-noise ratios of 3, 5, and 7. The root signal-to-noise ratio is defined as $\text{RSNR} = (\text{Var}(\mathbf{f}))^{1/2}/\sigma$. We set $\sigma = 1$ and adjust the scale of the function accordingly.

Table 1. Variance and mean-square-error for translation invariant wavelet shrinkage with an oracle-driven threshold (TIWS) and the universal threshold (TIWS-UNI), wavelet shrinkage using the orthogonal wavelet transform with an oracle-driven threshold (WS) and the universal threshold (WS-UNI), matching pursuit (MP), forward stepwise regression (Stepwise), and forward-backward stepwise regression (Forw-Back). The carrier selection methods use an oracle to determine the number of carriers. The values in this table are based on $M = 100$ simulations. The standard error for each value is given in parentheses. See Section 3.1 for details.

RSNR=5	Blacks	Bumps	Doppler	Heavisine
	Variance			
TIWS	0.309 (0.006)	0.389 (0.006)	0.201 (0.004)	0.092 (0.003)
TIWS-UNI	0.310 (0.006)	0.383 (0.006)	0.200 (0.004)	0.083 (0.003)
WS	0.502 (0.009)	0.550 (0.009)	0.269 (0.007)	0.131 (0.005)
WS-UNI	0.541 (0.010)	0.547 (0.011)	0.239 (0.007)	0.120 (0.004)
MP	0.639 (0.009)	0.660 (0.011)	0.493 (0.008)	0.284 (0.008)
Stepwise	0.637 (0.009)	0.613 (0.010)	0.479 (0.009)	0.298 (0.008)
Forw-Back	0.643 (0.009)	0.613 (0.010)	0.485 (0.009)	0.302 (0.008)
	MSE			
TIWS	0.360 (0.006)	0.440 (0.006)	0.252 (0.004)	0.138 (0.004)
TIWS-UNI	0.419 (0.007)	0.498 (0.006)	0.271 (0.005)	0.149 (0.004)
WS	0.588 (0.008)	0.631 (0.008)	0.367 (0.006)	0.190 (0.004)
WS-UNI	0.821 (0.010)	0.870 (0.012)	0.414 (0.007)	0.203 (0.005)
MP	0.692 (0.009)	0.729 (0.010)	0.571 (0.007)	0.328 (0.008)
Stepwise	0.680 (0.009)	0.679 (0.009)	0.539 (0.008)	0.233 (0.007)
Forw-Back	0.683 (0.009)	0.677 (0.009)	0.544 (0.008)	0.339 (0.007)

3.1. Simulation results

The variance and mean square error (MSE) of the methods for root signal-to-noise ratio RSNR = 5 are listed in Table 1. Results for RSNR = 3 and RSNR = 7 are similar; they can be found in Bruce, Gao and Stuetzle (1996).

From the table, we can draw the following conclusions:

1. Both universal and oracle-guided TIWS dominate the carrier selection methods in terms of MSE and variance for all functions examined. The MSE for carrier selection methods is 60%–170% higher than the MSE for oracle-guided

TIWS. Universal TIWS inflates the MSE by only about 10%–15% over oracle-guided TIWS.

2. The carrier selection methods are all roughly comparable, with matching pursuit performing slightly worse in terms of MSE. This is to be expected since matching pursuit does not orthogonalize the remaining carriers at each step.
3. In general, carrier selection methods have high variance and low bias. This is consistent with the finding of Breiman (1994a, b) and Tibshirani and Knight (1995) who show that subset selection is an unstable procedure.
4. Carrier selection methods perform better for estimating non-smooth functions ("blocks" and "bumps") as compared with estimating smooth functions ("doppler" and "heavisine").
5. The MSE for carrier selection methods is 10%–80% higher than the MSE for oracle-guided wavelet shrinkage. This is remarkable: Wavelet shrinkage with the hard shrinkage function is identical to carrier selection applied to the pool of orthogonal wavelets. Hence, in our examples, restricting the carrier selection methods to the smaller pool of orthogonal wavelets improves performance!
6. The MSE for wavelet shrinkage is 40%–60% higher than the MSE for TIWS. TIWS is superior in terms of both variance and bias. Since variance dominates bias, however, most of the reduction in MSE is due to the reduction in variance.

Remark 1. Coifman and Donoho (1995) report that while TIWS achieves low MSE estimates, it also results in a very large number of noise induced spikes. We did not observe this behavior.

Remark 2. We have found that these results extend to a broad range of functions and transform types, such as wavelet packets, cosine packets, and chirplets. Further simulation results are reported in Bruce, Gao and Stuetzle (1996).

3.2. Influence of sparsity on performance

It is, of course, possible to construct examples for which carrier selection methods applied to the pool of non-decimated wavelets have smaller MSE than TIWS. We have found, however, that one has to work remarkably hard to find such examples. To illustrate this, we constructed functions directly from the overcomplete set of carriers by defining $\mathbf{f} = \mathbf{X}_{TI}\mathbf{b}$ and $b_j \neq 0$ for only a few j. We let the number of non-zero coefficients range from 1 to 72 and selected the coefficients b_j to create a set of functions of increasing complexity. Several of the constructed functions are shown in Figure 1. We compare the MSE for TIWS,

wavelet shrinkage using an orthogonal transform, and matching pursuit using the carrier matrix $\mathbf{X}_{TI}$. We use the same experimental set-up as in Section 3.1.

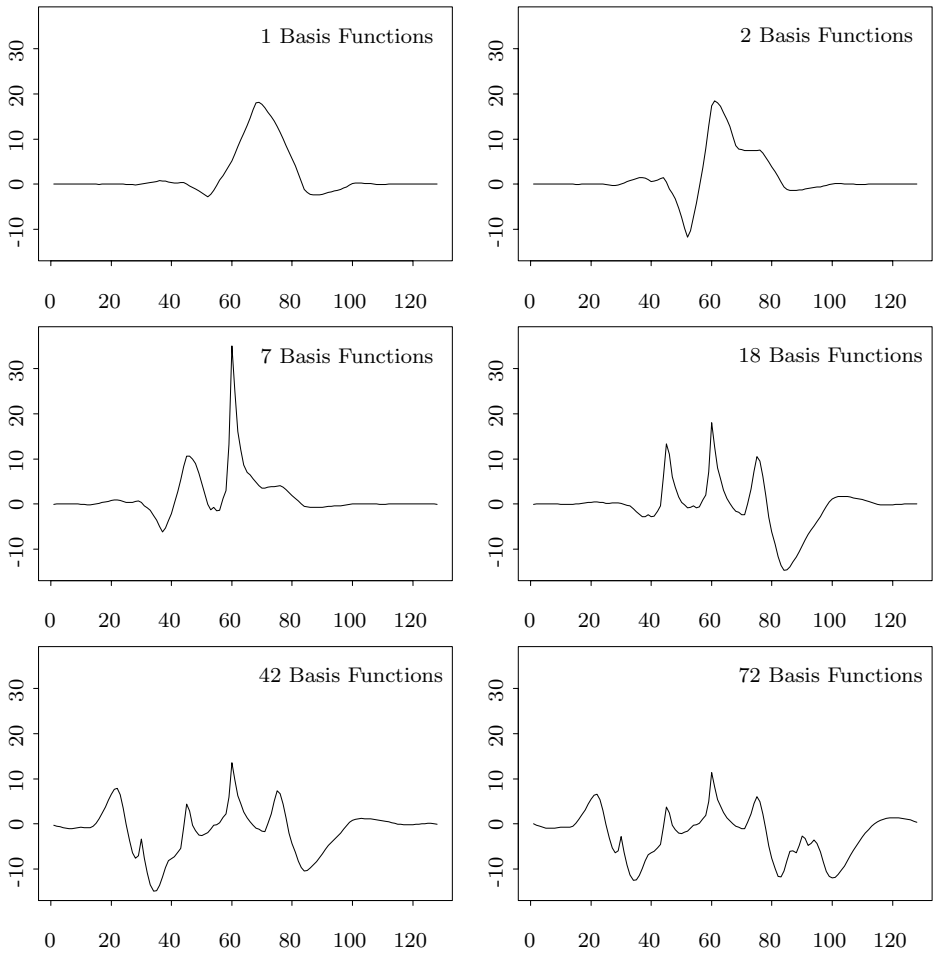

Figure 1. Plotted are functions $\mathbf{f} = \mathbf{X}_{TI}\mathbf{b}$ constructed directly from the over-complete set of carriers $\mathbf{X}_{TI}$ given by the non-decimated wavelet transform. The coefficients $\mathbf{b}$ for each function are selected to create a set of functions of increasing complexity, ranging from one non-zero coefficient (top left) to 72 non-zero coefficients (bottom right).

The results are displayed in Figure 2. Matching pursuit has significantly lower MSE than TIWS when $\mathbf{f}$ is composed of one basis function. When $\mathbf{f}$ is composed of between two and fifteen basis functions, the MSE performance of the methods is roughly comparable. When $\mathbf{f}$ is composed of more than fifteen basis functions, TIWS is the clear winner.

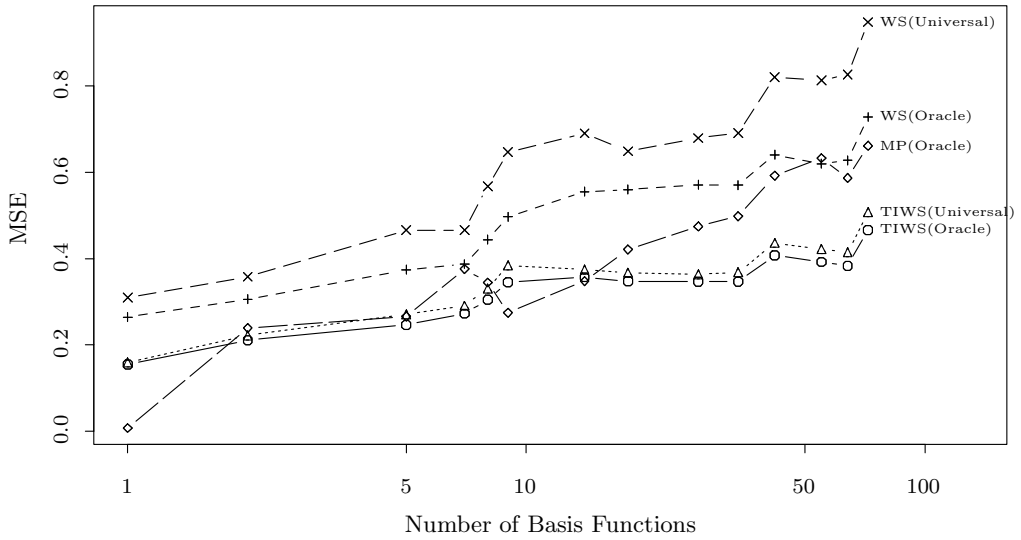

Figure 2. Mean-square-error for translation invariant wavelet shrinkage (TIWS), wavelet shrinkage using an orthogonal transform (WS), and matching pursuit (MP) plotted against number of basis elements in the function to be estimated. The functions are displayed in Fgure 1.

In all cases, TIWS performs almost as well using a universal threshold as with an oracle. By contrast, the performance of wavelet shrinkage using an orthogonal transform is much worse with the universal threshold as compared to an oracle. We further study this issue in the next section.

4. Choice of Shrinkage Function and Threshold

The results of the previous section were based on hard shrinkage with a threshold determined by an oracle and the universal threshold $\sigma(2\log n)^{1/2}$. In this section, we further study the choice of shrinkage function and threshold. We show that hard shrinkage δ_λ^H is clearly preferred over soft shrinkage δ_λ^S for TIWS. Our results also indicate that, compared with wavelet shrinkage using an orthogonal transform, the MSE performance of TIWS is less sensitive to choice of threshold.

For the blocks, bumps, doppler and heavisine functions, we compute the ideal thresholds for hard shrinkage and soft shrinkage. The ideal thresholds minimize the expected MSE $R(\hat{\mathbf{f}},\ \mathbf{f})$ defined by (3). Ideal thresholds were computed for TIWS and for wavelet shrinkage using an orthogonal transform applied to 16 translation shifts $\mathbf{S}_k$ with $k = 0, 1, \ldots, 15$. Table 2 gives the ideal thresholds for $n = 256$, RSNR $= 5$, the least asymmetric "s8" wavelet and resolution level $J = 4$. The corresponding expected MSE's are given in Table 3.

Table 2. Ideal thresholds for translation invariant wavelet shrinkage and (TIWS) the minimim, median and maximum of the ideal thresholds for different spin cycles $k = 0, 1, \ldots, 15$ of wavelet shrinkage using an orthogonal transform. The corresponding mean-square-error's are given in Table 3. See Section 4 for details.

	Blocks	Bumps	Doppler	Heavisine
TIWS-Hard	2.75	2.67	3.05	3.01
Minimum-Hard	2.30	2.14	2.61	3.18
Median-Hard	2.38	2.24	2.69	3.42
Maximum-Hard	2.57	2.29	2.91	∞
TIWS-Soft	0.96	0.89	1.22	1.81
Minimum-Soft	0.81	0.70	1.04	1.79
Median-Soft	0.84	0.72	1.06	1.87
Maximum-Soft	0.90	0.75	1.11	2.00

Table 3. MSE corresponding to ideal thresholds of Table 2. See Table 2 for details.

	Blocks	Bumps	Doppler	Heavisine
TIWS-Hard	0.378	0.462	0.263	0.140
Best-Hard	0.578	0.669	0.399	0.175
Median-Hard	0.618	0.697	0.452	0.199
Worst-Hard	0.646	0.732	0.479	0.218
TIWS-Soft	0.502	0.562	0.350	0.140
Best-Soft	0.542	0.637	0.413	0.153
Median-Soft	0.584	0.657	0.442	0.170
Worst-Soft	0.599	0.672	0.453	0.186

For TIWS, the MSE for ideal hard shrinkage is considerably smaller than the MSE for ideal soft shrinkage. This is not the case with wavelet shrinkage using an orthogonal transform: ideal soft and hard shrinkage have comparable MSE. Bruce and Gao (1996) show that, for orthogonal wavelets, hard shrinkage estimates have high variance while soft shrinkage estimates have high bias. Since the effect of cycle spinning is primarily to reduce the variance, the MSE of hard shrinkage is reduced more than the MSE of soft shrinkage.

The MSE performance of TIWS is more robust towards the choice of threshold for than wavelet shrinkage using an orthogonal transform. The expected MSE curve for TIWS is much flatter than for wavelet shrinkage using an orthogonal transform. Moreover, the ideal TIWS threshold is closer to the "minimax" threshold which minimizes a bound on the asymptotic minimax risk. The minimax threshold is a very simple shrinkage rule which ensures that wavelet shrink-

age using an orthogonal transform asymptotically achieves close to the optimal risk (Donoho and Johnstone (1994)).

For the blocks, bumps, and doppler functions, TIWS has larger ideal thresholds than the smallest ideal threshold for any spin cycle for both hard and soft shrinkage. The larger ideal thresholds can be attributed to a change in the variance–bias tradeoff due to cycle spinning, which reduces variance more than bias for these functions. For the heavisine function, TIWS has a smaller threshold. The heavisine function is distinguished from the other functions since it has relatively small discontinuities, and bias makes up a bigger fraction of MSE (see Table 1).

5. Weighted Cycle Spin Estimators

The translation invariant wavelet shrinkage (TIWS) estimator (4), which takes a simple average, is just one way to combine the spin cycle estimates. An alternative is to use a "spin selection estimator" $\hat{\mathbf{f}}_{k^*}$ where $k^* = \arg\min_{0\le k<2^J-1} \|\mathbf{y} - \hat{\mathbf{f}}_k\|^2$ with $\hat{\mathbf{f}}_k = (\mathbf{WS}_{k-1})^t \delta\,(\mathbf{WS}_{k-1}\mathbf{y})$. Whereas the TIWS is analogous to the bagging procedure of Breiman (1994a), the spin selector estimator is analogous to the bumping procedure of Tibshirani and Knight (1995). Another generalization of the simple average estimator (4) is to allow linear combinations of the spin cycle estimates $\hat{\mathbf{f}}_w = \sum_{k=0}^{2^J-1} w_k (\mathbf{WS}_{k-1})^t \delta\,(\mathbf{WS}_{k-1}\mathbf{y})$ where $\sum_{k=0}^{2^J-1} w_k = 1$.

In this section, we derive an approximate upper bound on the amount we can reduce prediction MSE using $\hat{\mathbf{f}}_{k^*}$ and $\hat{\mathbf{f}}_w$ in place of (4). If the sparsity of the function is relatively constant across spin cycles, our results indicate that only a modest reduction in prediction MSE can be achieved. On the other hand, if the function is sparser for certain spin cycles, then a greater reduction in MSE is possible.

5.1. Oracle guided spinning and weighting

Furnished with an oracle $\mathbf{f}$, it is possible to construct an oracle-guided spin selection estimator $\hat{\mathbf{f}}_{k^\#}$ where $k^\# = \arg\min_{0\le k<2^J-1} \|\mathbf{f} - \hat{\mathbf{f}}_k\|^2$. An ideally weighted estimator $\hat{\mathbf{f}}_{\mathbf{w}^\#}$ can be defined similarly. Let

$$d_{k\ell} = \frac{1}{n} E\Big\{ \sum_{i=1}^{n} \Big(\hat{f}_k(i) - f(i)\Big)\Big(\hat{f}_\ell(i) - f(i)\Big) \Big\}$$

and $D = (d_{k\ell})_{2^J\times 2^J}$. Then the L_2 risk of $\hat{f}_w$ is $R(\hat{\mathbf{f}}_w, f) = \mathbf{w}^t D\mathbf{w}$ where $\mathbf{w} = (w_1, \ldots, w_2^J)^t$. When D is full rank, then $\mathbf{w}^\# \equiv \arg\min_{\mathbf{w}^t\mathbf{1}=1}\{\mathbf{w}^t D\mathbf{w}\} = D^{-1}\mathbf{1}/\mathbf{1}^t D^{-1}\mathbf{1}$ where $\mathbf{1}$ is a vector of one's. The corresponding risk is $R(\hat{\mathbf{f}}_{\mathbf{w}^\#}, f) = 1/\mathbf{1}^t D^{-1}\mathbf{1}$. We use the estimtors $\hat{\mathbf{f}}_{k^\#}$ and $\hat{\mathbf{f}}_{\mathbf{w}^\#}$ in our simulation below to determine the biggest improvement we can expect by generalizing the TIWS estimator.

5.2. Simulation study

Table 4 compares the MSE of TIWS, the ideal weighted cycle spin estimator $\hat{\mathbf{f}}_{\mathbf{w}\#}$, the data-driven spin selector estimator $\hat{\mathbf{f}}_{k^*}$, and the oracle-driven spin selector estimator $\hat{\mathbf{f}}_{k\#}$. The estimates are based on $n = 256$ observations, RSNR = 5, $J = 4$ resolution levels, and the s8 wavelet. We use the hard shrinkage function with the minimax threshold ($\lambda = 3.117$).

Table 4. MSE for different spin cycle estimators: see Section 5 for details. The same set-up as in Table 1 is used with the hard shrinkage function and minimax threshold ($\lambda = 3.117$). The values in this table are based on $M = 100$ simulations. The standard error for each value is given in parentheses.

RSNR=5	Blocks	Bumps	Doppler	Heavisine
TIWS	0.399(0.007)	0.480(0.006)	0.268(0.005)	0.146(0.004)
Ideal Weight	0.351(0.007)	0.434(0.006)	0.232(0.005)	0.106(0.003)
Oracle Best Spin	0.591(0.008)	0.706(0.007)	0.379(0.005)	0.153(0.003)
Data Best Spin	0.661(0.011)	0.776(0.010)	0.451(0.008)	0 204(0.006)

The ideal weights give about 10%-15% reduction in MSE over TIWS for the blocks, bumps, and doppler functions. Apparently, the spatial inhomogeneity of these functions is spread evenly through the different spin cycles, and little improvement is seen by weighting the cycles differently. The ideal weights give over 30% reduction in MSE for the heavisine function. The inhomogeneity of the heavisine function is characterized by two jumps which is represented more sparsely — and hence can be better estimated — by some cycles.

TIWS has smaller MSE than the oracle-driven spin selector estimator. For the blocks, bumps, and doppler functions, the MSE for TIWS is over 40% smaller. In other words, even if we use an oracle to tell use which is the best single cycle, we are better off by averaging all cycles.

5.3. Discussion

In certain applications, when the sparsity of the transform coefficients varies substantially across spin cycles, the ideally weighted spin and spin selector estimators can have considerably smaller prediction MSE than the average spin estimator. Bruce, Gao, Mulligan and Satorius (1995) propose a data-driven spin selector estimator which significantly out-performs TIWS for binary signal demodulation of fractionally spaced channels, both in terms of mean-square-error and in terms of probability of symbol detection error. In fact, the spin selector estimator achieves close to the matched filter performance. For estimation of sinusoids, Bruce Gao and Stuetzle (1996) show that an ideally weighted estimator

can dramatically reduce the prediction error when cycle spinning in frequency with the cosine packet transform.

6. Conclusions and Discussion

In the examples studied in this paper, as well as a range of other situations (see Bruce, Gao and Stuetzle (1996)), translation invariant wavelet shrinkage (TIWS) has almost uniformly lower mean squared error (MSE) than matching pursuit and related carrier selection methods applied to the pool of non-decimated wavelets. Moreover, TIWS is computationally more efficient than carrier selection, and it is based on cycle spinning, a very simple and general technique. We caution that the observed improved MSE performance of TIWS is based on empirical results, and any conclusions should be drawn with care.

Our results confirm other studies (Breiman (1994a,b) and Tibshirani and Knight (1995)) which show that subset regression techniques tend to be unstable and can be improved by use of an ensemble estimator. To improve stability of estimators, Breiman (1994b) proposes the bagging procedure and Tibshirani and Knight (1995) propose the bumping procedure. These procedures combine multiple estimates obtained by applying the original estimator to bootstrap samples. Bagging averages the estimates from the bootstrap samples while bumping selects a best estimate from the bootstrap samples. We found that bagging and bumping do not improve the stability of subset selection procedures in the wavelet setting (Bruce, Gao and Stuetzle (1996)). In fact, for the spatially inhomogeneous functions, applying bagging to the matching pursuit procedure considerably inflates prediction mean-square-error. Bagging and bumping perform poorly in this setting because the bootstrap samples do not include local structure of the signal, such as the peaks in the bumps function or the steps in the blocks function. In Bruce, Gao and Stuetzle (1996), we propose an alternative resampling method called "random orthogonal basis" (ROB) which achieves the same performance as TIWS.

Ridge regression has been shown to have smaller mean-square-error than subset regression methods in some situations (Frank and Friedman (1993)). Ridge regression reduces the variability of the estimates by using a shrinkage estimator of the form $(X^tX+\lambda I)^{-1}X^tY$. Recently, other shrinkage methods have been proposed which compare favorably with ridge regression. These methods include the nonnegative garrote (Breiman (1995)), the lasso (Tibshirani (1996)), and basis pursuit de-noising(Chen, Donoho and Saunders (1996)). Another very promising direction is the use of Bayesian shrinkage rules (Vidakovic (1994), Clyde, Parmigiani and Vidakovic (1995), and Chipman, Kolaczyk and McCulloch (1996)). It is an open question whether these shrinkage methods give better MSE prediction

performance than translation invariant wavelet shrinkage and other ensemble estimators.

A Software

All plots and calculations are done using extensions to the S+WAVELETS software toolkit (Bruce and Gao (1994)). S+WAVELETS is a module in the S-Plus software system (Statistical Sciences, 1993). Software for the simulations and a more detail technical report Bruce, Gao and Stuetzle (1996) can be obtained by anonymous ftp to `ftp.statsci.com` in the directory `pub/WAVELETS`.

Acknowledgements

Dr. Bruce and Dr. Gao were supported by ONR contract N00014-93-C-0106, NSF grant DMI-94-61370. Dr. Bruce was also supported by a research fellowship with the U. S. Bureau of the Census. Dr. Stuetzle was supported by NSF grants DMS-9114027 and DMS-0203002.

References

Breiman, L. (1994a). Bagging predictors. Technical report, University of California, Berkeley.

Breiman, L. (1994b). Heuristics of instability in model selection. Technical report, University of California, Berkeley.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics* **37**, 373-384.

Bruce, A. G. and Gao, H.-Y. (1994). *S+WAVELETS Users Manual.* StatSci Division of MathSoft, Inc., 1700 Westlake Ave. N, Seattle, WA 98109-9891.

Bruce, A. G. and Gao, H.-Y. (1996). Understanding waveshrink: Variance and bias estimation. *Biometrika* **83**, 727-745.

Bruce, A. G., Gao, H.-Y., Mulligan, J. J. and Satorius, E. H. (1995). Application of wavelet De-noising to signal demodulation. In *Proc. 29th Asilomar Conf. on Signals, systems and computers*, Pacific Grove, CA.

Bruce, A. G., Gao, H.-Y. and Stuetzle, W. (1996). Subset selection and ensemble wavelet function prediction. Technical Report 45, Data Analysis products Division, MathSoft, Inc., 1700 Westlake Ave. N, Seattle, WA 98109-9891.

Chen, S., Cowan, C. F. N. and Grant, P. M. (1991). Orthogonal least squares learning algorithm for radial basis function networks. *IEEE Trans. Neural Networks* **2**, 302-309.

Chen, S., Donoho, D. L. and Saunders, M. (1996). Atomic decomposition by basis pursuit. Technical report, Stanford University.

Chipman, N. A., Kolaczyk, E. D. and McCulloch, R. E. (1996). Adaptive Bayesian wavelet shrinkage. Technical Report Technical Report 421, University of Chicago.

Clyde, M., Parmigiani, G. and Vidakovic, B. (1995). Multiple shrinkage and subset selection wavelets. Technical Report Discussion Paper 95-37, Duke University.

Coifman, R. R. and Donoho, D. L. (1995). Translation-invariant De-noising. In *Wavelets and Statistics* (Edited by A. Antoniadis and G. Oppenheim), 125-150. Springer-Verlag, New York.

Daubechies, I. (1992). *Ten Lectures on Wavelets.* Society for Industrial and Applied Mathematics, Philadelphia, PA.

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation via wavelet shrinkage. *Biometrika* **81**, 425-455.

Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.* **90**, 1200-1224.

Fan, J. and Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *J. Roy. Statist. Soc. Ser. B* **57**, 371-394.

Frank, E. and Friedman, J. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35**, 109-148.

Friedman, J. H. (1991). Multivariate adaptive regression splines (with discussion). *Ann. Statist.* **19**, 1-67.

Friedman, J. H., Grosse, E. H. and Stuetzle, W. (1983). Multidimensional additive spline approximation. *SIAM J. Sci. Statist. Comput.* **4**, 291-301.

Friedman, J. H. and Silverman, B. W. (1989). Flexible parsimonious smoothing and additive modeling (with discussion). *Technometrics* **31**, 3-39.

Mallat, S. and Zhang, Z. (1993). Matching pursuits with time frequency dictionaries. *IEEE Trans. Signal Processing* **41**, 3397-3415.

Nason, G. P. and Silverman, B. W. (1995). The stationary wavelet transform and some statistical applications. In *Wavelets and Statistics* (Edited by A. Antoniadis and G. Oppenheim), 281-300. Springer-Verlag, New York.

Pati, Y. C., Rezaiifar, R. and Krishnaprasad, P. S. (1993). Orthogonal matching pursuit: recusive function approximation with applications to wavelet decomposition. In *Conference Record of The Twenty-Seventh Asilomar Conference on Signals, Systems and Computers* (Edited by A. Singh).

Qian, S. and Chen, D. (1994). Signal representation using adaptive normalized Gaussian functions. *Signal Processing* **36**, 1-11.

Shensa, M. J. (1992). The discrete wavelet transorm: Wedding the A trous and mallat algorithms. *IEEE Trans. Signal Processing* **40**, 2464-2482.

Statistical Sciences (1993). *S-PLUS User's Manual, Version* 3.2. StatSci, A Division of MathSoft, Inc., 1700 Westlake Ave., Suite 500, Seattle, WA 98109-9891.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 267-288.

Tibshirani, R. and Knight, K. (1995). Model search and inference by bootstrap "bumping". Technical report, University of Toronto.

Vidakovic, B. (1994). Nonlinear wavelet shrinkage with Bayes rules and Bayes factors. Discussion Paper 24, Duke University.

Wang, L. X. and Mendel, J. M. (1992). Fuzzy basis functions, universal approximation, and orthogonal least squares learning. *IEEE Trans. Neuarl Networks.* **3**, 807-814.

Data Analysis Products Division, MathSoft, Inc., 1700 Westlake Ave. N, Suite 500, Seattle, WA 98109-9891, U.S.A.

E-mail: andrew@statsci.com

TeraLogic, Inc., 707 California Street, Mountain View, CA 94041-2005, U.S.A.

E-mail: hgao@teralogic-inc.com

Department of Statistics, Box 354322, University of Washington, Seattle, WA 98195-4322, U.S.A.

E-mail: wxs@stat.washington.edu

(Received January 1997; accepted April 1998)