Semi-supervised Wavelet Thresholding and Applications

Kichun Sky Lee and Brani Vidakovic

April 6, 2010
Author’s Footnote:

Kichun Sky Lee is a Graduate Student, Department of Industrial and Systems Engineering, Georgia Institute of Technology (E-mail: skylee1020@gmail.com). Brani Vidakovic is Professor, Department of Biomedical Engineering, Georgia Institute of Technology (E-mail: brani@bme.gatech.edu).
Abstract

Under the general regression setup, $y_i = g_i + \epsilon_i$, we are interested in estimating a possibly multivariate regression function $g$. The wavelet thresholding is a simple operation in the wavelet domain that selects the subset of wavelet coefficients corresponding to an estimator of $g$ when back-transformed. We propose the selection of this subset in a semi-supervised fashion, in which a neighbor structure and classification function appropriate for wavelet domains are utilized. The unlabeled coefficients are considered to be connected to neighboring coefficients and fall on a low dimensional manifold in the space of all wavelet coefficients. The decision to include a coefficient in the model depends not only on its magnitude but also on the labeled and unlabeled coefficients from its neighborhood. The theoretical properties of the method are discussed and its performance is demonstrated on simulated examples.

Key Words: K-NN, Manifold-regularization, Semi-supervised learning, Shrinkage, Wavelet denoising
1. INTRODUCTION

In recent years statistical wavelet modeling has attracted the attention of both theoretical and applied statisticians. The most important property of wavelets is their adaptive locality in both time and frequency, which helps in dealing with phenomena that change rapidly in both domains.

Shrinkage in the wavelet domain is a simple, yet powerful tool in nonparametric statistical modeling. It utilizes the fact that wavelet transforms are energy-compressing; that is, most of the signal variance is described by only a few wavelet coefficients. Many approaches for wavelet shrinkage are suggested in the literature: shrinkage by Bayes’ rules, unbiased estimator of the risk, multiple hypothesis testing, and cross-validation techniques, to name just a few (Antoniadis, Bigot and Sapatinas 2001). In almost all of the cases there is an underlying statistical model on the wavelet coefficients and the shrinkage rule represents the optimal action in the adopted statistical paradigm. Many wavelet shrinkage methods based on those approaches are proper thresholding rules, meaning that the inclusion of a wavelet coefficient in the model takes place if its magnitude exceeds a particular threshold. Perturbations on the threshold level always affect the selection of wavelet coefficients and proper strategies are needed to ensure that the model is neither over-fitted nor under-fitted.

In this paper, we propose a semi-supervised wavelet shrinkage. For the coefficients whose magnitudes are close to the adopted threshold, we seek additional information to decide if they are going to be retained in the model or not. We put this task in the framework of statistical learning and introduce labeled and unlabeled wavelet coefficients. For the labeled coefficients the membership in the model is decided, while for the unlabeled coefficients the decision is not clear and hence additional information is needed. The unlabeled coefficients are processed under a semi-supervised learning paradigm that incorporates information from the labeled neighboring coefficients.

The semi-supervised learning has become very popular in the area of machine learning. It comprises a wide range of methods aimed to enhance learning from both labeled and
unlabeled data and provide better inference (usually in classification and clustering tasks). While labeled data can be expensive or time-consuming to obtain, unlabeled data is usually easy to collect and may carry information useful for the inference, such as in the development of classifiers. In semi-supervised learning, information contained in unlabeled data can be incorporated by a variety of techniques, such as expectation maximization (EM) algorithm, transductive support vector machines (SVMs), graph regularization, and others (Zhu 2005).

We start with two thresholds $\lambda_1$ and $\lambda_2$, $\lambda_1 \leq \lambda_2$. The labeled coefficients would have two labels, 0 or 1, depending on whether they are excluded or included in the model. The exclusion is decided by a thresholding rule using the threshold $\lambda_1$. The coefficients whose magnitude is less than $\lambda_1$ carry the label 0, while the coefficients greater than threshold $\lambda_2$ in magnitude carry the label 1. The labels of some wavelet coefficients with their magnitude between $\lambda_1$ and $\lambda_2$ remain unassigned or undetermined. These will be treated as unlabeled.

By taking the approach of manifold regularization, the decisions for the unlabeled coefficients are based on the neighborhood content of manifold structure. We demonstrate that semi-supervised (SS) shrinkage based on background shrinkages with two different thresholds $\lambda_1$ and $\lambda_2$ possesses optimal properties of their background shrinkages and optimizes performance.

The paper is organized as follows: In Sections 2 and 3 short overviews of wavelet shrinkage and manifold-regularization approach in semi-supervised learning are provided. In Section 4 the semi-supervised shrinkage rule is explored. Asymptotic properties of SS rules are explored in a context of background thresholdings. Comprehensive simulations and comparisons are provided in Section 5. This section also contains a discussion on the selection of parameters, and implementation of extension to other background shrinkages. We conclude the paper with Section 6 in which some possible directions for future research are outlined. Technical points are deferred to the Appendix.
2. WAVELET SHRINKAGE

If wavelet transform $W$ is applied to noisy measurements $y_i = g_i + \epsilon_i, i = 1, \ldots, N$, or in vector notation $\mathbf{y} = \mathbf{g} + \mathbf{\epsilon}$, with normal noise $\mathbf{\epsilon}$, the transformed noise $W\mathbf{\epsilon}$ is normal, as well. This linear and orthogonal transformation can be described for the discrete inputs by an orthogonal matrix $W$ of dimension $N \times N$. With $\mathbf{d} = Wy$ and $\mathbf{\theta} = Wg$, the noise model in the time domain can be reformulated as $\mathbf{d} = \mathbf{\theta} + \mathbf{\epsilon}$ in the wavelet domain. Wavelet shrinkage methods, now widely utilized in nonparametric function estimation, estimate $\mathbf{\theta}$ using noisy observations $\mathbf{d}$. The simplest nonlinear wavelet shrinkage technique is thresholding. The two most common thresholding policies are hard and soft thresholding rules with corresponding rules given by:

$$\theta_{\text{hard}}(d, \lambda) = d 1(|d| > \lambda),$$
$$\theta_{\text{soft}}(d, \lambda) = (d - \text{sgn}(d)\lambda) 1(|d| > \lambda),$$

where $1(A)$ is an indicator of relation $A$.

These two shrinkage mechanisms with properly selected thresholds are characterized by exceptional statistical properties especially under i.i.d. Gaussian noise model (Donoho and Johnstone 1994). Under the correlated Gaussian noise model, they also have near optimal behavior in a wide range of function classes (Johnstone and Silverman 1997). Bayesian approaches (Chipman, Kolaczyk and McCulloch 1997; Vidakovic and Ruggeri 2001; Figueiredo and Nowak 2001) and minimax approaches (Donoho and Johnstone 1998) among others are suggested for accurate estimation of the true signal. The wavelet thresholding based on semi-supervised learning, which is developed in this paper, aims to exploit the hierarchical dependence structure of wavelet coefficients for more accurate model selection.

3. SEMI-SUPERVISED LEARNING

In the real world, we encounter both labeled and unlabeled observations. When we estimate the true link between a label (response) and attributes (variables), it is beneficial to incor-
porate the unlabeled observations conditional on whether labeled and unlabeled attributes come from the same population. This requires a formal model that is capable of handling both kinds.

In this section, Laplacian kernel, associated with manifold regularization (Belkin and Niyogi 2004), was used to incorporate the information contained in the unlabeled observations. This information refers to low dimensional manifolds in the attribute space spanned by both labeled and unlabeled data. When such manifolds are well estimated, the assignment of labels can be more precise.

For the illustration of our point, consider Figure 1. In panel (a) only the labeled observations are present and they suggest that the best decision for the new observation denoted by the question mark is the red rectangle. However, when unlabeled observations are added (blue dots) the intrinsic low-dimensional topology of the attributes suggests that the new observation in question should be the green square.

Suppose that empirical data are given as

\[
(x_1, y_1), \ldots, (x_l, y_l) \in X \times \{0, 1\},
\]

\[
x_{l+1}, \ldots, x_{l+u} \in X,
\]

where \(X\) is the domain of attributes \(x_i\) and the \(y_i\) is the label. There are \(l\) labeled and \(u\) unlabeled observations in (1).

The goal is to find a function \(f\) with domain \(X\) and range \(\{0, 1\}: f : X \mapsto \{0, 1\}\), which is well fitted to the labeled observations and regularized both in the ambient space (data space) and the intrinsic space (low-dimensional space of data) of \(X\). The labels selected are 0 and 1, but in principle the values can be arbitrary. An estimate of \(f\), \(\hat{f}\), is given by

\[
\hat{f} = \arg \min_f \frac{1}{l} \sum_{i=1}^l V(y_i, f) + \gamma_A \|f\|^2_A + \gamma_I \|f\|_I^2, \quad \gamma_A \geq 0, \gamma_I \geq 0,
\]

where \(V\) is a loss function. Possible choices for \(V\) are squared error loss \((y_i - f(x_i))^2\) for Ordinary Least Square (OLS), or hinge loss function \(\max\{0, 1 - y_i f(x_i)\}\) for SVMs. Param-
eter $\gamma_A$ controls the complexity of the solution in the ambient space while $\gamma_I$ controls the complexity of the solution in the intrinsic geometry of $X$.

The graph Laplacian $L$ is constructed with an adjacency matrix $W$ with entries $W_{ij}$ representing the closeness between $x_i$ and $x_j$, and a diagonal matrix $D$ with $D_{ii} = \sum_{j=1}^{I+U} W_{ij}$:

$$L = D - W.$$ While several natural choices for the norm $\| \|_I$ are possible, the graph Laplacian has an intuitive interpretation, in which the penalty is proportional to the adjacency $W_{ij}$ between $x_i$ and $x_j$ when the two are mapped apart.

$$\hat{f} = \arg \min_f \frac{1}{l} \sum_{i=1}^{l} V(y_i, f) + \gamma_A \| f \|_A^2 + \gamma_I \sum_{i \neq j} (f(x_i) - f(x_j))^2 W_{ij}$$

$$= \arg \min_f \frac{1}{l} \sum_{i=1}^{l} V(y_i, f) + f^T \tilde{L} f,$$

where $\tilde{L} = \gamma_A I + \gamma_I L$. Here the estimate for the labeled observations is the same as the labels: $f(y_i) = y_i, i = 1, \ldots, l$. We have $V(y_i, f) = 0$ and the minimization of (2) involves only unlabeled $f_u$:

$$\hat{f} = \begin{bmatrix} f^{(l)} \\ \hat{f}^{(u)} \end{bmatrix} = \arg \min_f f^T \tilde{L} f$$

$$= \arg \min_{f^{(u)}} \begin{bmatrix} (f^{(l)})^T & (f^{(u)})^T \end{bmatrix} \begin{pmatrix} L_1 & L_2 \\ L_2^T & L_3 \end{pmatrix} \begin{bmatrix} f^{(l)} \\ f^{(u)} \end{bmatrix} = \begin{bmatrix} f^{(l)} \\ -L_3^{-1} L_2^T f^{(l)} \end{bmatrix}.$$

The quantity $-L_3^{-1} L_2^T f^{(l)}$ is interpreted as a confidence measure of support of label 1 against label 0, which can be justified by the approach of optimal graph cut, as in Belkin and Niyogi (2004).

4. SEMI-SUPERVISED WAVELET SHRINKAGE

Although wavelet transforms are whitening, in most cases the wavelet coefficients still show inter-dependence. There are several shrinkage mechanisms that utilize the hierarchical dependence of coefficients. Some of the approaches are Hall, Kerkyacharian and Picard (1998) and Cai (1999). The underlying idea is that the inclusion/exclusion of a coefficient should
depend not only on its magnitude, but also on magnitudes of its neighbors. The proposed procedure for semi-supervised shrinkage uses the geometry of coefficients implied by their hierarchical structure. The methodology of semi-supervised wavelet shrinkage is outlined next.

In what follows, the double index \( jk \) representing the scale/shift indices in \( d_{jk} \) is omitted and a “typical” wavelet coefficient is denoted as \( d \). Assume that \( d \) is modeled as

\[
d = \theta + \epsilon, \quad \epsilon \sim N(0, \sigma^2),
\]

where we are interested in estimating the location \( \theta \). The SS shrinkage consists of the following three steps.

**Step-1.** Specify a thresholding rule of \( \delta(d, \lambda) = d \mathbf{1}(|d| \leq \lambda) \). Other choices of thresholding policies (soft, semisoft(Bruce and Gao 1995), garrotte(Gao 1998), et al.) can be used as well.

**Step-2.** Find an estimate \( \hat{\sigma} \) of the standard deviation \( \sigma \) of \( d \), and set two threshold levels \( \lambda_1 \) and \( \lambda_2 \) such that \( \lambda_1 \leq \lambda_2 \). This can be done by selecting a \( (\tau_1, \tau_2) \), where \( \tau_1 \leq \tau_2 \), so that the two threshold levels are

\[
\begin{align*}
\lambda_1 &= \hat{\sigma} \sqrt{(2 + \tau_1) \log N}, \\
\lambda_2 &= \hat{\sigma} \sqrt{(2 + \tau_2) \log N}.
\end{align*}
\]

**Step-3.** Estimate \( \theta \) by \( \theta^{SS} \) as

\[
\theta^{SS}(d|\delta, \lambda_1, \lambda_2) = \begin{cases} 
0, & \text{if } |d| < \lambda_1, \\
d f(d), & \text{if } \lambda_1 \leq |d| < \lambda_2, \\
d, & \text{if } |d| \geq \lambda_2,
\end{cases}
\]

where \( f \) is a decision function depending on neighbors \( d \) of \( d \) that takes values 0 or 1 indicating whether \( d \) is excluded or included, respectively. We can generalize \( \theta^{SS} \) with other choices of thresholding policy by applying the rule to the right-hand side of (6) accordingly.
The coefficient-by-coefficient estimator at Step-1 with thresholding level will be background thresholding of the SS rule. The rule in (6) defines labeled (included and excluded) and unlabeled (undetermined) coefficients. An illustration of $\theta_{SS}$ is given in Figure 2(a) where $\delta$ is hard thresholding and in Figure 2(b) where $\delta$ is semisoft thresholding.

For an example, thresholding levels can be set as follows: $\lambda_2$ as $\sigma \sqrt{2 \log N}$ by universal shrinkage; $\lambda_1$ as $\arg \min_{0 \leq \lambda \leq \sigma \sqrt{2 \log N}} \text{SURE}(\lambda, d)$ by SURE shrinkage paradigm (Donoho and Johnstone 1995) or as $\sigma \sqrt{2(1 - \varepsilon) \log N}$ by a minimax approach (Donoho and Johnstone 1994).

To obtain the estimate $\hat{f}$ of the decision function, the following semi-supervised learning mechanism is adopted. As discussed before, we treat included and excluded coefficients as labeled and coefficients for which $\lambda_1 \leq |d| \leq \lambda_2$ as unlabeled. Let $s$, $k$, and $q$ be the number of included, excluded, and undetermined coefficients, respectively ($N = s + k + q$). Also let $I_i$ and $I_e$ be the index sets of indices for included and excluded coefficients, respectively. The learning mechanism consists of the following three steps.

Step-1. For $N$ wavelet coefficients, define an adjacency matrix $W$. The size of the matrix is $N$ by $N$ and possible entries are given as follows. For a constant $\rho \in \mathbb{R}^+$,

$$W_{nm} = \begin{cases} 
\exp \left( -\frac{|d_m - d_n|^2}{\rho} \right), & d_n \text{ and } d_m \text{ in the same neighborhood,} \\
0, & \text{otherwise.}
\end{cases} \tag{7}$$

The parameter $\rho$ is called the diffusion parameter and its choice depends on the assumption of how much the mutual influence between two coefficients diminishes as their distances increase. When $\rho$ is $\infty$, $W_{nm}$ simply becomes 1 if the two wavelet coefficients are in the same neighborhood and 0 otherwise. In practical applications, we fix the number of neighbors to be $k$ (the $k$ nearest neighbors) and set $\rho$ to be $\infty$. 

10
**Step-2.** Construct the graph Laplacian matrix $L$ ($N$ by $N$). Rearrange $W$ so that $L$ has a block of $L_1$ ($s+k$ by $s+k$) corresponding to the included and the excluded coefficients and $L_3$ ($q$ by $q$) corresponding to the undetermined coefficients:

$$L := \text{diag}(\sum_i W_{1i}, \ldots, \sum_i W_{Ni}) - W = \begin{pmatrix} L_1 & L_2 \\ L_2^T & L_3 \end{pmatrix}.$$  \hspace{1cm} (8)

**Step-3.** Obtain an estimate $\hat{f}_j^{(u)}$ for the undetermined coefficients by

$$\hat{f}_j^{(u)} = 1 \left( \sum_p \Delta_{jp} f_p^{(\text{included})} > \frac{1}{2} \sum_p \Delta_{jp} \right),$$  \hspace{1cm} (9)

where $\Delta$ is $-L_3^{-1} L_2^T$, and $f^{(\text{included})}$ is ($s+k$) by 1 column vector with elements $f_p^{(\text{included})} := 1(p \in I_i)$.

In the Figure 2(c), neighbors for one unlabeled coefficient are depicted: The neighbors are located at the same level, at its upper level and lower level. Derivation of the decision rule (9) is provided in the Appendix A. In the following section, properties of SS shrinkage are discussed.

### 4.1 Interpretation of SS Rule

**K-nearest neighbor (k-NN) algorithm.** The SS rule for an undetermined coefficient can be interpreted as $k$-NN algorithm when its neighborhood consists of labeled coefficients only. In this case, the decision condition (9) proves to be a $k$-NN algorithm that compares the number of included neighbors with the number of excluded neighbors when adjacency between the two neighbors is constant.

**Theorem 4.1.** Let $\delta$ be the background thresholding for $\delta^{SS}$. For a coefficient $d$ ($\lambda_1 < |d| < \lambda_2$) whose neighbors are either included or excluded and in which constant adjacency exists between $d$ and each of its neighbors, the SS estimator at $d$ in (9) becomes a $k$-NN rule;

$$\theta^{SS} = d \mathbf{1}(\#\{\text{included neighbors of } d\} > \#\{\text{excluded neighbors of } d\}).$$
The proof is provided in the Appendix. When the adjacency is not constant but varies across its neighbors, the rule becomes a weighted $k$-NN algorithm. When an undetermined coefficient has another undetermined coefficient as its neighbor, the SS shrinkage can not be expressed simply as $k$-NN but as an estimator based on geometry of all coefficients imposed by the complete adjacency structure. By adjusting the neighborhood in (7), and hence the adjacency matrix, the SS shrinkage generalizes the nearest neighborhood algorithms.

4.2 Optimality and Risk Analysis

**Asymptotic near-optimality.** We demonstrate that the SS rule achieves the same asymptotic convergence rate as its background thresholding rule. Without loss of generality, assume that $\hat{\delta}$ is a hard thresholding estimator, $\delta^{\text{hard}}$, that is proven to be near-optimal as in (12).

Under the model in (4) and the squared error loss, the risk of diagonal projection (DP), $\hat{\theta} = d \mathbb{1}(|\theta| \geq \sigma)$, is

$$R(\hat{\theta}, \theta) = E||\hat{\theta} - \theta||_2^2 = \sum_{i=1}^{N} \min(\theta_i^2, \sigma^2) := R_{\text{oracle}}(\text{DP}, \theta).$$

(10)

The risk is called the oracle risk ($R_{\text{oracle}}$) since it is unachievable because the true $\sigma$ and $\theta$ are unknown.

Donoho and Johnstone (1994) prove that traditional hard thresholding estimators with threshold $\lambda$ exhibit good asymptotic optimality when $\lambda$ is sufficiently close to $\sigma \sqrt{2 \log N}$, which means that for some $\gamma > 0$,

$$(1 - \gamma) \log \log N \leq (\lambda/\sigma)^2 - 2 \log N \leq o(\log N).$$

(11)

Following this result, it will be shown that when risks of $\delta^{\text{hard}}(d, \lambda_1)$ and $\delta^{\text{hard}}(d, \lambda_2)$ are within the log $N$ factor of the ideal risk, the same holds for $\delta^{SS}(d|\delta^{\text{hard}}, \lambda_1, \lambda_2)$ rule.

**Theorem 4.2.** Under the model in (4), the SS estimator $\delta^{SS}(d|\delta^{\text{hard}}, \lambda_1, \lambda_2)$, as defined in (6) above, satisfies the inequality

$$R(\delta^{SS}, \theta) \leq L \left\{ \sigma^2 + \sum_{i=1}^{n} \min(\theta_i^2, \sigma^2) \right\}$$

(12)
for an \( L \sim \log N \) and all \( \theta \in \mathbb{R}^N \), where \( \lambda_1 \) and \( \lambda_2 \) are sufficiently close to \( \sigma \sqrt{2 \log N} \) by (11).

The proof of the theorem, given in the Appendix, is based on bounding \( \delta^{SS}(d|\delta^{hard}, \lambda_1, \lambda_2) \) with \( \delta^{hard}(d, \lambda_1) \) and \( \delta^{hard}(d, \lambda_2) \). In the ‘Oracle’ notation of (10), the equation (12) is

\[
R(\delta^{SS}, \theta) \leq L\left\{ \sigma^2 + R_{oracle}(DP, \theta) \right\}
\]

The inequality indicates that the SS estimator can mimic, within the \( \log N \) factor, the performance of an oracle plus one extra parameter induced by unbiased estimation of \( \theta \). The theorem can be extended to any \( \lambda_1 \) and \( \lambda_2 \) that satisfy asymptotic near-optimality in the sense of (12). For example, the threshold \( \hat{\sigma} \sqrt{(2 + \tau) \log N} \) as in (5) satisfies the condition (11) whenever \( \tau \sim o(1) \).

5. EXAMPLES

In this section we apply the proposed thresholding rules. First, we discuss the selection of parameters \( \lambda_1 \) and \( \lambda_2 \), and the neighborhood structure. This consideration helps in automating the methodology. Next, we compare performance of the proposed rules to eight other popular methods (both global and adaptive). In the simulations we set the primary resolution level \( j_0 \) to be \( \log_2(\log(N)) + 1 \), following the asymptotic considerations given in Chapter 10 of Hardle, Kerkyacharian and Tsybakov (1998).

5.1 Selection of Parameters

In any shrinkage task, selection of the parameters is essential for satisfactory performance of the model. It is also preferable to have guidelines that indicate how to select the parameters, thus providing automaticity to the shrinkage procedure (i.e., user intervention is not required). In SS shrinkage, the neighborhood structure for a coefficient and two threshold parameters should be preset.

The characteristics of the signal can guide the shape and size of the neighborhood, while the selection of two threshold levels can be guided by the variance of the signal and the
The number of neighbors at the upper and the lower levels should depend on the support of the decomposing wavelet. The support size is linked to the cone of influence of a wavelet coefficient Mallat (1999). Therefore, it is preferable to use wavelets with compact support and a sufficient number of vanishing moments. As for the number of neighbors at the upper and lower levels, the suggestion is as follows: the more irregular the signal is locally, the more vertical neighbors the local wavelet coefficient should have. Prior information about smoothness of the signal can be useful in guiding this setup.

Threshold levels. We extend the arguments in Bruce and Gao (1995) to SS shrinkage by computing optimal lower thresholds for a fixed background thresholding $\delta$ from the perspective of the average mean squared error (AMSE). As an example, the background thresholding $\delta$ can be hard thresholding with the upper threshold level $\lambda_2$ as $\sqrt{2 \log N}$. The choice for $\lambda_2$ is motivated by an argument used for universal thresholding: when $Z_1, \ldots, Z_n$ are i.i.d $\mathcal{N}(0,1)$, then as $N \to \infty$, $P\left\{ \max_i |Z_i| > \sqrt{2 \log N} \right\} \to 0$. In other words, when the underlying signal is exactly zero, the hard thresholding estimator with $\sqrt{2 \log N}$ will be zero with high probability. It was observed that the universal threshold is under-fitting, or larger than the optimal threshold. The $\lambda_1$ can be selected as an over-fitting threshold: an example is a threshold set by the cross-validation approach (Nason 1996a).

Distance between two threshold levels. Parameters $\lambda_1$ and $\lambda_2$, equivalently $\tau_1$ and $\tau_2$ as in (5), can be set to minimize AMSE according to the types of signal and the length. The lower threshold $\lambda_1$ can also be chosen accordingly when the upper threshold $\lambda_2$ is set to be $\sqrt{u \log N}$ for $u = 2, 3, 4$. Figure 3 shows the Piecewise-Regular signal at signal-to-noise ratio (SNR) = 5 and sample size $n = 256, 1024, \text{ and } 4048$, smoothed by SS shrinkage with hard thresholding as its background shrinkage for various values of $(\tau_1, \tau_2)$ with $\tau_2$ being 1. The optimal parameters for the signals of various sizes can be selected by minimizing AMSE. As the distance between $\lambda_1$ and $\lambda_2$ increases, the AMSE first decreases and then increases. The parameters are tabulated in Table 1 according to the signal length. Additional information
about the numerical experiments is found in Section 5.2. It is noted that the optimal distance between two threshold levels to minimize AMSE varies as the length of signal changes.

[Figure 3 about here.]

[Table 1 about here.]

5.2 Simulations and Comparisons

We present a simulation study of the performance of the SS method. The simulation is done with the “known truth”, that is, with test functions and SNR specified. We compare AMSEs of SS and several other popular methods. Comparisons between the performance of SS and its own background thresholding method are also made.

**Simulation setting.** For our simulation study, six standard test functions (Bumps, Blocks, HeaviSine, Doppler, Piecewise-Regular, Piecewise-Polynomial) were perturbed by scaled normal noise in order to produce a preassigned SNR. For each method, the six test functions were generated at equally-spaced $N$ points on the unit interval. The true and noisy signals are shown in Figure 4(a) at $N = 1024$ and SNR = 5. Denoised signals with SS shrinkage based on VisuShrink are presented in Figure 4(b). The decomposing wavelets were chosen in a standard way: Symmlet 8 for HeaviSine, Doppler, and Piecewise-Polynomial; Daubechies 6 for Bumps and Piecewise-Regular; and Haar for Blocks. The neighborhood structure and thresholding levels were set as follows: the number of horizontal neighbors was 4; the number of vertical neighbors was 12; the diffusion parameter $\rho$ was 20; $(\tau_1, \tau_2)$ was $(-0.2593, 0)$ for the SS rule based on Hybrid-SureShrink, while it was $(-0.4222, 0)$ for the SS rule based on VisuShrink; VisuShrink was set with the hard thresholding option. The accuracies of the estimated signals were measured by an AMSE over 1000 simulation runs. All computations were carried out using MATLAB with the Wavelab toolbox (Buckheit and Donoho 1995).

[Figure 4 about here.]
Comparisons with other methods. We compare SS shrinkage with several established wavelet-based denoising methods. In particular, we consider the classical term-by-term estimators VisuShrink of Donoho and Johnstone (1994); Hybrid-SureShrink of Donoho and Johnstone (1995); the scale invariant term-by-term Bayesian ABE method of Figueiredo and Nowak (2001); LPM of Cutillo, Jung, Ruggeri and Vidakovic (2008); the “leave-out-half” version of the Cross-Validation method of Nason (1996b); the term-by-term False Discovery Rate (FDR) method of Abramovich and Benjamini (1995); the term-by-term Bayesian estimator BAMS of Vidakovic and Ruggeri (2001); NeighCoeff of Cai and Silverman (2001); and finally BlockJS of Cai (1999). NeighCoeff and BlockJS represent classical estimators that incorporate the blocking procedure to achieve a better performance. We consider the CrossValidation method with the hard thresholding policy, the BlockJS with the option ‘Augment’ (Antoniadis et al. 2001), and SS shrinkage with Hybrid-SureShrink as its background shrinkage.

Figure 5 presents the boxplots of the AMSE computed for the above 9 methods with $N = 1024$ at SNR = 3. We observe that the SS shrinkage performed comparably to other established methods. For some signals, it outperformed several methods such as CrossValidation, FDR, and BlockJS. In particular, it performed best for the HeaviSine signal and surpassed the background threshold Hybrid-SureShrink for all the six signals in terms of the sample average of AMSEs. Its ratios are shown in Figure 7(a).

Comparisons with background thresholding. Figure 6 illustrates the comparisons between SS shrinkage and its background shrinkage (hard thresholding) with a part of standard “Blocks” signal and SNR = 3. Panels (a) and (c) present background estimators for $\lambda_1 = \sqrt{(2 - 0.9300) \log N}$ and $\lambda_2 = \sqrt{(2 + 1) \log N}$, respectively, while panel (b) gives an estimator from the associated SS shrinkage. In contrast to each of the background estimators, the estimation by semi-supervised shrinkage compromises between the two signals in
panel (a) and (b) and is more sensitive to the overall geometry of the true signal, which is evident by signal features and AMSE.

We observed that SS shrinkage outperformed its own background thresholding method. To demonstrate this point, the thresholding level $\lambda^*$ for hard thresholding was preset in such a way that it minimized AMSE from 100 simulations. The simulation was conducted using signals of size $N = 2048$ and with noise such that SNR is 5. Then, the AMSE for each signal was computed based on 1000 simulations. The neighborhood structure was characterized as follows: the number of horizontal neighbors was 20; the number of vertical neighbors was 20; the diffusion parameter $\rho$ was 5000. The threshold levels for the SS rule were obtained by perturbing $\lambda^*$ with estimated noise level $\hat{\sigma}$: $\lambda_1 = \lambda^* - k_1 \hat{\sigma}$ and $\lambda_2 = \lambda^* + k_2 \hat{\sigma}$; $k_1 = k_2 = 0.1$ for Bumps; $k_1 = k_2 = 0.2$ for Blocks, Doppler, and Piecewise-Polynomial; $k_1 = 0.5, k_2 = 0$ for HeaviSine; $k_1 = 0.3, k_2 = 0$ for Piecewise-Regular. The estimate of the noise level was calculated by the median absolute deviation of the wavelet coefficients in the finest level of detail. The threshold levels in terms of $\tau_1$ and $\tau_2$ are shown in Figure 7(c) and AMSE ratios of the SS rule compared to the background method are shown in Figure 7(b). This simulation shows that the SS rule outperformed the background method in terms of AMSE for the six test signals.

[Figure 7 about here.]

6. CONCLUSIONS

In this paper we developed a new method for wavelet-filtering of noisy signals that combines semi-supervised learning with wavelet thresholding. Construction of labeled and unlabeled coefficients for semi-supervised learning and its justification is provided. Both interpretations of the method and its theoretical properties have been explored. We have demonstrated that the performance of semi-supervised shrinkage is comparable to several existing shrinkage methods in terms of average mean-squared error. Evidence has also been provided to show that the method preserves geometries of signals better than underlying background
shrinkage method and that it improves its own background shrinkage method. For future research, we envision combining different decision rules of a general type, not necessarily thresholding, with machine learning algorithms. The implementation is fast in computations for sample sizes up to 4096 and feasible for 8192. We want to explore computationally efficient algorithms for samples of larger sizes.

7. SUPPLEMENTAL MATERIALS

Computer Code: MATLAB package of SSWS that implements semi-supervised wavelet shrinkage. The package also generates the optimal distance between two thresholds and contains comparison tests in this paper. (SSWS.zip) In the spirit of reproducible research, the code is also available at http://www.isye.gatech.edu/~brani/wavelet.html.

APPENDIX A. DERIVATION OF EQUATION (8)

Define \( f^{(\text{included})} \) and \( f^{(\text{excluded})} \) as \((s + k)\) by 1 column vector, respectively,

\[
    f_p^{(\text{included})} = \begin{cases} 
    1, & \text{if } p \in I_i, \\
    0, & \text{if not,}
    \end{cases} \quad \text{and} \quad
    f_p^{(\text{excluded})} = \begin{cases} 
    1, & \text{if } p \in I_e, \\
    0, & \text{if not.}
    \end{cases}
\]  

(A.1)

By equation (3), \(-L_3^{-1}L_2^T f^{(l)}\) is a measure to support the label \( f^{(l)} \) against all the other labels. Then, the \( j \)th undetermined coefficient is to be included if

\[
    [-L_3^{-1}L_2^T f^{(\text{included})}]_j > [-L_3^{-1}L_2^T f^{(\text{excluded})}]_j, \quad (A.2)
\]

or, equivalently, (recall \( \Delta \) is denoted to be \(-L_3^{-1}L_2^T\))

\[
    \sum_p \Delta_{jp} f_p^{(\text{included})} > \sum_p \Delta_{jp} f_p^{(\text{excluded})}. \quad (A.3)
\]

By definition of \( f^{(\text{included})} \) and \( f^{(\text{excluded})} \) in equation (A.1), it follows

\[
    f_p^{(\text{included})} = 1 - f_p^{(\text{excluded})} \quad \text{for all } p \in I_i \cup I_e. \quad (A.4)
\]

From equations (A.4) and (A.3), equation (9) follows.
APPENDIX B. PROOF OF THEOREM 4.1

Since neighbors of undetermined coefficients are either included or excluded, the matrix $L_3$ in (8) becomes a diagonal matrix: $L_3 = \text{diag}(\sum_i W_{ij})$. By definition, the elements of $L_2$ become $-W_{ij}$. The right hand side of the decision condition for the $b$-th undetermined wavelet coefficient in (A.2) writes:

$$[-L_3^{-1} L_2^T f(\text{included})]_b = -\sum_{l \in I_i} \sum_h W_{hb}(-W_{lb}) = \sum_h W_{hb} \sum_{l \in I_i} W_{lb}.$$  

Similarly, the right-hand side of (A.2) becomes $\sum_h W_{hb} \sum_{l \in I_e} W_{lb}$. Consequently, the decision rule for the $b$-th undetermined wavelet coefficient is

$$\hat{f}_b^{(u)} = 1\left(\sum_{l \in I_i} W_{lb} > \sum_{l \in I_e} W_{lb}\right),$$

which can be simplified into a $k$-NN algorithm because of constant adjacency among neighbors ($W_{lb} = \text{constant}$, $\forall l, b$):

$$\hat{f}_b^{(u)} = 1\left(\#\{\text{included neighbors of } b\} > \#\{\text{excluded neighbors of } b\}\right).$$

APPENDIX C. PROOF OF THEOREM 4.2

Let $\delta^{\text{hard}}_{\lambda_1}$ and $\delta^{\text{hard}}_{\lambda_2}$ be two hard thresholding estimators, respectively. By the definition of $\delta^{SS}$ at (6), for all $d$ we have

$$|\delta^{\text{hard}}_{\lambda_2}(d)| < |\delta^{SS}(d|\delta^{\text{hard}}, \lambda_1, \lambda_2)| < |\delta^{\text{hard}}_{\lambda_1}(d)|. \quad (A.5)$$

Then, by (A.5), the square error loss of $\delta^{SS}$ at $d$ is bounded,

$$\left(\theta - \delta^{SS}(d|\delta^{\text{hard}}, \lambda_1, \lambda_2)\right)^2 < \max\left\{\left(\theta - \delta^{h}_{\lambda_1}(d)\right)^2, \left(\theta - \delta^{h}_{\lambda_2}(d)\right)^2\right\}.$$  

The risk of $\delta^{SS}$ is bounded;

$$R(\delta^{SS}, \theta) < R(\delta^{\text{hard}}_{\lambda_1}, \theta) + R(\delta^{\text{hard}}_{\lambda_2}, \theta).$$

Since risks of $\delta^{\text{hard}}_{\lambda_1}$ and $\delta^{\text{hard}}_{\lambda_2}$ are within the log $N$ factor of the ideal risk, it follows that $R(\delta^{SS}, \theta)$ is within the log $N$ factor of it.
REFERENCES


# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Illustration of semi-supervised learning where the new observation (the black diamond with question mark) is acquired when the shrinkage is performed on the existing data.</td>
</tr>
<tr>
<td>2</td>
<td>Illustration of estimator $\hat{\delta}<em>{SS}$ when the background shrinkage is (a) hard thresholding, $\hat{\delta}</em>{hard}$ and (b) minimax estimation.</td>
</tr>
<tr>
<td>3</td>
<td>AMSE for Piecewise-Regular signal of (a) length of 256, (b) length of 1024, and (c) length of 4048.</td>
</tr>
<tr>
<td>4</td>
<td>All true signals (in blue line) and noised signals (in black dots) for simulation at the panel (a); estimated signals by semi-supervised shrinkage and its background shrinkage, where the true signal is (a) 1024 and (b) 4048.</td>
</tr>
<tr>
<td>5</td>
<td>Boxplots of AMSE for (1) LPM (GAMMARULE), (2) BAMS, (3) VisuShrink, (4) Hybrid-SureShrink.</td>
</tr>
<tr>
<td>6</td>
<td>Estimated signals by semi-supervised shrinkage and its background shrinkage, where the true signal is (a) 1024 and (b) 4048.</td>
</tr>
<tr>
<td>7</td>
<td>(a) Comparison (AMSE ratio) of the SS rule with its background thresholding (Hybrid-SureShrink).</td>
</tr>
</tbody>
</table>
Figure 1: Illustration of semi-supervised learning where the new observation (the black diamond with question mark) is to be classified in the presence of two labels (a red square and a green circle): (a) when only labeled data are considered, the classification is a red square; (b) when both labeled and unlabeled data are considered, the classification is a green circle.
Figure 2: Illustration of estimator $\delta^{SS}$ when the background shrinkage is (a) hard thresholding, $\delta^{hard}$ and (b) semisoft thresholding, $\delta^{semisoft}$; estimators for undetermined coefficients are red points. In the panel (c), a neighborhood for a wavelet coefficient in a rectangle is illustrated. It contains 6 neighbors at the same level and 5 neighbors each at the upper and lower levels.
Figure 3: AMSE for Piecewise-Regular signal of (a) length of 256, (b) length of 1024, and (c) length of 4048 at SNR = 5; The $\lambda_1$ is selected so that the distance between $\lambda_1$ and $\lambda_2$ is the minimizer of AMSE.
Figure 4: All true signals (in blue line) and noised signals (in black dots) for simulation at the panel (a); estimated signals with SS rule based on VisuShrink at the panel (b).
Figure 5: Boxplots of AMSE for (1) LPM (GAMMARULE), (2) BAMS, (3) VisuShrink, (4) Hybrid-SureShrink, (5) ABE, (6) CV, (7) FDR, (8) NC (9) BJS (10) SS rule based on Hybrid-SureShrink, with $n = 1024$ at SNR=3.
Figure 6: Estimated signals by semi-supervised shrinkage and its background shrinkage, where the true signal is Blocks with SNR = 3, \( \lambda_1 = \sqrt{(2 - 0.9300) \log N} \), and \( \lambda_2 = \sqrt{(2 + 1) \log N} \); (a) for hard thresholding with \( \delta_{\lambda_1}^{\text{hard}} \), (b) for semi-supervised shrinkage based on hard thresholding, \( \delta_{\lambda_1, \lambda_2}^{\text{SS}} \), and (c) for hard thresholding with \( \lambda_2 \), \( \delta_{\lambda_2}^{\text{hard}} \). Note that the signal in panel (b) compromises between the two signals in panel (a) and (b).
Figure 7: (a) Comparison (AMSE ratio) of the SS rule with its background thresholding (Hybrid-SureShrink as for Figure 5); (b) comparison (AMSE ratio) of the SS rule with its background thresholding (hard thresholding with a threshold level that minimizes AMSE) at $N = 2048$ and SNR=5; the SS rule outperformed its background thresholding for all test signals; (c) threshold levels of the background hard thresholding and the SS rule in terms of $\tau$; numbers represent: 1, Bumps; 2, Blocks; 3, HeaviSine; 4, Doppler; 5, Piecewise-Regular; 6, Piecewise-Polynomial.
List of Tables

1. Optimal parameters for semi-supervised shrinkage: column $SS^{(2)}$ has $\lambda_2 = \sqrt{2 \log N}$; column $SS^{(3)}$
Table 1: Optimal parameters for semi-supervised shrinkage: column SS\(^{(2)}\) has $\lambda_2 = \sqrt{2 \log N}$; column SS\(^{(3)}\), $\lambda_2 = \sqrt{3 \log N}$; and column SS\(^{(4)}\), $\lambda_2 = \sqrt{4 \log N}$.

<table>
<thead>
<tr>
<th></th>
<th>SS(^{(2)})</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$N$</td>
<td>$\lambda_1$</td>
<td>$\lambda_2$</td>
<td>$\lambda_1$</td>
<td>$\lambda_2$</td>
<td>$\lambda_1$</td>
</tr>
<tr>
<td>256</td>
<td>256</td>
<td>2.1402</td>
<td>3.3302</td>
<td>2.1687</td>
<td>4.7096</td>
<td>3.3996</td>
</tr>
<tr>
<td>512</td>
<td>512</td>
<td>2.6222</td>
<td>3.5322</td>
<td>2.6961</td>
<td>4.3261</td>
<td>3.4853</td>
</tr>
<tr>
<td>1024</td>
<td>1024</td>
<td>2.8333</td>
<td>3.7233</td>
<td>2.9701</td>
<td>4.5601</td>
<td>3.4555</td>
</tr>
<tr>
<td>2048</td>
<td>2048</td>
<td>3.0150</td>
<td>3.9050</td>
<td>3.1527</td>
<td>4.7827</td>
<td>3.6125</td>
</tr>
<tr>
<td>4096</td>
<td>4096</td>
<td>3.1887</td>
<td>4.0787</td>
<td>3.3653</td>
<td>4.9953</td>
<td>3.6581</td>
</tr>
</tbody>
</table>