Generalized Cross-Validation for Wavelet Shrinkage in Nonparametric Mixed Effects Models

Henry HORNG-SHING LU, Su-Yun HUANG, and Fang-Jiun LIN

A nonlinear wavelet shrinkage estimator was proposed in an earlier article by Huang and Lu. Such an estimator combined the asymptotic equivalence to the best linear unbiased prediction and the Bayesian estimation in nonparametric mixed-effects models. In this article, a data-driven GCV method is proposed to select hyperparameters. The proposed GCV method has low computational cost and can be applied to one or higher dimensional data. It can be used for selecting hyperparameters for either level independent or level dependent shrinkage. It can also be used for selecting the primary resolution level and the number of vanishing moments in the wavelet basis. The strong consistency of the GCV method is proved.

Key Words: Asymptotic BLUP; Bayesian wavelet shrinkage; Soft thresholding.

1. INTRODUCTION

In the last decade wavelets have been an important and successful tool in signal and image processing, especially for denoising and compression. In denoising and compression, the wavelet coefficients are truncated or shrunk toward zero. There are different approaches for truncating and shrinking wavelet coefficients. Hard and soft thresholding schemes were proposed and studied in a series of papers by Donoho and Johnstone (see, e.g., Donoho and Johnstone 1994; Donoho 1995). Later, various Bayesian wavelet shrinkage methods were studied by several authors (Chipman, Kolaczyk, and McCulloch 1997; Abramovich, Sapatinas and Silverman 1998; Vidakovic 1998a,b; Huang and Lu 2000, 2001). The works of Huang and Lu proposed an adaptive nonlinear shrinkage method, BLUPWAVE, based on

Henry Horng-Shing Lu is Professor, Institute of Statistics, National Chiao Tung University, 1001 Ta Hsueh Road, Hsinchu 30050, Taiwan, ROC (E-mail: hslu@stat.nctu.edu.tw). Su-Yun Huang is Associate Research Fellow, Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan, ROC (E-mail: syhuang@stat.sinica.edu.tw). When this work was done, Fang-Jiun Lin was Graduate Student, Institute of Statistics, National Chiao Tung University, 1001 Ta Hsueh Road, Hsinchu 30050, Taiwan, ROC.

Journal of Computational and Graphical Statistics, Volume 12, Number 3, Pages 714–730
DOI: 10.1198/1061860032076

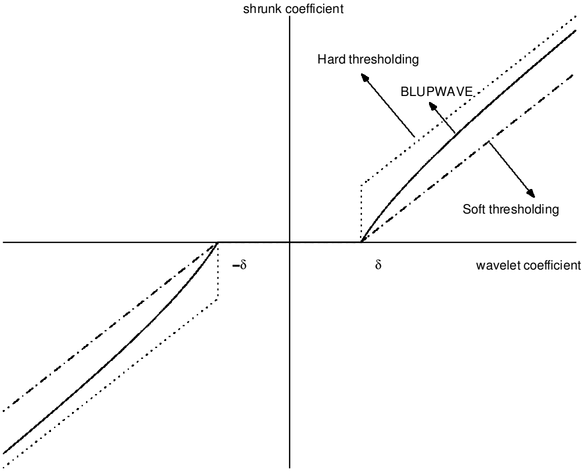

Figure 1. Shrinkage schemes by hard, soft, and BLUPWAVE thresholdings.

perspectives of the Bayesian estimation and the Gauss–Markov estimation (i.e., the BLUP). The shrinkage curve of BLUPWAVE falls between those of hard and soft thresholdings, as seen in Figure 1. These three different thresholding approaches can be summarized by the following.

$$\text{hard thresholding: } \Delta^H(d,\delta) = d\, I(|d| \geq \delta);$$

$$\text{soft thresholding: } \Delta^S(d,\delta) = \text{sign}(d)\ (|d| - \delta)\ I(|d| \geq \delta);$$

and

$$\text{BLUPWAVE: } \Delta^D(d,\delta) = \left(1 - \frac{\delta^2}{d^2}\right) d\, I(|d| \geq \delta);$$

where d is a wavelet coefficient, δ is a threshold parameter, $I(\cdot)$ is the indicator function, and sign$(\cdot)$ is the signum function. The BLUPWAVE shrinkage rule compromises between hard and soft thresholding by drawing positive aspects of both strategies.

The generalized cross-validation method for parameter selection was proposed and its related consistency was studied in the literature (Craven and Wahba 1979; Golub, Heath, and Wahba 1979; Li 1985, 1986, 1987; Wahba 1990). In wavelet shrinkage estimation, data-driven procedures based on cross-validation or generalized cross-validation for selecting parameters of soft and hard thresholdings were studied by Weyrich and Warhola (1995, 1998); Nason (1996, 1999); Jansen, Malfait, and Bultheel (1997); Jansen and Bultheel (1999, 2001); and Jansen (2001). In this article we propose a generalized cross-validation method for selecting parameters encountered in the BLUPWAVE scheme, including selection of threshold parameters at different resolution levels, the primary resolution level and the number of vanishing moments in the wavelet basis. Extending from consistency results in Li (1985, 1986, 1987) for linear estimation, we present the consistency of GCV for the nonlinear BLUPWAVE. Simulation studies are performed to explore the finite sample behavior in practice.

2. BLUPWAVE

We consider the following model of a discrete noisy signal:

$$y_i = f(t_i) + \epsilon_i, \quad i = 1, \ldots, n, \tag{2.1}$$

where t_i's are equally spaced design points over an interval $[a, b]$ and the errors ϵ_i's are iid normal random variables with zero mean and variance σ^2. In vector form, we use the notation

$$y = f + \epsilon, \tag{2.2}$$

where $y = (y_1, \ldots, y_n)'$, $f = (f(t_1), \ldots, f(t_n))'$ and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)'$. The sample size is assumed $n = 2^{m+1}$, $m \in N$. Expand the mean function $f(t)$ in terms of wavelet basis as

$$f(t) = \sum_k \beta_{j,k}\phi_{j,k}(t) \ + \sum_{\ell=j}^{\infty}\sum_k \gamma_{\ell,k}\psi_{\ell,k}(t), \tag{2.3}$$

where j is the primary resolution level, $\{\phi(\cdot), \psi(\cdot)\}$ are a pair of orthogonal scaling function and mother wavelet that generate a multiresolution analysis, and $\phi_{j,k}(t) = 2^{j/2}\phi(2^j t - k)$, $\psi_{\ell,k}(t) = 2^{j/2}\psi(2^j t - k)$.

We adopt a Bayesian approach, with coefficients $\{\beta_{j,k}\}$ being treated as unknown scalars (i.e., fixed effects) and $\{\gamma_{\ell,k}\}$ being treated as random variables (i.e., random effects) having prior distribution $N(0, \eta_\ell)$, where η_ℓ is a variance. An adaptive reconstruction scheme, called BLUPWAVE, was proposed by Huang and Lu (2000) for this model (also known as a nonparametric mixed-effects model). The reconstruction therein is a thresholding rule that combines keep-or-kill and shrinkage with the following form

$$\hat{f}(t) = \sum_{k=1}^{2^j} \hat{\beta}_{j,k}\phi_{j,k}(t) + \sum_{\ell=j}^{m}\sum_{k=1}^{2^\ell}\left(1 - \frac{\lambda}{\hat{\gamma}^2_{\ell,k}}\right)_+ \hat{\gamma}_{\ell,k}\psi_{\ell,k}(t), \tag{2.4}$$

where $\hat{\beta}_{j,k}$'s and $\hat{\gamma}_{\ell,k}$'s are the empirical scaling and wavelet coefficients, respectively, λ is a certain parameter involving σ, and $(\cdot)_+$ means $\max(\cdot, 0)$. The Bayesian estimation for nonparametric mixed effects is also the best linear unbiased prediction (BLUP). Such a predictor has the following asymptotic form

$$\begin{aligned}\hat{f}_{\mathrm{BLUP}}(t) &= \sum_{k=1}^{2^j} \hat{\beta}_{j,k}\phi_{j,k}(t) \\ &\quad + \sum_{\ell=j}^{m}\sum_{k=1}^{2^\ell}\left(1 - \frac{\lambda}{\eta_\ell + \lambda}\right)\hat{\gamma}_{\ell,k}\psi_{\ell,k}(t) + O\left(\frac{1}{n\lambda}\right) \quad \text{almost surely.}\end{aligned}$$

Every $\hat{\gamma}^2_{\ell,k}$, $k = 1, \ldots, 2^{\ell}$, is an asymptotically unbiased estimator for $\eta_\ell + \lambda$. Hence, it leads to the nonlinear BLUPWAVE scheme (2.4). The details were studied by Huang and Lu (2000). The empirical coefficients, $\hat{\beta}_{j,k}$ and $\hat{\gamma}_{\ell,k}$, can be obtained in linear complexity by a discrete wavelet transform (DWT) of y

$$\begin{pmatrix} \hat{\beta} \\ \hat{\gamma} \end{pmatrix} = \frac{Wy}{\sqrt{n}},$$

where W is the orthogonal matrix associated with the DWT (Mallat 1989). Hereafter, we work on the coefficients w which comes from $w = Wy$, instead of the coefficients $(\hat{\beta}, \hat{\gamma})$. Model (2.2) can then be represented as

$$w = Wf + W\epsilon = u + \epsilon^*, \tag{2.5}$$

where the errors ϵ^* are still iid normal with zero mean and variance σ^2. For notational simplicity, we make no distinction between ϵ and ϵ^* henceforth, and use the notation ϵ for both of the errors ϵ and ϵ^*. The BLUPWAVE thresholding scheme becomes

$$\hat{u}_{\lambda,i} = \begin{cases} w_i & \text{for} \quad i = 1, \ldots, 2^j, \\ \left(1 - \frac{n\lambda}{w_i^2}\right)_+ w_i & \text{for} \quad i = 2^j + 1, \ldots, n. \end{cases} \tag{2.6}$$

The focus of this article is on GCV selection of the parameter λ. The GCV method can be also extended to select level dependent (i.e., scale dependent) thresholds for one or higher dimensional data.

3. THE GENERALIZED CROSS-VALIDATION METHOD

3.1 GCV for Level Independent Thresholding

Define the mean square error $R_n(\lambda) = \frac{1}{n}\|\hat{u}_\lambda - u\|_2^2 = \frac{1}{n}\sum_{i=1}^n \left(\hat{u}_{\lambda,i} - u_i\right)^2$. The GCV score is defined by

$$\begin{aligned} \mathrm{GCV}_n(\lambda) &= \frac{\frac{1}{n}\|\hat{u}_\lambda - w\|_2^2}{\left(1 - \frac{1}{n}\sum_{i=1}^n \frac{\partial \hat{u}_{\lambda,i}}{\partial w_i}\right)^2} \\ &= \frac{\frac{1}{n}\|\hat{u}_\lambda - w\|_2^2}{\left(1 - \frac{2^j}{n} - \frac{1}{n}\sum_{i=2^j+1}^n (1 + \frac{n\lambda}{w_i^2}) I(w_i^2 > n\lambda)\right)^2}. \end{aligned}$$

The GCV estimate of λ, given by $\hat{\lambda}_n = \arg\min_{\lambda \geq 0} \mathrm{GCV}_n(\lambda)$, serves as an estimate of the argument minimizer for $R_n(\lambda)$. The strong consistency of the GCV theorem (see Theorem 1) assures that the GCV estimate $\hat{\lambda}_n$ achieves the minimum value for $R_n(\lambda)$ asymptotically under Assumptions 1–3.

Assumption 1. *Assume that the underlying function $f(t)$ is in the Sobolev space $W_2^s[a, b]$, where $s > 0$ is the degree of regularity (or smoothness). Also assume that the wavelet basis used has the number of vanishing moments v with $v > s$.*

Assumption 2. *Assume that $\lambda \to 0$ and $n\lambda \to \infty$, as $n \to \infty$; also assume that the primary resolution level $j \to \infty$, as $n \to \infty$.*

Assumption 3. *Assume that the threshold parameter λ and the primary resolution level j satisfy the constraint* $\lim_{n\to\infty} 2^{2js}\lambda = \infty$.

By Assumption 1, f is in a certain Sobolev space W_2^s. There are also some related Besov spaces $B_{p,q}^s$ imbedded in W_p^s. As the error criterion adopted in this article is the mean square error, we take $p = 2$. Let C^α denote the class of functions that are uniformly Lipschitz with exponent $\alpha > 0$. The following imbeddings are useful: $W_2^s = B_{2,2}^s$, $B_{2,q^-}^{s^+} \subset B_{2,q}^s$, and $C^s \subset W_2^{s^-}$, where $0 < s^- < s \le s^+ < \infty$, and $0 < q^- \le q$. We know that piecewise Lipschitz functions with exponent s are not in W_2^s. They are not in $B_{2,q}^s$, either. The discontinuity at jump point(s) will lower the global regularity. However, such functions are in some other Sobolev spaces with a smaller regularity depending on the type(s) of singularity at the jump point(s). For examples, we discuss four test functions, Blocks, Bumps, HeaviSine, and Doppler, which are used later in the simulation studies. For the Blocks, every block is C^∞ in the interior of its block interval. However, at jump points, the regularity is less than 1/2. The entire Blocks signal belongs to the space $W_2^s[0,1]$ for any $0 < s < 1/2$, but the regularity will never reach 1/2. By the above imbedding results, the Blocks signal is also in $B_{2,q}^s$ for any $0 < s < 1/2$ and $2 \le q \le \infty$. The HeaviSine signal has the same phenomenon. It is C^∞ in the interiors of intervals $(0, 0.3)$, $(0.3, 0.72)$, and $(0.72, 1)$. At jump points, 0.3 and 0.72, the regularity is less than 1/2. The HeaviSine signal belongs to the same Sobolev space as the Blocks does. As for the Bumps, the signal is in $W_2^s[0,1]$ for any $0 < s < 3/2$, and hence in $B_{2,q}^s[0,1]$ for any $0 < s < 3/2$ and $2 \le q \le \infty$. As for the Doppler, the signal is in $C^\infty[h, 1-h]$ for arbitrarily small $h > 0$, but the regularity near the right boundary is less than 3/2. Hence, the Doppler signal belongs to $W_2^s[h,1]$ for any $0 < s < 3/2$ and $B_{2,q}^s[h,1]$ for any $0 < s < 3/2$ with $2 \le q \le \infty$.

The following Sobolev characterization by Mallat (1989) will be used. A function f is in $W_2^s[a,b]$ if and only if it satisfies the condition: $\sum_{\ell,k\in Z}(1 + 2^{2\ell s})|\langle f, \psi_{\ell,k}\rangle|^2 < \infty$. Therefore, for $f \in W_2^s[a,b]$, we have

$$\frac{1}{n}\sum_{i=2^j+1}^{n} u_i^2 = \sum_{\ell=j}^{m}\sum_{k=1}^{2^\ell} |\langle f, \psi_{\ell,k}\rangle|^2 = O\left(2^{-2js}\right).$$

As required by Assumption 2, the primary resolution level j increases to infinity as the data size $n \to \infty$. Usually, j goes to infinity at a somewhat slow rate. An intuitive explanation for this slow rate is as follows. For smoother functions (i.e., larger s), j goes to infinity at a slower rate to allow a wider smoothing bandwidth. Assumption 2 requires that $\lambda \to 0$ and $n\lambda \to \infty$, while Assumption 3 controls the convergence speed so that λ goes to zero at a rate slower than $O(2^{-2js})$. If the primary resolution level is of the optimal order $2^j = O(n^{1/(2s+1)})$, then λ is required to go to zero at a rate slower than $O(n^{-2s/(2s+1)})$ and $n\lambda$ is required to go to infinity at a rate faster than $O(n^{1/(2s+1)})$. That is, the threshold $n\lambda$ for BLUPWAVE is theoretically approaching infinity at a faster rate than the universal threshold $2\sigma^2 \log n$. From Figure 1, the BLUPWAVE must have a larger threshold value than soft thresholding has, if one intends to have the same shrinkage effect (i.e., have the same resulting shrunk coefficient) for both methods. If we draw a horizontal line in the upper half plane in Figure 1, we have to shift the curve of soft thresholding to the left, which means that the soft thresholding must have a smaller threshold parameter δ, to meet with the BLUPWAVE curve at the same intersection point with the horizontal line.

Theorem 1. *(GCV Theorem): Under Assumptions 1, 2, and 3, we have*

$$\lim_{n\to\infty} \frac{R_n(\hat{\lambda}_n)}{R_n(\lambda_n^*)} = 1 \quad \textit{almost surely},$$

where $\hat{\lambda}_n = \arg\min_{\lambda\geq 0} \mathrm{GCV}_n(\lambda)$ *and* $\lambda_n^* = \arg\min_{\lambda\geq 0} R_n(\lambda)$.

Theorem 2. *Under Assumptions 1, 2, and 3, we have*

$$\lim_{n\to\infty} \mathrm{GCV}_n(\lambda) = \lim_{n\to\infty} R_n(\lambda) + \sigma^2 \quad \textit{almost surely.}$$

3.2 PROOFS FOR THEOREM 1 AND THEOREM 2

Define the following notation:

$$\mu_n(\lambda) = \frac{1}{n} \sum_{i=2^j+1}^{n} \left(1 + \frac{n\lambda}{w_i^2}\right) I(w_i^2 > n\lambda),$$

$$p_n(\lambda) = \frac{1}{n} \sum_{i=2^j+1}^{n} (\hat{u}_{\lambda,i} - u_i)(w_i - u_i),$$

and

$$h_n(\lambda) = R_n(\lambda) - \mathrm{GCV}_n(\lambda) + \sigma^2 - 2p_n(\lambda).$$

The next lemmas are established for Theorem 1 and Theorem 2.

Lemma 1. *Under Assumptions 1, 2, and 3, we have*

$$\frac{1}{n} \sum_{i=2^j+1}^{n} \Pr\{w_i^2 > n\lambda\} = O\left(\frac{e^{-n\lambda/(8\sigma^2)}}{\sqrt{n\lambda}}\right).$$

Proof: Recall that $n^{-1}\sum_{i=2^j+1}^{n} u_i^2 = O\left(2^{-2js}\right)$. By Assumption 3, $\lambda = o(2^{-2js})$, then we have $|u_i| < \sqrt{n\lambda}/2$ for $i = 2^j + 1, \ldots, n$, for sufficiently large n. Then

$$\begin{aligned} \frac{1}{n} \sum_{i=2^j+1}^{n} \Pr\{w_i^2 > n\lambda\} &= \frac{1}{n} \sum_{i=2^j+1}^{n} \Pr\{(\epsilon_i + u_i)^2 > n\lambda\} \\ &\leq \frac{2}{n} \sum_{i=2^j+1}^{n} \int_{\sqrt{n\lambda}/2}^{\infty} \frac{e^{-y^2/(2\sigma^2)}}{\sqrt{2\pi}\,\sigma} dy \leq \frac{1}{\sqrt{\pi}} \int_{n\lambda/(8\sigma^2)}^{\infty} \frac{e^{-t}}{\sqrt{t}}\, dt \\ &\leq \frac{2\sqrt{2}\,\sigma e^{-n\lambda/(8\sigma^2)}}{\sqrt{\pi n\lambda}} = O\left(\frac{e^{-n\lambda/(8\sigma^2)}}{\sqrt{n\lambda}}\right). \end{aligned}$$

The last inequality follows from the inequality $1/\sqrt{t} \leq 2\sqrt{2}\,\sigma/\sqrt{n\lambda}$, valid for t in the limits of integration. □

Lemma 2. *Assume Assumptions 1, 2, and 3. For any arbitrary $\alpha > 0$, we have*

$$\frac{1}{n}\sum_{i=2^j+1}^{n} I(w_i^2 > n\lambda) = o\left(\frac{1}{(n\lambda)^\alpha}\right) \qquad \textit{almost surely.}$$

Proof: For arbitrary $\alpha > 0$ and $q > 2\alpha$, we have

$$\begin{aligned} \max_{\substack{2^j+1\leq i\leq n \\ |u_i|<\sqrt{n\lambda}/2}} (n\lambda)^{2\alpha} E\, I(w_i^2 > n\lambda) &= \max_{\substack{2^j+1\leq i\leq n \\ |u_i|<\sqrt{n\lambda}/2}} (n\lambda)^{2\alpha}\, \Pr\{w_i^2 > n\lambda\} \\ &\leq \max_{2^j+1\leq i\leq n} (n\lambda)^{2\alpha}\, \Pr\{|\epsilon_i| > \sqrt{n\lambda}/2\} \\ &\leq \max_{2^j+1\leq i\leq n} \frac{4^q E|\epsilon_i|^{2q}}{(n\lambda)^{q-2\alpha}} < \infty. \end{aligned}$$

The second inequality above is a Markov inequality. We then apply the strong law of large numbers for uncorrelated random variables having a common upper bound to their second moments (Chung 1974, theorem 5.1.2), and get

$$\begin{aligned} &\lim_{n\to\infty} \frac{1}{n}\sum_{i=2^j+1}^{n} (n\lambda)^\alpha I(w_i^2 > n\lambda) \\ &\qquad = \lim_{n\to\infty} \frac{1}{n}\sum_{i=2^j+1}^{n} (n\lambda)^\alpha E\, I(w_i^2 > n\lambda) \quad \text{almost surely.} \qquad (3.1) \end{aligned}$$

By Lemma 1, the limit in (3.1) is zero. Therefore, we can conclude Lemma 2. □

Lemma 3. *Assume Assumptions 1, 2, and 3. For any arbitrary $\alpha > 0$, we have*

$$\mu_n(\lambda) = o\left(\frac{1}{(n\lambda)^\alpha}\right) \qquad \textit{almost surely.}$$

Proof: Note that $(1+\frac{n\lambda}{w_i^2})I(w_i^2 > n\lambda) \leq 2I(w_i^2 > n\lambda)$. Thus, $\mu_n(\lambda) \leq \frac{2}{n}\sum_{i=2^j+1}^{n} I(w_i^2 > n\lambda)$. We can conclude Lemma 3 from Lemma 2. □

Lemma 4. *Assume Assumptions 1, 2, and 3. For any arbitrary $\lambda_1 > 0$, $\lambda_2 > 0$ and $\alpha > 0$, we have $|p_n(\lambda_1) - p_n(\lambda_2)| = o\left((n\lambda_2)^{1/2}/(n\lambda_1)^\alpha\right)$ almost surely.*

Proof: Without loss of generality, we may assume that $\lambda_1 \leq \lambda_2$. For $\kappa = 1, 2$, let $A_{\kappa,n} = \{i : 2^j + 1 \leq i \leq n \text{ and } w_i^2 > n\lambda_\kappa\}$ and $A_{\kappa,n}^c = \{i : 2^j + 1 \leq i \leq n\} \setminus A_{\kappa,n}$.

$$\begin{aligned} |p_n(\lambda_1) - p_n(\lambda_2)| &= \frac{1}{n}\left|\sum_{i=2^j+1}^{n} (\hat{u}_{\lambda_1,i} - \hat{u}_{\lambda_2,i})(w_i - u_i)\right| \\ &= \frac{1}{n}\left|\sum_{A_{2,n}} \frac{(n\lambda_1 - n\lambda_2)\epsilon_i}{w_i} + \sum_{A_{1,n}\cap A_{2,n}^c} \left(w_i - \frac{n\lambda_1}{w_i}\right)\epsilon_i\right|. \quad (3.2) \end{aligned}$$

Similar to the proof for Lemma 2, we have

$$\frac{1}{n}\sum_{A_{2,n}}\left|\frac{(n\lambda_1 - n\lambda_2)\epsilon_i}{w_i}\right| \le \sqrt{n\lambda_2}\left(\frac{\sum_{A_{2,n}}|\epsilon_i|}{n}\right) = o\left((n\lambda_2)^{-\alpha+1/2}\right) \quad \text{almost surely.}$$

for arbitrary $\alpha > 0$, and

$$\frac{1}{n}\sum_{A_{1,n}\cap A^c_{2,n}} w_i\left(1-\frac{n\lambda_1}{w_i^2}\right)\epsilon_i = o\left(\frac{(n\lambda_2)^{1/2}}{(n\lambda_1)^{\alpha}}\right) \quad \text{almost surely.} \tag{3.3}$$

Therefore, we can conclude Lemma 4. □

Lemma 5. *Assume Assumptions 1, 2, and 3. For any arbitrary* $\lambda_1 > 0$ *and* $\lambda_2 > 0$, *we have*

$$|h_n(\lambda_1) - h_n(\lambda_2)| = O\left(\frac{4^j}{n^2}\right) + O\left(\frac{1}{n\,2^{j(s-1)}}\right) \quad \textit{almost surely.}$$

Proof: Without loss of generality, we may assume $\lambda_1 \le \lambda_2$. Let $\nu_n(\lambda) = 2^j/n + \mu_n(\lambda)$. Observe that $\frac{1}{n}\|\hat{u}_\lambda - w\|_2^2 = R_n(\lambda) - 2p_n(\lambda) - \frac{1}{n}\sum_{i=1}^{2^j}\epsilon_i^2 + \frac{1}{n}\sum_{i=2^j+1}^{n}\epsilon_i^2$. Then $h_n(\lambda)$ and $h_n(\lambda_1) - h_n(\lambda_2)$ can be expressed as follows

$$\begin{aligned}
h_n(\lambda) &= \frac{(R_n(\lambda)+\sigma^2)(1-\nu_n(\lambda))^2 - n^{-1}\|\hat{u}_\lambda - w\|_2^2 - 2p_n(\lambda)(1-\nu_n(\lambda))^2}{(1-\nu_n(\lambda))^2}, \\
h_n(\lambda_1) - h_n(\lambda_2) &= \frac{\{R_n(\lambda_1)+\sigma^2-2p_n(\lambda_1)\}\{-2\nu_n(\lambda_1)+\nu_n^2(\lambda_1)\}}{\{1-\nu_n(\lambda_1)\}^2} \\
&\quad - \frac{\{R_n(\lambda_2)+\sigma^2-2p_n(\lambda_2)\}\{-2\nu_n(\lambda_2)+\nu_n^2(\lambda_2)\}}{\{1-\nu_n(\lambda_2)\}^2} \\
&= \sigma^2\left(\frac{\{-2\nu_n(\lambda_1)+\nu_n^2(\lambda_1)\}}{\{1-\nu_n(\lambda_1)\}^2} - \frac{\{-2\nu_n(\lambda_2)+\nu_n^2(\lambda_2)\}}{\{1-\nu_n(\lambda_2)\}^2}\right) \\
&\quad + \frac{\{R_n(\lambda_1)-2p_n(\lambda_1)\}\{-2\nu_n(\lambda_1)+\nu_n^2(\lambda_1)\}}{\{1-\nu_n(\lambda_1)\}^2} \\
&\quad - \frac{\{R_n(\lambda_2)-2p_n(\lambda_2)\}\{-2\nu_n(\lambda_2)+\nu_n^2(\lambda_2)\}}{\{1-\nu_n(\lambda_2)\}^2} \overset{\text{def}}{=} I + II - III.
\end{aligned} \tag{3.4}$$

As $R_n(\lambda) = O\left(\frac{2^j}{n}\right) + O\left(2^{-2js}\right)$, $p_n(\lambda) = O\left(2^{-js}\right)$, and $\nu_n(\lambda) = O\left(\frac{2^j}{n}\right)$, almost surely, we have

$$|I| = O(\nu_n(\lambda_1)) + O(\nu_n(\lambda_2)) = O\left(\frac{2^j}{n}\right), \quad \text{and}$$

$$\begin{aligned}
|II| &= O\left(R_n(\lambda_1) - 2p_n(\lambda_1)\right) O\left(\nu_n(\lambda_1)\right) \\
&= \left(O\left(\frac{2^j}{n}\right) + O\left(2^{-js}\right)\right) O\left(\nu_n(\lambda_1)\right) \\
&= O\left(\frac{4^j}{n^2}\right) + O\left(\frac{1}{n2^{j(s-1)}}\right).
\end{aligned} \tag{3.5}$$

Similarly, III has the same order as II. Therefore, we can conclude Lemma 5. □

Proof for the GCV Theorem: Notice that

$$\begin{aligned} 1 &\leq \frac{R_n(\hat{\lambda}_n)}{R_n(\lambda_n^*)} = \frac{\mathrm{GCV}_n(\hat{\lambda}_n) - \sigma^2 + 2p_n(\hat{\lambda}_n) + h_n(\hat{\lambda}_n)}{R_n(\lambda_n^*)} \\ &\leq \frac{\mathrm{GCV}_n(\lambda_n^*) - \sigma^2 + 2p_n(\hat{\lambda}_n) + h_n(\hat{\lambda}_n)}{R_n(\lambda_n^*)} \\ &= \frac{R_n(\lambda_n^*) - 2p_n(\lambda_n^*) + 2p_n(\hat{\lambda}_n) + h_n(\hat{\lambda}_n) - h_n(\lambda_n^*)}{R_n(\lambda_n^*)} \\ &= 1 + \frac{-2p_n(\lambda_n^*) + 2p_n(\hat{\lambda}_n) + h_n(\hat{\lambda}_n) - h_n(\lambda_n^*)}{R_n(\lambda_n^*)}. \end{aligned} \tag{3.6}$$

Let $A_n = \{i : 2^j + 1 \leq i \leq n \text{ and } w_i^2 > n\lambda\}$ and $A_n^c = \{i : 2^j + 1 \leq i \leq n\} \setminus A_n$. Notice that

$$R_n(\lambda) = \frac{1}{n}\sum_{i=1}^{2^j} \epsilon_i^2 + \frac{1}{n}\sum_{i \in A_n} (\hat{u}_i - u_i)^2 + \frac{1}{n}\sum_{i \in A_n^c} u_i^2 \geq \frac{1}{n}\sum_{i=1}^{2^j} \epsilon_i^2 \sim \frac{2^j \sigma^2}{n}, \tag{3.7}$$

where $a_n \sim b_n$ means that $\lim_{n\to\infty} a_n/b_n = 1$. By Lemmas 4 and 5 and inequalities (3.6) and (3.7), we have

$$\lim_{n\to\infty} \frac{-2p_n(\lambda_n^*) + 2p_n(\hat{\lambda}_n) + h_n(\hat{\lambda}_n) - h_n(\lambda_n^*)}{R_n(\lambda_n^*)} = 0 \quad \text{almost surely.}$$

Therefore, $\lim_{n\to\infty} R_n(\hat{\lambda}_n)/R_n(\lambda_n^*) = 1$, almost surely. □

Proof for Theorem 2: This follows immediately from the proof for Theorem 1 by observing that $\lim_{n\to\infty} h(\lambda) = 0$ almost surely and that $\lim_{n\to\infty} p_n(\lambda) = 0$ almost surely □

3.3 Level-Dependent Thresholding

A more flexible approach, introduced by Johnstone and Silverman (1997), is to allow the threshold parameter of wavelet shrinkage to be level dependent. We will consider the following level-dependent BLUPWAVE scheme

$$\hat{u}_{\lambda,i} = \begin{cases} w_i & \text{for} \quad i = 1, \ldots, 2^j, \\ \left(1 - \frac{n\lambda_\ell}{w_i^2}\right)_+ w_i & \text{for} \quad 2^\ell + 1 \leq i \leq 2^{\ell+1},\ \ell = j, \ldots, m. \end{cases} \tag{3.8}$$

The thresholds, $\Lambda = (\lambda_j, \ldots, \lambda_m)$, form an array of nonnegative parameters.

Since the GCV score function for level dependent thresholding is multivariate, it is difficult to minimize. To reduce the multivariate minimization problem to a univariate problem iteratively, we minimize the GCV score function by coordinatewise descent. This turns the multivariate minimization problem into a sequence of easily solved one-dimensional problems. The initial threshold parameters are set equal to the level independent threshold, that is, $\hat{\Lambda}_0 = (\hat{\lambda}, \ldots, \hat{\lambda})$ and $k = 0$. The next iteration, $\hat{\Lambda}_{k+1}$ is computed as follows.

- $\hat{\lambda}_{k+1,m} = \arg\min_{\lambda_m \geq 0} \mathrm{GCV}(\hat{\lambda}_{k,j}, \ldots, \hat{\lambda}_{k,m-1}, \lambda_m), \quad \lambda_m \leftarrow \hat{\lambda}_{k+1,m}$.
- $\hat{\lambda}_{k+1,m-1} = \arg\min_{\lambda_{m-1} \geq 0} \mathrm{GCV}(\hat{\lambda}_{k,j}, \ldots, \hat{\lambda}_{k,m-2}, \lambda_{m-1}, \hat{\lambda}_{k+1,m}), \quad \lambda_{m-1} \leftarrow \hat{\lambda}_{k+1,m-1}$.

  $\vdots$
- $\hat{\lambda}_{k+1,j} = \arg\min_{\lambda_j \geq 0} \mathrm{GCV}(\lambda_j, \hat{\lambda}_{k+1,j+1}, \ldots, \hat{\lambda}_{k+1,m}), \quad \lambda_j \leftarrow \hat{\lambda}_{k+1,j}$.

For each level, the golden section search method is used to solve the one-dimensional minimization problem. We found that GCV scores are close to convergence after one iteration of coordinatewise descent. Furthermore, the average square errors (ASEs) are found to be reduced by this method. Hence, in principle, one iteration of coordinatewise descent is used in our simulation studies reported later. The iteration order can go from m to j or from j to m. Our simulation results shows that different iteration orders are not significantly different.

3.4 GCV SELECTION FOR THE PRIMARY RESOLUTION LEVEL AND FOR THE NUMBER OF VANISHING MOMENTS IN THE WAVELET BASIS

Nason (1999) discussed choosing the number of vanishing moments in the wavelet basis, primary resolution j, and the threshold in wavelet shrinkage by cross-validation. He considered wavelet shrinkage using the universal threshold $\delta = \sqrt{2 \log n}\, \sigma$ (with $\sigma = 1$ in his simulation study) and applied the leave-one-out cross-validation method to select the level j as well as the number of vanishing moments v in the basis. Conditioned on the selected values of (j, v), he then minimized the leave-one-out cross-validation score function to select the level independent threshold λ.

To reduce the computational complexity, we use the GCV method to select level dependent thresholds $\Lambda = (\lambda_j, \ldots, \lambda_m)$ and parameters (j, v) for the BLUPWAVE. We first compute the minimums of the GCV score functions for the BLUPWAVE with a level independent threshold with respect to all possible values of the primary resolution level ranging from 0 to m and the number of vanishing moments in the Symmlet basis ranging from 4 to 10, where Symmlet refers to the least asymmetric wavelet described in Daubechies (1992). The GCV scores are denoted as $\mathrm{GCV}(\hat{\lambda}(j, v); j, v)$, $j = 0, \ldots, m$, and $v = 4, \ldots, 10$. Then, we select the primary resolution level $\hat{\jmath}$ and the number of vanishing moments $\hat{v}$ which achieves the minimum among $\mathrm{GCV}(\hat{\lambda}(j, v); j, v)$. Conditioned on the selected $(\hat{\jmath}, \hat{v})$, we then consider the level dependent thresholds and apply the multivariate GCV method described in the previous section to select the threshold for every level.

The foregoing discussion for 1-D signals can be extended naturally to 2-D images or higher p-dimensional data. The details are reported in the technical report by Lu, Huang, and Lin (2002).

4. SIMULATION RESULTS AND DISCUSSIONS

4.1 ONE-DIMENSIONAL SIGNALS

Four test signals, Blocks, Bumps, HeaviSine, and Doppler, from Donoho and Johnstone

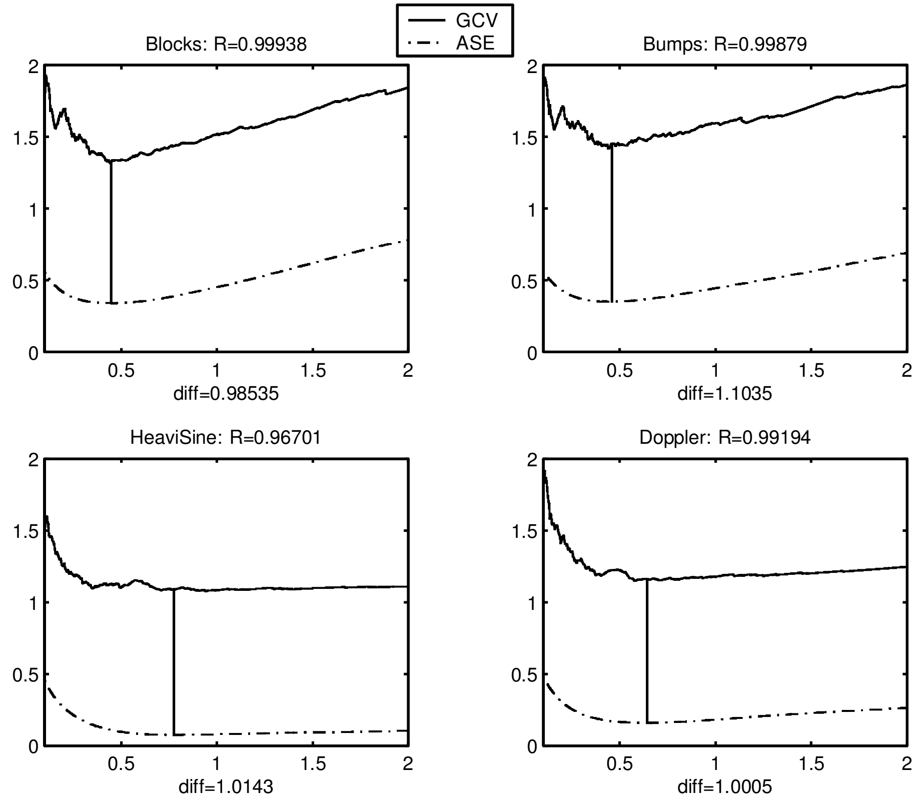

Figure 2. The GCV/σ^2 and ASE/σ^2 curves by BLUPWAVE for four noisy signals when n = 1,024 and SNR = 7.

(1994), are used in the simulation study. The periodic Symmlet wavelet basis over $[0, 1]$ is applied, since these four examples are periodic. The computation is based on the WaveLab package developed by Donoho, Duncan, Huo, and Levi-Tsabari (1999) for MATLAB.

4.1.1 Level Independent Thresholding for 1-D Data

Define the average square error by ASE $= \sum_{i=1}^{n}\{\hat{f}(t_i) - f(t_i)\}^2/n$ and the standardized ASE by

$$\text{standardized ASE} = \frac{1}{n\sigma^2}\sum_{i=1}^{n}\{\hat{f}(t_i) - f(t_i)\}^2.$$

Notice that the ASE is the same as the mean square error $R_n(\cdot)$, since $f = W'u$ and $\hat{f} = W'\hat{u}$, where W is the orthogonal matrix associated with the DWT. The GCV/σ^2 and ASE/σ^2 curves by BLUPWAVE for four test noisy signals are plotted in Figure 2, where $n = 1024$ and the signal to noise ratio is SNR $= 7$. The SNR is defined by $(\sqrt{n}\,\sigma)^{-1}\|f - \bar{f}\|_{\ell_2}$ with $\bar{f} = \sum_{i=1}^{n} f(t_i)/n$. Two reference statistics are given for each plot to check the performance of the GCV procedure in finite samples. They are the difference, diff =

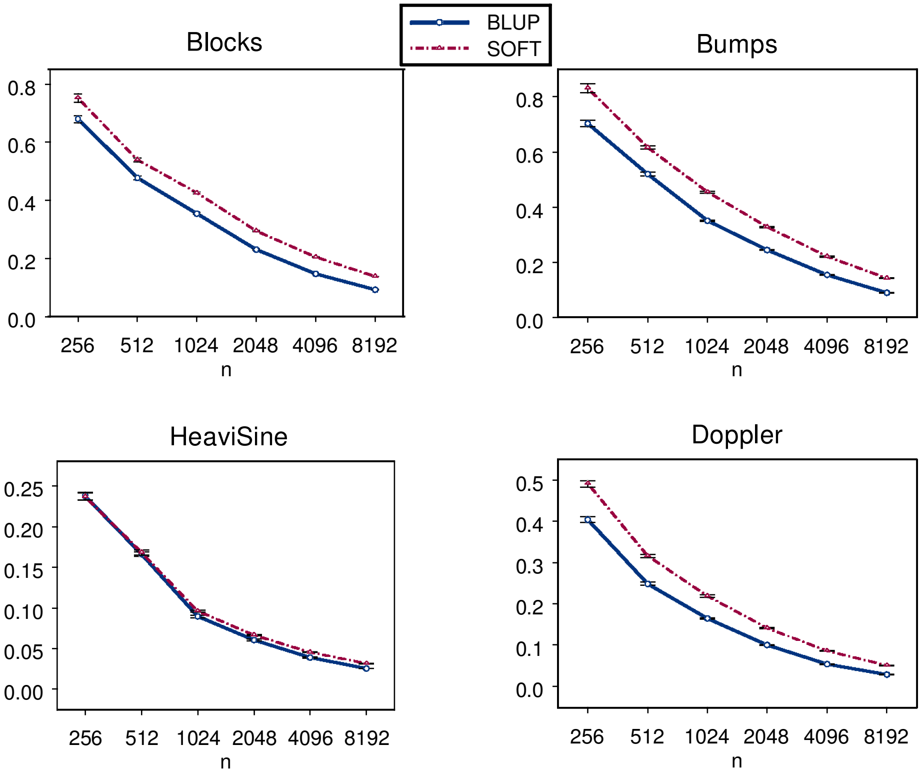

Figure 3. Averages and standard errors of standardized ASE's based on 100 replications of BLUPWAVE and soft thresholding, with SNR = 7.

$\min_\lambda \mathrm{GCV}(\lambda)/\sigma^2 - \min_\lambda \mathrm{ASE}(\lambda)/\sigma^2$, and the relative efficiency

$$\mathcal{R} = \frac{\min_{\lambda \geq 0} \mathrm{ASE}(\lambda)}{\mathrm{ASE}(\hat{\lambda})},$$

where $\hat{\lambda}$ is the GCV selection of λ. Both quantities, diff and $\mathcal{R}$, should approach 1 as n goes to infinity. The search domain of the threshold parameter, $\delta^2 = n\lambda$, in these four test signals is set to $[0, 2\sigma^2 \log n]$. For the purpose of presentation, the horizontal axis in Figure 2 is taken to be $x = n\lambda/(\sigma^2 \log n)$, $0 \leq x \leq 2$.

These four test signals are further studied with sample sizes varying from $n = 256$ to $n = 8192$, SNRs equal to 3, 5, 7 and 10. The Symmlets with eight vanishing moments are used. The primary resolution level is $j = 5$ and the remaining fine scales are all included up to the finest possible resolution. The averages and standard errors of standardized average square errors based on 100 replication runs with various sample sizes and SNRs were reported by Lu, Huang, and Lin (2002) for BLUPWAVE and soft thresholding. Note that the reported numbers are based on standardized ASEs, which are ASEs scaled by error variance σ^2. In other words, the larger SNR cases have the ASEs divided by smaller σ^2 and result in inflated standardized ASEs. Therefore, the reader may find larger standardized

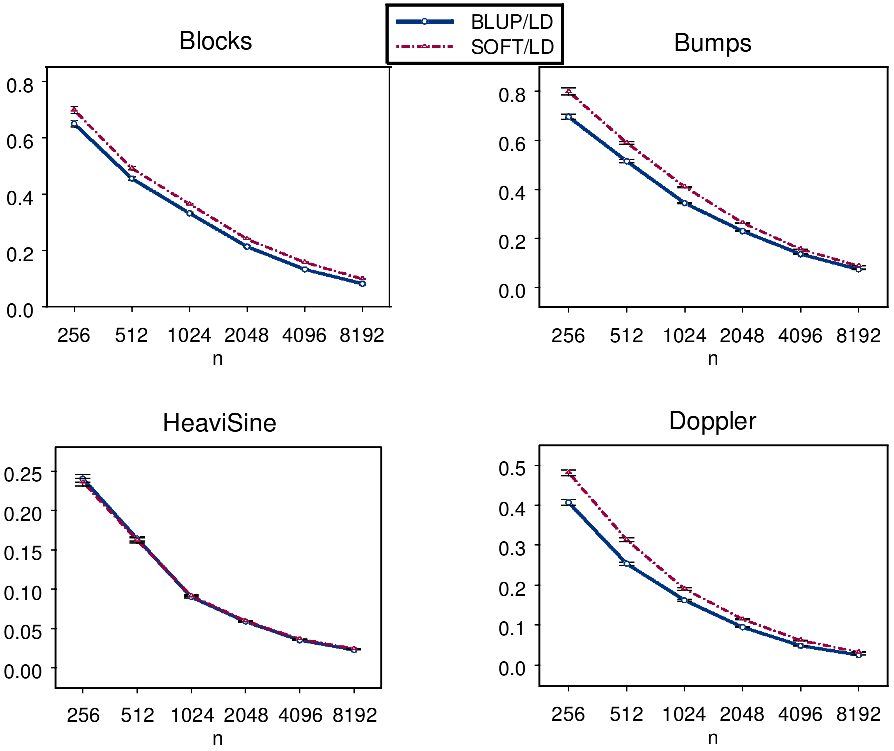

Figure 4. Comparison of the standardized ASEs for BLUPWAVE and soft thresholding with level dependent (LD) thresholds, with SNR = 7.

ASEs for higher SNR cases. The purpose of standardization is to facilitate comparison of the GCV and ASE curves. Plots of standardized ASEs along with ±1.96 times standard errors are given in Figure 3. In the case of Blocks, Bumps, and Doppler, the BLUPWAVE performs better than the soft thresholding does, while in the case of HeaviSine both methods perform with the same quality.

4.1.2 Level Dependent Thresholding for 1-D Data

The averages and standard errors of standardized ASEs based on 100 replications of test signals with various sample sizes and SNRs are reported in the technical report for BLUPWAVE and soft thresholding (Lu, Huang, and Lin 2002).

The comparison of standardized ASEs of BLUPWAVE with or without level dependent (LD) thresholding based on 100 replications when SNR $= 7$ is reported in the technical report (Lu, Huang, and Lin 2002). We observe that the BLUPWAVE has similar quality of performance for level independent and level dependent thresholding. The primary resolution level here is $j = 5$. The comparison plots for BLUPWAVE and soft thresholding are in Figure 4. Again, with level dependent thresholding, the BLUPWAVE has smaller standardized ASEs than soft thresholding does in most cases.

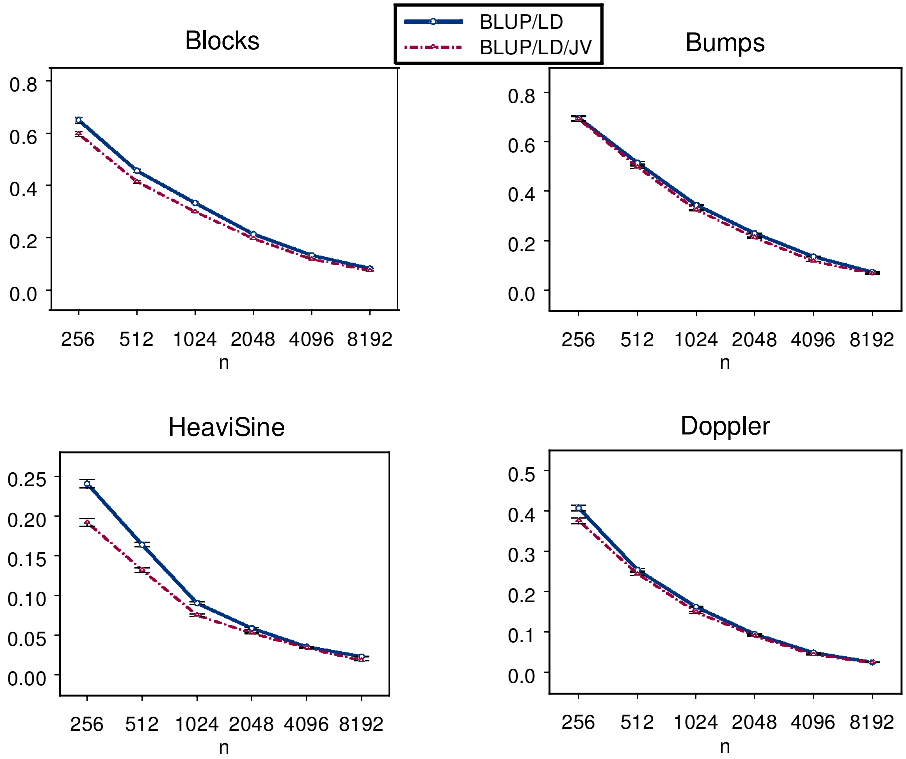

Figure 5. Comparison of the standardized ASEs for BLUP/LD (with pre-assigned $j = 5$ and $v = 8$) versus BLUP/LD/JV (using GCV selection for j and v), with SNR = 7.

4.1.3 Adjustment for the Primary Resolution Level and the Number of Vanishing Moments in the Wavelet Basis for 1-D Data

We take the case of Blocks with $n = 1{,}024$ and SNR $= 7$ in this simulation as an example. The minimums of GCV$/\sigma^2$ values for the BLUPWAVE with level independent thresholding for various values of (j, v) are reported in the technical report (Lu, Huang, and Lin 2002). Among these values the minimum in this table is marked with an asterisk $*$, which corresponds to $\hat{\jmath} = 3$ and $\hat{v} = 8$. Then, conditioned on $\hat{\jmath} = 3$ and $\hat{v} = 8$, the level dependent thresholds $\hat{\lambda}_3, \ldots, \hat{\lambda}_9$ are computed and the standardized ASE is found to be 0.3054. The corresponding value for the case of level dependent thresholding with preassigned primary resolution level $j = 5$ and preassigned number of vanishing moments $v = 8$ is 0.3344 (Lu, Huang, and Lin 2002). Thus the further adjustment of (j, v) does indeed reduce the standardized ASE.

The standardized ASE for BLUPWAVE with preassigned values $j = 5$ and $v = 8$ using level dependent thresholding (denoted by BLUP/LD) is compared in Figure 5 with the standardized ASE with further adjustment for (j, v) (BLUP/LD/JV) based on 100 replications and with SNR $= 7$. It is observed that the GCV selection for j and v do reduce

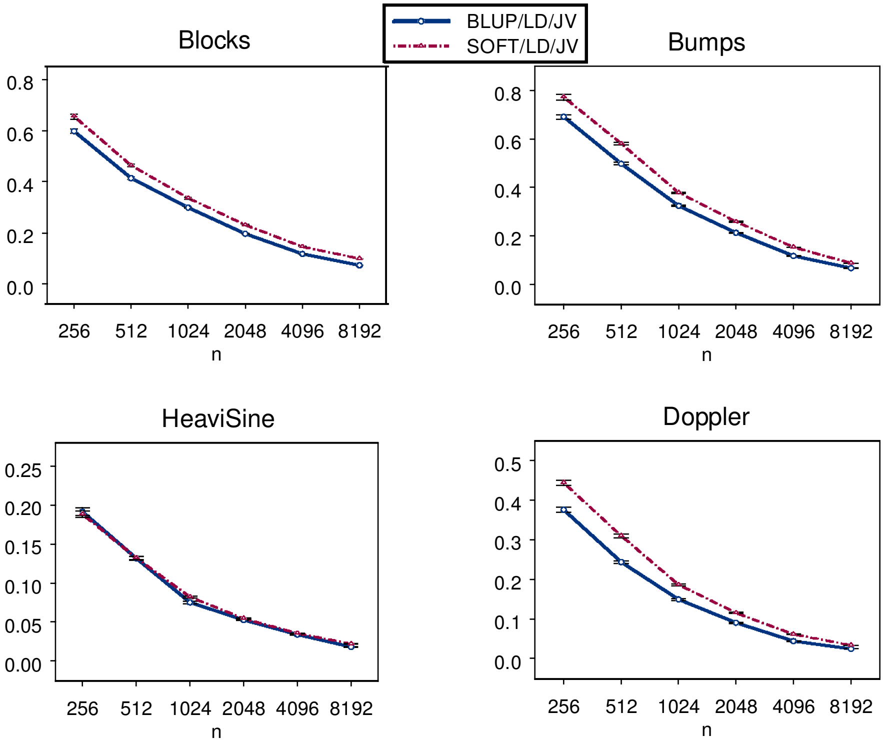

Figure 6. Comparison of the standardized ASEs for BLUP/LD/JV versus SOFT/LD/JV based on 100 replications, with SNR = 7.

the standardized ASE, especially when n is small. For the case of SNR $= 7$, the averages and standard errors of standardized ASEs based on 100 replications for various sample sizes are reported in the technical report for BLUPWAVE and soft thresholding (Lu, Huang, and Lin 2002). Comparisons of BLUP/LD/JV and SOFT/LD/JV are shown in Figure 6. The standardized ASEs of BLUP/LD/JV are smaller than those of SOFT/LD/JV in most cases.

4.2 2-D Images

We also investigated the performance of GCV selection on eight standard test images and obtained results that demonstrate the usefulness of BLUPWAVE here also. Details about this work were given by Lu, Huang, and Lin (2002).

5. CONCLUDING DISCUSSION

The BLUPWAVE shrinkage scheme is based on the use of best linear unbiased prediction as well as Bayesian modeling and estimation. The GCV selection for hyperparameters is proposed. Based on our simulation experience, we find the GCV selection method fast and reliable, and it can also be applied to the selection of level-dependent thresholds, the primary resolution level and the number of vanishing moments in the wavelet basis. With the help of the GCV selection for parameters involved in the BLUPWAVE scheme, we find that BLUPWAVE performs well. Compared to other Bayesian shrinkage schemes, the BLUPWAVE has a very simple asymptotic form. Though its derivation and theoretical background may seem complicated, its final form for shrinkage is really straightforward and easy to implement.

For one-dimensional test signals, BLUPWAVE has smaller standardized ASEs than soft thresholding does. Moreover, the BLUPWAVE has similar performance quality for both the level independent thresholding and the level dependent thresholding. Furthermore, standardized ASEs can be reduced by using GCV to also select the primary resolution level and the number of vanishing moments in the wavelet basis. Especially when n is small, the reduction is quite significant. BLUPWAVE also offers some advantages for denoising 2-D images.

ACKNOWLEDGMENTS

This research was partially supported by the National Science Council, Taiwan, R.O.C. The authors thank the associate editor and referees for their valuable comments and editorial suggestions.

[Received February 2001. Revised July 2002.]

REFERENCES

Abramovich, F., Sapatinas, T., and Silverman, B. W. (1998), "Wavelet Thresholding via a Bayesian Approach," *Journal of the Royal Statistical Society*, Ser. B, 60, 725–749.

Chipman, H. A., Kolaczyk, E. D., and McCulloch, R. E. (1997), "Adaptive Bayesian Wavelet Shrinkage," *Journal of the American Statistical Association*, 92, 1413–1421.

Chung, K. L. (1974), *A Course in Probability Theory* (2nd ed.), Boston: Academic Press.

Craven, P., and Wahba, G. (1979), "Smoothing Noisy Data with Spline Functions: Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-Validation," *Numerische Mathematik*, 31, 377–403.

Daubechies, I. (1992), *Ten Lectures on Wavelets*, CBMS-NSF Series of Applied Mathematics, Philadelphia: SIAM.

Donoho, D. L. (1995), "De-noising by Soft-Thresholding," *IEEE Transactions on Information Theory*, 41, 613–627.

Donoho, D. L., Duncan, M., Huo, X., and Levi-Tsabari, O. (1999), "About WaveLab," Technical Report, Department of Statistics, Stanford University.

Donoho, D. L., and Johnstone, I. M. (1994), "Ideal Spatial Adaptation by Wavelet Shrinkage," *Biometrika*, 81, 425–455.

Golub, G., Heath, M., and Wahba, G. (1979), "Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter," *Technometrics*, 21, 215–223.

Huang, S. Y., and Lu, H. H.-S. (2000), "Bayesian Wavelet Shrinkage for Nonparametric Mixed-Effects Models," *Statistica Sinica*, 10, 1021–1040.

——— (2001), "Extended Gauss-Markov Theorem for Nonparametric Mixed-Effects Models," *Journal of Multivariate Analysis*, 76, 249–266.

Jansen, M. (2001), *Noise Reduction by Wavelet Thresholding*, Lecture Notes in Statistics, 161, New York: Springer-Verlag.

Jansen, M., and Bultheel, A. (1999), "Multiple Wavelet Threshold Estimation by Generalized Cross-Validation for Images with Correlated Noise," *IEEE Transactions on Image Processing*, 8, 947–953.

——— (2001), "Asymptotic Behavior of the Minimum Mean Squared Error Threshold for Noisy Wavelet Coefficients of Piecewise Smooth Signals," *IEEE Transactions on Signal Processing*, 49, 1113–1118.

Jansen, M., Malfait, M., and Bultheel, A. (1997), "Generalized Cross-Validation for Wavelet Thresholding," *Signal Processing*, 56, 33–44.

Johnstone, I. M., and Silverman, B. W. (1997), "Wavelet Threshold Estimators for Data With Correlated Noise," *Journal of the Royal Statistical Society*, Ser. B, 59, 319–351.

Li, K. C. (1985), "From Stein's Unbiased Risk Estimates to the Method of Generalized Cross Validation," *The Annals of Statistics*, 13, 1352–1377.

——— (1986), "Asymptotic Optimality of C_L and Generalized Cross-Validation in Ridge Regression With Application to Spline Smoothing," *The Annals of Statistics*, 14, 1101–1112.

——— (1987), "Asymptotic Optimality for C_p, C_L, Cross-Validation and Generalized Cross-Validation: Discrete Index Set," *The Annals of Statistics*, 15, 958–975.

Lu, H. H.-S., Huang, S. Y., and Lin, F. J. (2002), "Generalized Cross-Validation for Wavelet Shrinkage in Nonparametric Mixed-Effects Models (extended version)," available at http://www.stat.nctu.edu.tw/faculty/hslu/techreport.htm.

Mallat, S. G. (1989), "Multiresolution Approximations and Wavelet Orthonormal Bases of $L^2(R)$," *Transactions of the American Mathematical Society*, 315, 69–87.

Nason, G. P. (1996), "Wavelet Shrinkage Using Cross-Validation," *Journal of the Royal Statistical Society*, Ser. B, 58, 463–479.

——— (1999), "Fast Cross-Validatory Choice of Wavelet Smoothness, Primary Resolution and Threshold in Wavelet Shrinkage using the Kovac–Silverman Algorithm," technical report, Department of Mathematics, University of Bristol, United Kingdom.

Vidakovic, B. (1998a), "Non-linear Wavelet Shrinkage with Bayes Rules and Bayes Factors," *Journal of the American Statistical Association*, 93, 173–179.

——— (1998b), "Wavelet-Based Nonparametric Bayes Methods," in *Practical Nonparametric and Semiparametric Bayesian Statistics*, eds. D. Dey, P. Müller, and D. Sinha, New York: Springer-Verlag, pp. 133–155.

Wahba, G. (1990), *Spline Models for Observational Data*, CBMS-NSF Series of Applied Mathematics, Philadelphia: SIAM.

Weyrich, N., and Warhola, G. T. (1995), "De-noising Using Wavelets and Cross Validation," *Approximation Theory, Wavelets and Applications*, 523–532.

——— (1998), "Wavelet Shrinkage and Generalized Cross Validation for Image Denoising," *IEEE Transactions on Image Processing*, 7, 82–90.