Journal of the Korean Statistical Society (2007), 36: 1, pp 1–29

EMPIRICAL BAYES THRESHOLDING: ADAPTING TO SPARSITY WHEN IT ADVANTAGEOUS TO DO SO

BERNARD W. SILVERMAN[1]

ABSTRACT

Suppose one is trying to estimate a high dimensional vector of parameters from a series of one observation per parameter. Often, it is possible to take advantage of sparsity in the parameters by thresholding the data in an appropriate way. A marginal maximum likelihood approach, within a suitable Bayesian structure, has excellent properties. For very sparse signals, the procedure chooses a large threshold and takes advantage of the sparsity, while for signals where there are many non-zero values, the method does not perform excessive smoothing. The scope of the method is reviewed and demonstrated, and various theoretical, practical and computational issues are discussed, in particularly exploring the wide potential and applicability of the general approach, and the way it can be used within more complex thresholding problems such as curve estimation using wavelets.

AMS 2000 subject classifications. Primary 62C12; Secondary 62G20.
Keywords. Adaptivity, curve estimation, marginal maximum likelihood, nonlinear smoothing, sparsity, wavelets.

1. INTRODUCTION

1.1. Background

There are many statistical problems where the object of interest is a sequence of parameters μ_i on each of which we have a single observation X_i subject to noise, so that

$$X_i = \mu_i + \epsilon_i, \tag{1.1}$$

where the ϵ_i are $N(0,1)$ random variables.

Problems of this kind arise, for example, in astronomical and other image processing contexts, in data mining, in model selection, and in function estimation

Received October 2006; accepted November 2006.
[1]St Peter's College, University of Oxford, OX1 2DL U.K. (e-mail: bernard.silverman@spc.ox.ac.uk)

using wavelets and other dictionaries. In many practical contexts, the sequence μ_i may be sparse, in some sense, and it is important to take advantage of this aspect if possible. This paper reviews a body of joint work with Iain Johnstone on this general topic. For further reading, see the papers Johnstone and Silverman (2004a,b,2005a,b), from which most of the material of this paper is drawn, often verbatim. The `EbayesThresh` package (Silverman, 2005) provides an implementation in R of the methods discussed, and throughout this paper we shall refer to our approach as *the EbayesThresh approach.* A MATLAB translation has been provided in Antoniadis *et al.* (2004).

A natural approach that can make use of sparsity is *thresholding*: if the absolute value of a particular X_i exceeds some threshold t then it is taken to correspond to a a nonzero μ_i which is then estimated, most simply by X_i itself. If $|X_i| < t$ then the coefficient $|\mu_i|$ is estimated to be zero. The quality of estimation is sensitive to the choice of threshold, with the best choice being dependent on the problem setting. In general terms, "sparse" signals call for relatively high thresholds (3, 4, or even higher) while "dense" signals might demand choices of 2 or even lower.

In essence, the EbayesThresh approach is a thresholding method with a threshold estimated from the data. Both theoretical and practical considerations show that the approach has excellent adaptivity properties, in particular by adjusting stably to the sparsity or otherwise of the underlying signal, by choosing an appropriate threshold from the data.

1.2. Examples

Before explaining the method in detail, we consider two examples that give a feeling for the way that the method works in practice.

The first example is of data with a relatively sparse mean vector of length 1000, with 25 nonzero entries uniformly distributed on $(-7, 7)$. To this mean vector is added a sequence of independent normally distributed noise errors with variance 1. The resulting data are plotted in the left panel of Figure 1.1. The EbayesThresh estimate of the mean vector is given in the right panel for comparison. It can be seen that relatively stringent thresholding has been applied to the data; the numerical value of the threshold chosen by the procedure turns out to be 2.99, in the sense that any data point within 2.99 standard deviations of zero is presumed to be pure noise.

The corresponding procedure was carried out for a data set constructed in

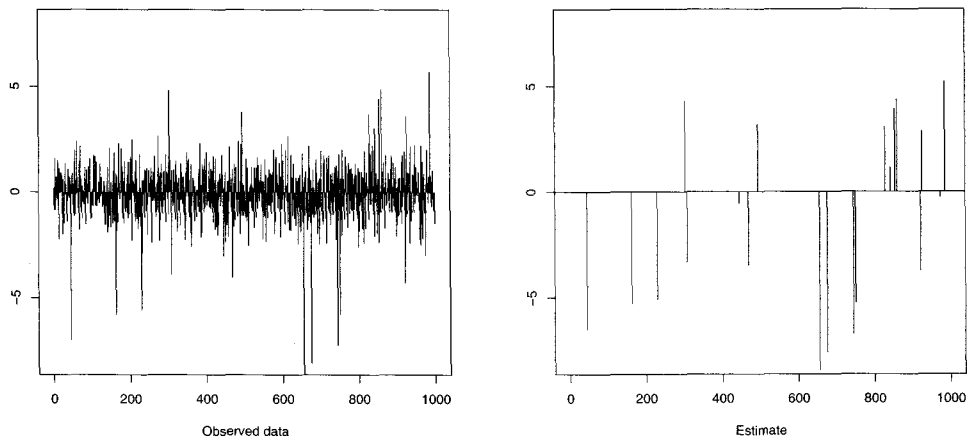

FIGURE 1.1 *Simulated data and estimate for sparse example. Only 25 of the 1000 underlying parameters μ_i are nonzero.*

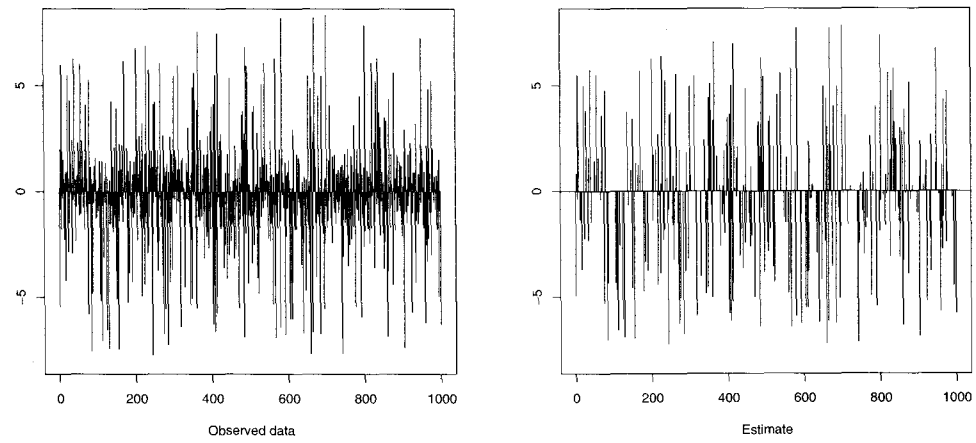

FIGURE 1.2 *Simulated data and estimate for dense example. In this case 250 of the 1000 underlying parameters μ_i are nonzero.*

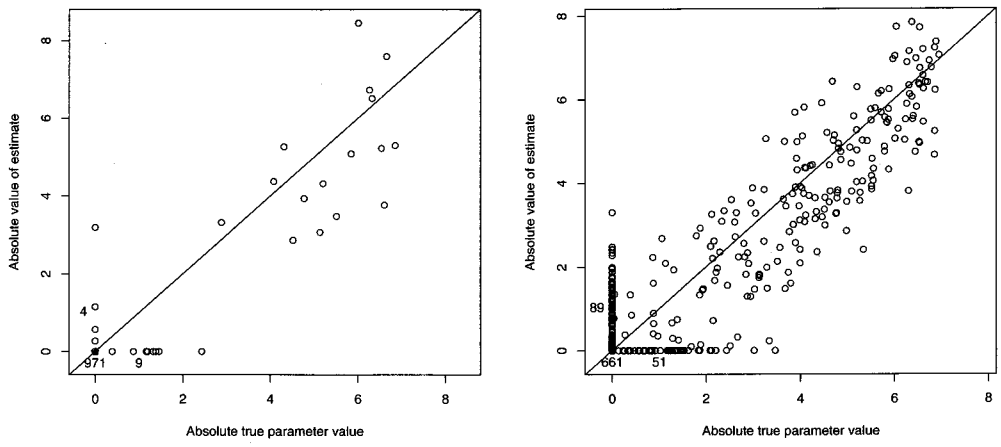

FIGURE 1.3 *Comparison of the performance of the estimation for the sparse and dense examples. In each figure, the absolute value of the estimates is plotted against the absolute value of the corresponding parameters. In the left panel, there are 971 parameters correctly estimated to be zero, 4 zero parameters estimated to be nonzero, and 9 nonzero parameters estimated to be zero. The corresponding quantities for the dense signal considered in the right panel are 661, 89 and 51.*

exactly the same way containing 250 nonzero mean values, a much denser signal. This time the numerical value of the threshold was only 1.67, and so it was much easier for an observation with zero mean to be considered as containing some signal. The results are shown in Figure 1.2. Further insight into the comparison of these examples is given in Figure 1.3. For the sparse signal, the high threshold has the effect that all but 4 of the 975 zero parameters are estimated to be zero. The other 971 are estimated perfectly. However 9 of the 25 nonzero parameters are estimated to zero, thereby presumably incurring slightly more error than if they were not thresholded. In the case of the dense signal, a far higher proportion of the zero parameters, 89 out of 750, are estimated to be nonzero, but the proportion of nonzero parameters incorrectly classified as zeroes is lower: 51 out of 250. The way in which the empirical Bayes method automatically adjusts this tradeoff will be discussed more systematically in Section 3.1 below.

1.3. Brief overview of the method

Very briefly, the main aspects of our method are as follows; they will be discussed further in Section 2.

- A Bayesian model is used for the parameters μ_i. Under this model, each μ_i is zero with probability $(1 - w)$, while, with probability w, μ_i is drawn from a symmetric heavy-tailed density γ.
- The mixing weight w is the key parameter in the prior. It is chosen automatically from the data, using a marginal maximum likelihood approach, and then substituted back into the Bayesian model.
- Estimation within the Bayesian model is a thresholding procedure, and the choice of w is equivalent to a choice of threshold $t(w)$. The method uses this data-based threshold in estimating the underlying vector of parameters from the data.

Within the `EbayesThresh` package, the master routine ***ebayesthresh*** takes a vector `x` and returns an estimate of the parameter values μ_i, by simply carrying out

```
> mu <- ebayesthresh(x)
```

The fuller syntax of the routine is

```
> ebayesthresh(x, prior = "laplace", a = 0.5, bayesfac = FALSE,
+        sdev = NA, verbose = FALSE, threshrule = "median")
```

and reviewing the optional arguments gives an overview of some of the topics discussed in more detail below. Further details are given in the help file for the routine.

The argument ***prior*** specifies the density $\gamma(u)$, the default choice for which is a double exponential, or Laplace, density $(1/2)a \exp(-a|u|)$. The parameter a is given by the argument `a`. In Section 2.1, this choice of prior is discussed further, together with an alternative possibility.

The arguments ***bayesfac*** and ***threshrule*** determine the exact way in which the data are processed once the threshold has estimated. For most practical purposes their default values can be used, but details of other possible approaches are given in Sections 2.2.

The argument ***sdev*** gives the standard deviation of the noise $X_i - \mu_i$ in the data; the default is for this to be estimated from the observed data, from the median absolute value of the X_i. The motivation for this is that even if the sequence μ_i is only reasonably sparse, the median absolute value will not be affected by those observations that have nonzero means μ_i. Clearly some care

may be needed, if there is a possibility that the signal is very far from sparse. If a numerical value of the standard deviation is known, or is estimated by other means, then it can be supplied as the value of ***sdev***.

Finally, the argument ***verbose***, if set to ***TRUE***, causes the routine to produce a list containing a number of different aspects of the thresholding, such as the numerical value of the threshold used, the estimated standard deviation, and so on. This is most useful for research purposes rather than for the actual processing of data.

2. DESCRIPTION OF THE METHOD

In this section, we describe and explain the various aspects of the method. Full details of the various algorithms and calculations required are set out in Johnstone and Silverman (2005a).

2.1. The Bayesian model

In this discussion, we assume throughout that the observations $X_i \sim N(\mu_i, 1)$. If the observations have variance equal to σ^2 rather than 1, then we renormalize the data by dividing by σ, and then multiply the resulting estimates of the means by σ.

Within a Bayesian context, the notion of sparsity is naturally modeled by a suitable prior distribution for the parameters μ_i. We model the μ_i as having independent prior distributions each given by the mixture

$$f_{\text{prior}}(\mu) = (1 - w)\delta_0(\mu) + w\gamma(\mu). \tag{2.1}$$

The nonzero part of the prior, γ, is assumed to be a fixed unimodal symmetric density.

We concentrate on two particular possibilities for the function γ, both of which allow for the necessary calculations to be tractable. Further details of the all the calculations needed for the implementation of the EbayesThresh approach are given in Johnstone and Silverman (2005a).

The Laplace density with scale parameter $a > 0$

$$\gamma_a(u) = \tfrac{1}{2} a \exp(-a|u|)$$

can be used. In the author's experience, a good value for the parameter a is 0.5, and this is the default value in the `EbayesThresh` package.

Another possibility for $\gamma(\mu)$ is specified by the mixture

$$\mu|\Theta = \theta \sim N(0, \theta^{-1} - 1) \text{ with } \Theta \sim \text{Beta}(\tfrac{1}{2}, 1). \tag{2.2}$$

This yields the density

$$\gamma(u) = (2\pi)^{-1/2}\{1 - |u|\tilde{\Phi}(|u|)/\phi(u)\}, \tag{2.3}$$

which has tails that decay as u^{-2}, the same weight as those of the Cauchy distribution. We refer to the density (2.3) as the *quasi-Cauchy* density. Some further discussion is provided in Section 2.3 of Johnstone and Silverman (2004a). The density is one of a family whose tails decay at polynomial rates; the main motivation for the quasi-Cauchy is its combination of heavy tails with reasonable tractability in the present context. It is the heaviest tailed density satisfying the theoretical assumptions made in Johnstone and Silverman (2004a).

2.2. Thresholding rules

Suppose μ has the prior distribution (2.1) and $X \sim N(\mu, 1)$. We can now find the posterior distribution of μ conditional on $X = x$; see Johnstone and Silverman (2005a) for algorithmic details. Define $\hat{\mu}(x; w)$ to be the median of this posterior distribution; for any fixed w, the estimation rule $\hat{\mu}(x; w)$ will be a monotonic function of x with the *thresholding property* that there exists $t(w) > 0$ such that $\hat{\mu}(x; w) = 0$ if and only if $|x| \leq t(w)$. A plot of the posterior median function for a particular value of w is given in Figure 2.1. It is clear that this function is a thresholding rule and that for a range of data values the estimate is exactly zero.

Given a sequence of observations, we can apply the Bayesian procedure separately to each observation X_i to yield an estimate of the corresponding parameter μ_i. The posterior median $\hat{\mu}(X_i; w)$ can be used as this estimate. This is an exact Bayesian procedure if the X_i are independent; if the X_i are not exactly independent then there is some loss of information in the estimation procedure, but if there is not too much dependence then the method will give at least reasonable results.

The posterior median is not the only possible estimation rule once w has been specified; for example one could use the posterior mean $\tilde{\mu}(x; w)$ of μ given $X = x$, but this will not be an exact thresholding rule.

There are other possibilities, for example to determine the threshold $t(w)$ associated with the posterior median with weight w, but then to carry out the

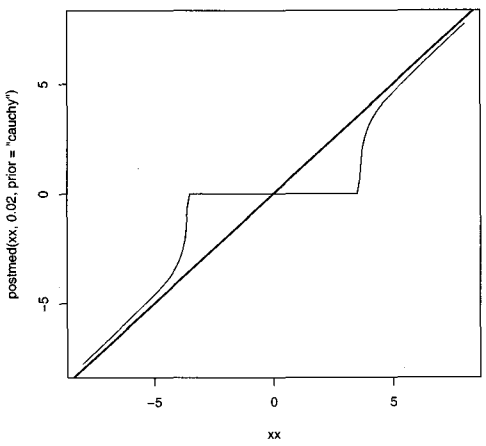

FIGURE 2.1 *Posterior median thresholding function, for quasi-Cauchy prior with mixing weight 0.02.*

actual estimation by hard or soft thresholding with threshold $t(w)$. Another alternative to the posterior median threshold is the *Bayes factor threshold*, defined as the value $\tau_b(w) > 0$ such that

$$P(\mu \neq 0 | X = \tau_b(w)) = 0.5.$$

2.3. Choosing the threshold

The key aspect of the empirical Bayes approach is the choice of mixing weight w, or equivalently of threshold $t(w)$. Assume the X_i are independent; we then estimate w by a marginal maximum likelihood approach. Let $g = \gamma \star \phi$, where $\star$ denotes convolution.

The marginal density of the observations X_i will then be

$$(1 - w)\phi(x) + wg.$$

We define the marginal maximum likelihood estimator $\hat{w}$ of w to be the maximizer of the marginal log likelihood

$$\ell(w) = \sum_{i=1}^{n} \log\{(1 - w)\phi(X_i) + wg(X_i)\}, \tag{2.4}$$

subject to the constraint on w that the threshold satisfies $t(w) \leq \sqrt{2 \log n}$. For our priors, the derivative $\ell'(w)$ is a monotonic function of w, so its root is very

easily found numerically, since the function g is tractable in each case. The bound $\sqrt{2\log n}$ on the threshold is the so-called *universal threshold* for a sample of size n. As explained by Donoho and Johnstone (1994b, p. 445), it is, asymptotically, the maximum absolute value of a sequence of n independent $N(0,1)$ random variables. If the universal threshold is used, then with high probability every zero signal value will be estimated correctly. Therefore, in simple terms, if we wish to take advantage of the possible economy of a signal by thresholding, there is no need to consider thresholds any larger than the universal threshold.

Having used the data once to obtain the estimate $\hat{w}$ by marginal maximum likelihood, we then plug the value $\hat{w}$ back into the prior and then estimate the parameters μ_i by a Bayesian procedure using this value of w, for example as the posterior median $\hat{\mu}(X_i; \hat{w})$. In our implementation the cost of both parts of the procedure is linear in the number of observations considered.

Other parameters of the prior can also be estimated by marginal maximum likelihood. In particular, if the Laplace prior is used, then the scale parameter a can be estimated as follows. Define g_a to be the convolution of $a\gamma(a\cdot)$ with the normal density. Then estimate both a and w by finding the maximum over both parameters of

$$\ell(w,a) = \sum_{i=1}^{n} \log\{(1-w)\phi(X_i) + w g_a(X_i)\}.$$

2.4. Wavelet thresholding and other extensions

Though the basic approach is much more widely applicable, our original motivation for embarking on this work was function estimation using wavelets. In the wavelet context, it is typical that the wavelet coefficients of a true signal will be sparse at the fine resolution scales, and dense at the coarser scales. It is therefore desirable to develop threshold selection methods that adapt the threshold level by level, and so our approach in the wavelet case is to apply the Empirical Bayes method separately to each level of the transform. A full discussion is given in Section 4. A detailed treatment of the approach, including both theoretical and practical aspects, is given in Johnstone and Silverman (2005b).

Another possible extension is to allow the threshold to increase as i increases, reflecting the notion that early μ_i have a reasonably large probability of being nonzero, but that as one proceeds along the sequence nonzero μ_i becomes rarer. For example, μ_i may be the coefficients of a function in a dictionary where the early terms in the sequence describe large-scale aspects of the function or phe-

nomenon of interest, while as we proceed further along the sequence, the terms describe finer and finer detail. For example, Jansen *et al.* (2004) construct just such a basis for the analysis of data observed on an irregular set of points in two dimensions. (Discrete wavelet transforms have something of this character, but are ordered in blocks rather than in a single order.)

Within this general paradigm, we model μ_i as having prior distributions of the same form as previously, but with weight w_i depending on i, so that μ_i has prior density

$$(1 - w_i)\delta(u) + w_i\gamma(u),$$

where δ is a Dirac delta function at zero. If we assume only that the weights w_i decrease as i increases, then we can, in principle, estimate the weights by marginal maximum likelihood. The estimating sequence $\hat{w}_i$ will be chosen to maximize the log marginal likelihood

$$\ell(w_1, \ldots, w_n) = \sum_{i=1}^{n} \log\{(1 - w_i)\phi(x_i) + w_i g(x_i)\}, \tag{2.5}$$

subject to the constraint $w_1 \geq w_2 \geq \cdots \geq w_n$. Once the weights have been estimated, we estimate each μ_i separately, using a thresholding rule based on the Bayesian model with mixing parameter w_i.

It is also possible to allow the prior mixing weight to vary in a more constrained way, but still typically to decrease as i increases, but by constraining it to be proportional to some prescribed sequence c_i, subject to the constraint that it remains bounded between some reasonable lower limit and 1.

Full details of algorithms to implement these approaches, and examples of their use, are provided in Johnstone and Silverman (2005a).

3. Examples and Further Aspects of the Package

In this section, we first consider in detail an example from Johnstone and Silverman (2004a), and then go on to explore some other aspects of the methodology and of the `EbayesThresh` package.

3.1. Finding a threshold for a simulated signal of given sparsity

In Johnstone and Silverman (2004a) a number of artificial images of varying degrees of sparseness are considered. These are shown in Figure 3.1, and the result of adding noise to these images is shown in Figure 3.2.

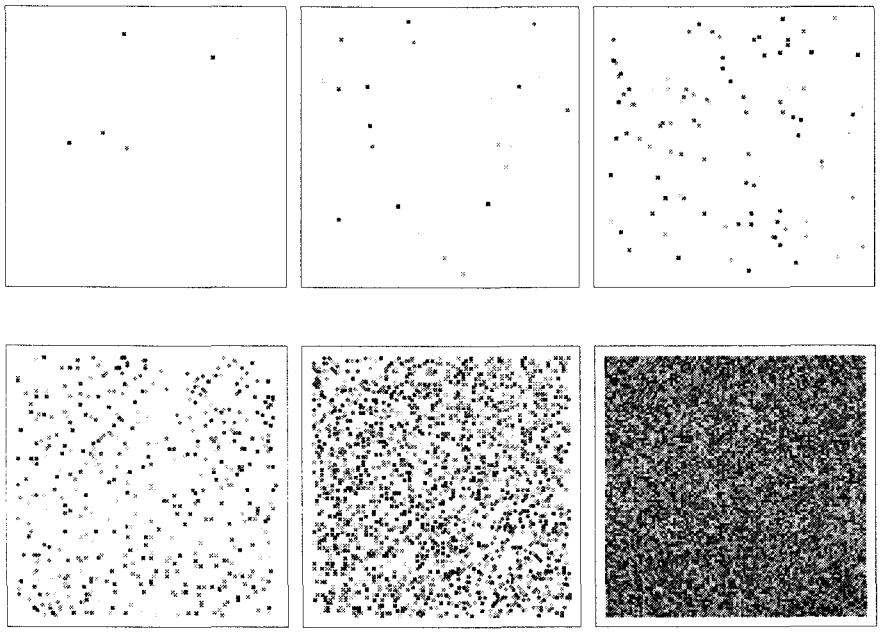

FIGURE 3.1 *Absolute value of parameter images of various sparsity. Out of 10000 pixels, the number of nonzero parameters is, from left to right: 5, 20, 100 in the top row and 500, 2000, 10000 in the bottom row. Each nonzero parameter is chosen independently from a uniform distribution on* $(-5, 5)$.

The average square estimation error yielded by thresholding X_i with varying thresholds is plotted in Figure 3.3. The number in the top right of each panel is the value of the number of nonzero parameters, m, so $m = 5$ corresponds to a very sparse model, while $m = 10000$ corresponds to a very dense model, with no zero parameter values at all. The naive estimator, estimating each μ_i by the corresponding X_i without performing any thresholding at all, will produce an expected mean square error of 1. The scales in each panel are the same, and the threshold range is from 0 to $\sqrt{2 \log 10000} \doteq 4.292$, the so-called *universal threshold* for a sample of this size. The arrow shows the threshold chosen by the EbayesThresh approach.

From this figure we can draw the following conclusions:

- The potential gain from thresholding is very large if the true parameter space is sparse. For the sparsest signals considered in Figures 3.1 and 3.3, the minimum average square error achieved by a thresholding estimate is 0.01 or even less.

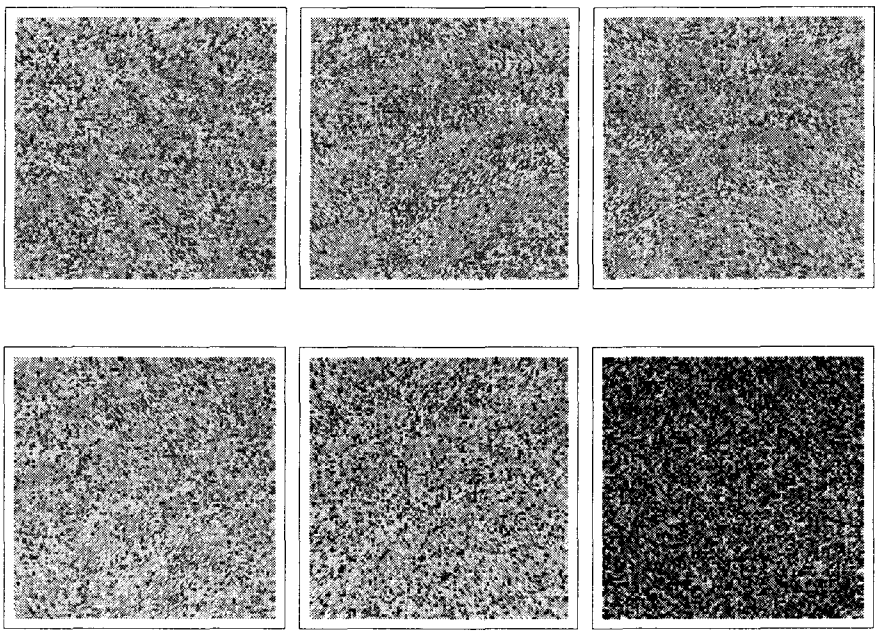

FIGURE 3.2 *Absolute values of data X_i, result of adding Gaussian white noise to the images depicted in Figure 3.1.*

- The appropriate threshold increases as the signal becomes more sparse. For the fully dense signal, no thresholding at all is appropriate, while for the sparsest signals, the best results are obtained using the universal threshold.
- It is important for the threshold to be tuned to the sparsity of the signal; if a threshold appropriate for dense signals is used on a sparse signal, or vice versa, the results are disastrous.
- The EbayesThresh method does an excellent job of tracking the best possible threshold over the whole range from very sparse to completely dense.

4. WAVELET THRESHOLDING

In this section, we go beyond estimation of a single vector or sequence of parameters. We consider the application of the approach to wavelet thresholding, as investigated in Johnstone and Silverman (2005b).

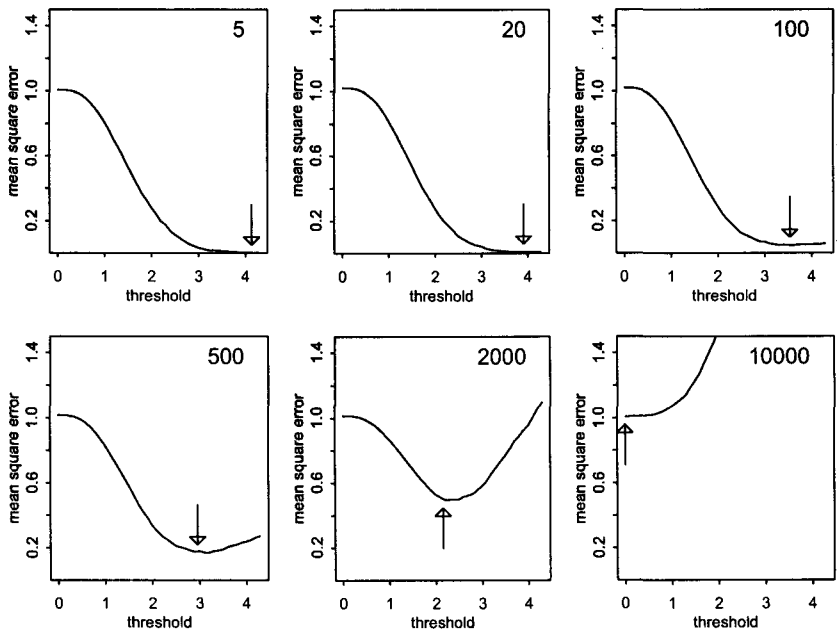

FIGURE 3.3 *Mean square error of thresholding data obtained from the images in Figure 3.1 by adding Gaussian white noise. In each panel, the arrow indicates the threshold chosen by the empirical Bayes approach. The prior used for the nonzero part of the distribution was a Laplace distribution with scale parameter* $a = 1/2$. *Each plot is labeled by the number of nonzero pixels, out of 10000, in the underlying signal.*

4.1. The inductance plethysmography data

The first example is a data set described by Nason (1996) in an anesthesiological study using inductance plethysmography. The data were collected in an investigation of the recovery of patients after general anesthesia. The data are available as part of the `wavethresh3` package (Nason, 1998). This example is used to explain the general approach in the wavelet context. The original data are plotted in Figure 4.1.

4.2. Empirical Bayes thresholding of the discrete wavelet transform

The basic steps in the approach as applied in this case are as follows:

1. Calculate the discrete wavelet transform of the data. In our particular example, we calculate six levels of the transform applying reflection boundary conditions.

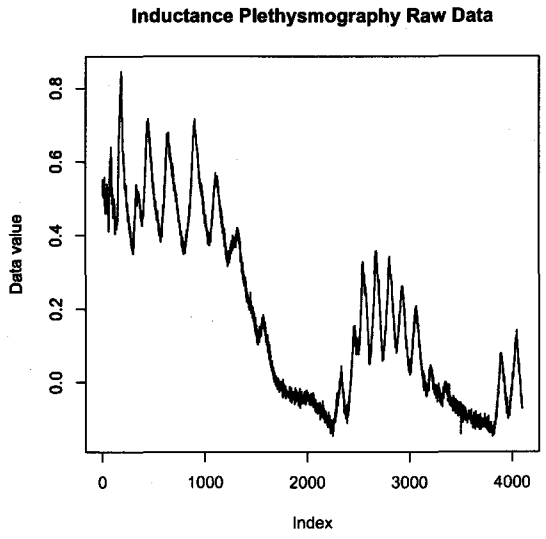

FIGURE 4.1 *Inductance plethysmography data.*

2. Estimate the noise variance from a median-absolute-deviation estimator applied to the coefficients at the finest scale level. This approach gives a value of 0.0108 for the noise standard deviation of the wavelet coefficients.

3. Apply the EbayesThresh method level-by-level to obtain a new array of discrete wavelet coefficients.

4. Invert the resulting discrete wavelet transform to obtain the smoothed estimate of the curve underlying the observed data.

A plot of the resulting curve is given in Figure 4.2.

4.3. The estimated thresholds

In this section we investigate in more detail the way that the EbayesThresh approach deals with the wavelet coefficients at various levels in the given example. Label the finest scale coefficients as level 1, and then subsequently coarser scales consecutively. With this labelling, the thresholds estimated by the method are as in the following table, which gives the thresholds both in absolute terms and in terms of the estimated standard deviation of the noise:

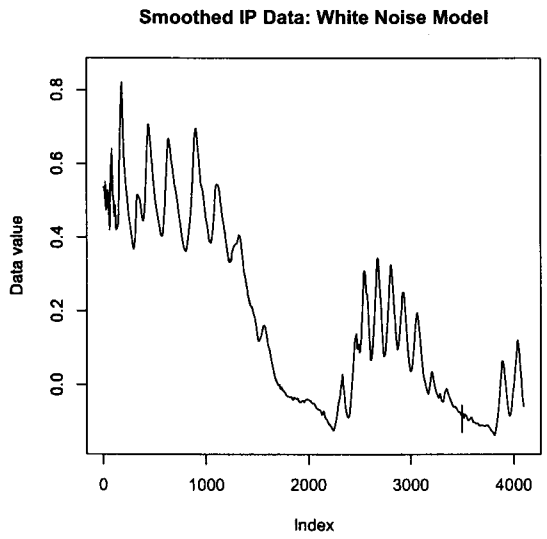

FIGURE 4.2 *Smoothed inductance plethysmography data, obtained by applying the EbayesThresh approach to each level of the wavelet transform, with the default option of the same noise standard deviation at all levels.*

Level	1	2	3	4	5	6
Threshold (as multiple of noise std dev)	4.08	3.91	3.16	2.29	0	0
Threshold (in absolute terms)	.044	.042	.034	.025	0	0

These thresholds are instructive. At the two finest scales of the transform, the threshold chosen is around four standard deviations, and is in each case the *universal* threshold $\sqrt{2 \log n}$ where n is the length of the vector of coefficients at the relevant level. This is the largest threshold that the method can choose, and is the value appropriate to a very sparse signal. On the other hand, at levels 5 and 6, the chosen threshold is zero, so that essentially no thresholding is carried out; this is the treatment appropriate for a signal that contains no zeroes at all. On the other hand at the intermediate levels 3 and 4, a threshold is chosen between these extremes, corresponding to the notion that the signal is moderately sparse.

Further insight can be gained by examining Figure 4.3. This gives a normal quantile plot of the wavelet coefficients at each level of the transform. In every case, the dashed line shows the expected plot that would be obtained if the relevant coefficients were all normally distributed with mean zero and standard deviation equal to the value $\hat{\sigma}_1$ estimated from the coefficients at level 1. (Ignore

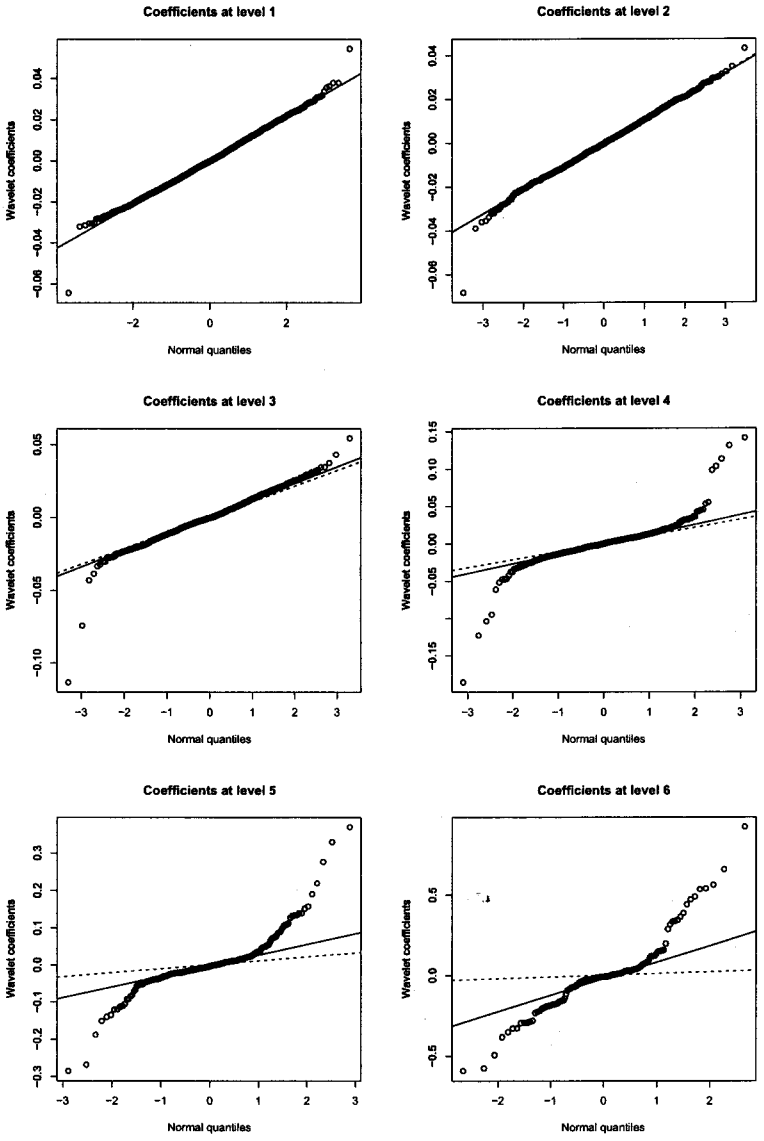

FIGURE 4.3 *Normal plots of the coefficients at each level of the discrete wavelet transform of the inductance plethysmography data. In each case the solid line gives the expected plot that would be obtained for a normally distributed data set with median absolute deviation equal to that of the given data; the dashed line gives the corresponding line for the median absolute deviation of the coefficients at the finest level.*

the solid lines for the moment.) It can immediately be seen that, if the noise in the data is assumed to be $N(0, \hat{\sigma}_1^2)$ at every level, it is reasonable to assume that virtually all the underlying signal values at levels 1 and 2 are zero, so that the observed data, except for a very small number of extreme values, come from the noise distribution. On the other hand, the dashed line is a poor fit at levels 5 and (especially) 6, even in the part of the distribution near zero, so it is reasonable that the empirical Bayes method chooses a prior probability of one that the signal values are nonzero. Finally, one can see the appropriateness of considering the signals at levels 3 and 4 to be mixtures of a mass at zero and a nonzero distribution. As one moves from coarser to finer levels, the treatment chosen by the method corresponds to increasing sparsity, and hence to a higher choice of threshold at finer levels.

Even if one did not constrain the noise at all levels to have the same standard deviation, the plots still indicate that the distributions are further from normal at the coarser levels. The solid lines show the expected plots that would be obtained if the data were normally distributed with standard deviation estimated separately at each level; the increasingly heavy tails of the observations at coarser levels are clear.

4.4. The stationary noise model

Johnstone and Silverman (1997) considered the use of wavelet thresholding methods for data where the original noise is stationary but correlated. They showed that an appropriate approach is to carry out wavelet thresholding as if the noise were independent, but to allow different noise variances at different levels. In our context, this would correspond to estimating the noise variance using the median absolute deviation separately at each level. If this is used then the resulting estimated thresholds are as follows:

Level	1	2	3	4	5	6
Threshold (as multiple of noise std dev)	4.08	3.91	3.40	2.61	2.07	1.70
Threshold (in absolute terms)	0.044	0.042	0.039	0.033	0.064	0.184

It can be seen that the treatment of the finest two levels is the same as previously, but that coarser levels are thresholded somewhat more severely (higher thresholds) than before, whether the thresholds are expressed in terms of the individual standard deviations or in absolute terms. Another interesting feature is the way that the signal is judged to be progressively less sparse as the scale becomes coarser, again bearing out the impression given by Figure 4.3.

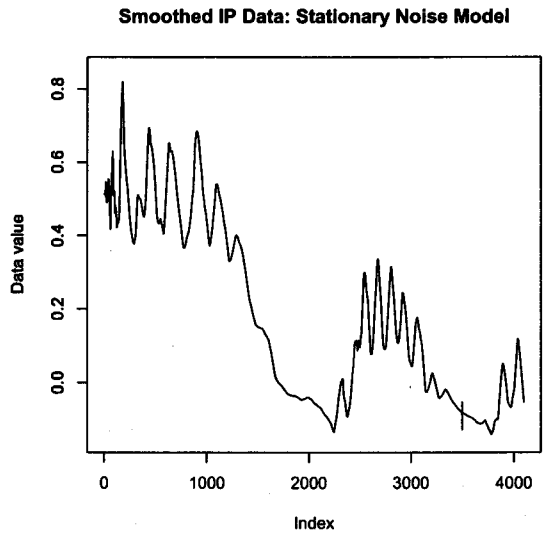

FIGURE 4.4 *Smoothed inductance plethysmography data, obtained by allowing different noise standard deviations at each level of the wavelet transform. This corresponds to an assumption of stationary correlated noise.*

A plot of the resulting estimate is given in Figure 4.4. A comparison between this estimate and the estimate based on a white noise error model is given in Figure 4.5. The first segment presented there is the one containing the highest peak in the data. There is little noticeable difference between the two estimates in this region, but if anything the peak is more sharply estimated in the stationary noise model. The second segment is one in which there is regular oscillatory variation at a fairly low frequency; the slight additional smoothness in the stationary noise estimate perhaps yields slightly preferable estimates. In the third short segment, including the high frequency glitch at time 3500, both methods retain the presumably spurious high frequency effect, but the stationary noise method removes the other local variability. Overall the stationary noise plots remove some moderately high frequency effects still present in the white noise plots.

The 'glitch' around point 3500 is caused by a single wavelet coefficient at the finest level taking a value six estimated standard deviations from zero, and this, and its partner in the reflected sequence of coefficients, are the only wavelet coefficients at the finest level that survives the thresholding; such a coefficient is

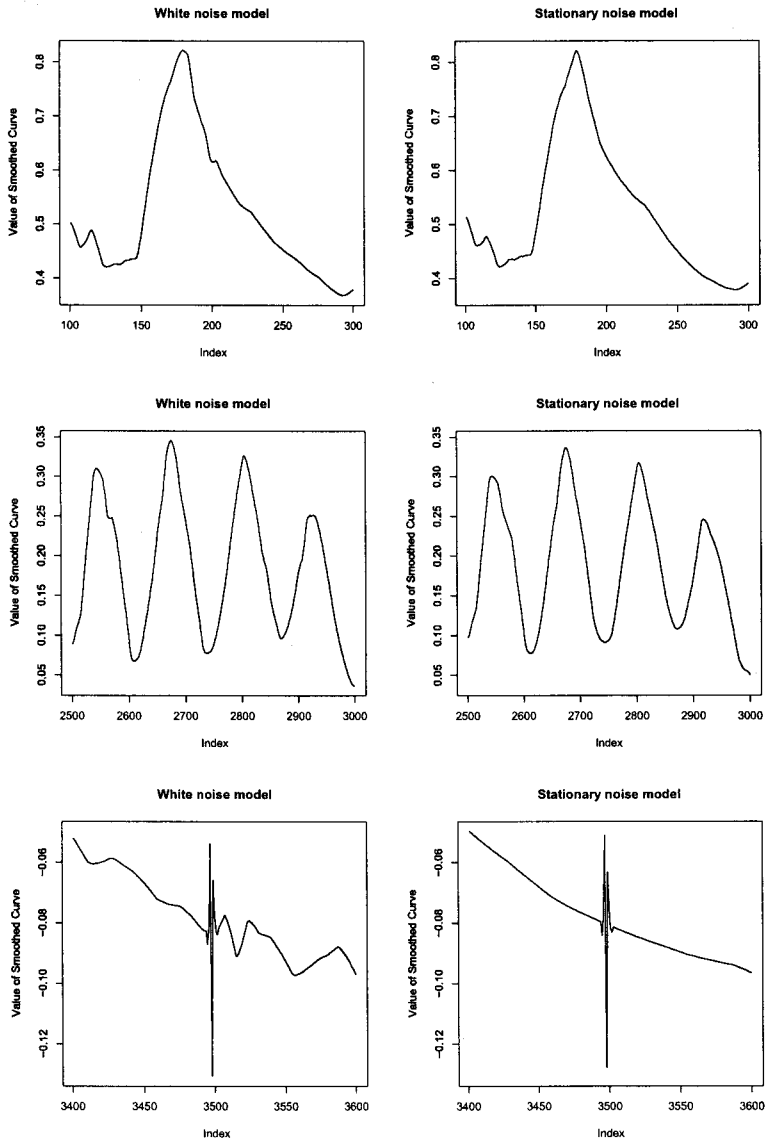

FIGURE 4.5 *Comparison between wavelet smoothing with a white noise model (left column of plots) for the error and a stationary noise model (right column), where the noise variance in the wavelet coefficients is estimated separately at each level. The comparison is made for various short segments of the data, intervals* $(101, 300)$, $(2501, 3000)$ *and* $(3401, 3600)$ *of the original index set.*

highly significant by any accounts. The numerical values of the observations in the interval [3491, 3510] are

```
[3491]  -.078  -.088  -.076  -.086  -.090
[3496]  -.083  -.054  -.142  -.081  -.071
[3501]  -.098  -.086  -.083  -.078  -.059
[3506]  -.086  -.090  -.073  -.086  -.076
```

Thus, observation 3497 is somewhat higher than its neighbours, and is immediately followed by the anomalously low observation $-.142$ at time 3498. Given that the instrumentation is generally more stable than this, one possible safeguard in future data analysis would be specifically to look out for outliers of this kind; a simple way of doing this would be to zero out all the wavelet coefficients at the finest level, in other words to use an infinite threshold, and the effect of this in the current plots would be just to remove the glitch.

4.5. The translation-invariant wavelet transform

It is generally recognized that improved smoothing results can often be obtained using the *translation-invariant* wavelet transform; see, for example, Coifman and Donoho (1995). This transform is also called the non-decimated, maximal overlap, or stationary wavelet transform. Given an original sequence of length N, this transform yields a sequence of N coefficients at each scale, rather than a pyramid of coefficient vectors whose length divides by two at successively coarser levels.

As discussed in detail in Johnstone and Silverman (2005b), the most straightforward way of applying the empirical Bayes approach is to threshold the coefficient vector at each level as if it were an independent sequence. To obtain the estimated curve, we then use the standard inversion algorithm for the translation-invariant wavelet transform.

The result of this procedure is plotted in Figure 4.6. It can be seen that the estimated curve is somewhat smoother than those obtained using the standard discrete wavelet transform. The high frequency effect at index 3500 is reduced in size. This is because, at the finest level, the inverse of the translation-invariant wavelet transform involves averaging the inverse of discrete transforms at two different positions, corresponding to basing the wavelets at odd or even positions in the original data sequence. The very large wavelet coefficient only occurs in one of these sequences, and so in the estimated curve the amplitude of the 'glitch' is halved.

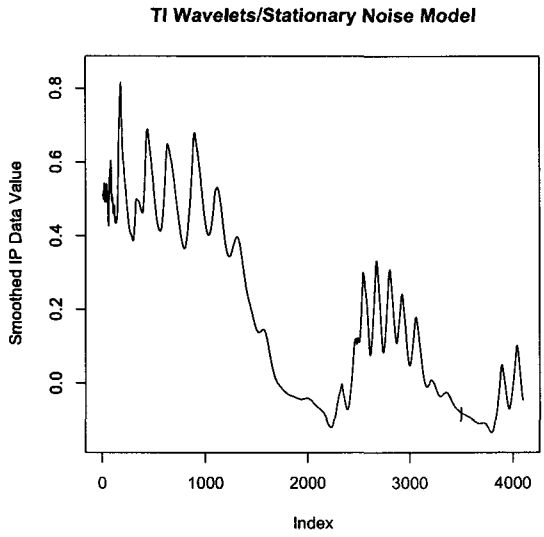

FIGURE 4.6 *Smoothed inductance plethysmography data, obtained from a translation-invariant wavelet transform and then applying the EbayesThresh approach estimating the noise variance separately at each level.*

4.6. Smoothing an image

We now move to the consideration of the possible use of the method for the processing of the wavelet transform of a two-dimensional image. The example we use will be the image of Ingrid Daubechies contained in the `waveslim` package in R.

Especially when processing images, it may be appropriate to use dictionaries other than the standard two-dimensional wavelet transform. Therefore this section should be read in a 'tutorial' way; its purpose is not to set out a black box recipe, but to illustrate how the basic EbayesThresh approach can be used in a broader context.

The standard deviation of the original image is about 35 and for this example we use noise with standard deviation 10. Fuller details are given in Johnstone and Silverman (2005a). The two-dimensional wavelet expansion of an image yields three arrays of coefficients at each level. We consider the finest four levels, therefore smoothing twelve arrays of coefficients altogether, but preserving the matrix of 'scaling' coefficients at the coarsest level.

In order to estimate the noise standard deviation from the data, we apply the

median absolute deviation function to the all the wavelet coefficients at level 1 (combining the three arrays for this purpose). This gives the result 10.25, very close to the theoretical value 10. We apply the EbayesThresh approach to the twelve matrices of coefficients with this noise standard deviation.

It is of interest to consider the thresholds estimated by the method; these are given as follows. In this list, the letters L and H indicate whether the coefficients result from a low or high pass filter, in the x and y directions respectively. The number refers to the level, with 1 being the finest level. The scaling coefficients at the coarsest scale would correspond to an entry LL4, which is not shown.

```
 LH1  HL1  HH1  LH2  HL2  HH2  LH3  HL3  HH3  LH4  HL4  HH4
4.41 3.81 4.41 2.42 2.58 3.26 1.01 0.92 1.85 0.00 0.00 0.00
```

These thresholds are interesting. There is no thresholding at level 4. At the finest level 1, on the other hand, the LH and HH coefficients are thresholded at the universal threshold $\sqrt{2\log(128 \times 128)}$ for data sets of their size, while the HL coefficients are subject to thresholding nearly as stringent. At levels 2 and 3 the thresholding applied to the HH coefficients is higher than that applied to the LH or HL coefficients. It is reasonable to consider the HH2 coefficients as being at a level intermediate between level 1 and 2, and the HH3 as being intermediate between levels 2 and 3, and so on; with this convention, the estimated thresholds increase monotonically as one moves from coarser to finer levels.

Figure 4.7 shows the original image, the noisy image, and the estimate of the image as obtained using EbayesThresh as set out in this section. In addition, we show a kernel smooth with bandwidth parameter adjusted to minimize the L_1 distance between the estimate and the original image `dau`; the average L_1 error of both the kernel smooth and the fully automatic EbayesThresh wavelet smooth is 3.07.

In the kernel smooth, there is considerable remaining random error in the flat parts of the plot, and some of the highlights and details are slighly more smoothed out than in the wavelet plot. On the other hand, in the wavelet plot there are some spurious artefacts. In addition, it is encouraging that the EbayesThresh method has succeeded in automatically achieving the L_1 error of the kernel estimate chosen by reference to the true image. Furthermore, it is now well understood that the standard two-dimensional wavelet transform is not a very good dictionary for the representation of images. In principle, our empirical Bayes approach is equally applicable whatever the transform used, and will take advantage of sparsity in the representation of the function or image being estimated, and so if a dictionary

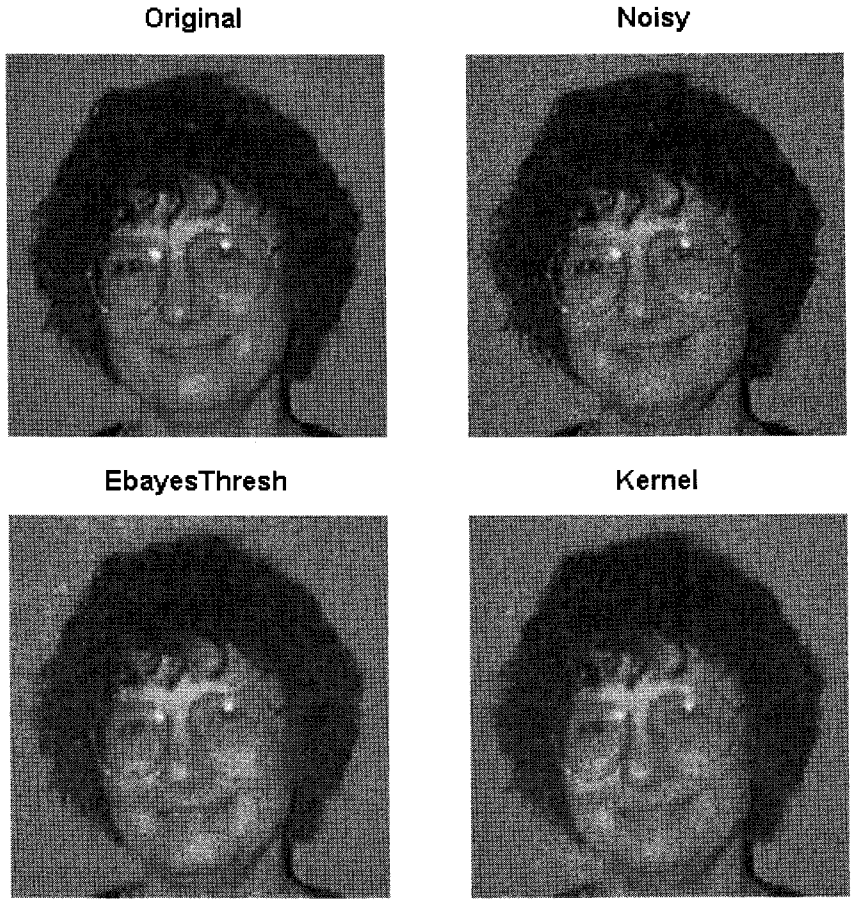

FIGURE 4.7 *Top left: original image of Ingrid Daubechies; top right: effect of adding normal independent noise; bottom left: result of applying the empirical Bayes smoothing method to the individual matrices of coefficients; bottom right: kernel smooth of noisy image, with bandwidth chosen to minimize the average absolute error.*

more specifically suited to the representation of this type of image were used, one could expect even better results.

5. THEORETICAL RESULTS

In this section, we review some of the theoretical properties of the EbayesThresh method. We first consider the estimation of a single sequence μ_i of parameters from a sequence of observations X_i distributed independently $N(\mu_i, 1)$. The theory explores how well the sequence μ_i is estimated by the EbayesThresh proce-

dure, and compares this quality of estimation with the rate that can be obtained by the best possible estimator under given assumptions.

5.1. *Sparsity*

The sparsity of a signal is not just a matter of the proportion of μ_i that are zero or very near zero, but also of more subtle ways in which the energy of the signal μ is distributed among the various components. Our theory will demonstrate that the empirical Bayes choice of estimated threshold yields a highly adaptive procedure, with excellent properties for a wide range of conditions on the underlying signal.

A natural notion of sparsity is the possibility that μ is a *nearly black* signal, in the sense that the number of indices i for which μ_i is nonzero is bounded. We define

$$\ell_0[\eta] = \{\mu : n^{-1} \sum_{i=1}^{n} I[\mu_i \neq 0] \leq \eta\}. \tag{5.1}$$

A more subtle characterization of sparsity will not require any μ_i to be exactly zero, but still control the concentration of the energy of the signal by placing bounds on the p-norm of μ for $p > 0$. In this case, we suppose the signal is belongs to an ℓ_p norm ball of small radius η,

$$\ell_p[\eta] = \{\mu : n^{-1} \sum |\mu_i|^p \leq \eta^p\}. \tag{5.2}$$

For reasons set out in Johnstone and Silverman (2004a), for $0 < p < 2$, among all signals with a given energy, the sparse ones are those with small ℓ_p norm.

5.2. *Quality of estimation*

We shall compare the EbayesThresh estimator with the minimax estimators over sets of sequences of particular sparsity. The estimator that attains the ideal performance over a nearly black class, or over an ℓ_p ball for some $p > 0$, will in general depend on p and on η. The minimax rate is a benchmark for the estimation of signals that display the sparseness characteristic of membership of an ℓ_p class. Our main theorem will show that, under mild conditions, an empirical Bayes thresholding estimate will essentially achieve the minimax rate over η simultaneously for all p in $[0, \infty]$, including the nearly black class as the case $p = 0$.

It should also be added that we do not restrict attention to losses based on squared errors, but we can measure risk by the average expected q^{th} power loss

for any q in $(0, 2]$,

$$R_q(\hat{\mu}, \mu) = n^{-1} \sum_{i=1}^{n} E|\hat{\mu}_i - \mu_i|^q. \tag{5.3}$$

We set two goals for estimation using the empirical Bayes threshold: 'uniform boundedness of risk', and 'flexible adaptation'. To explain what we mean by flexible adaptation, suppose that the signal is sparse in the sense of belonging to an ℓ_p norm ball $\ell_p[\eta]$ as defined in (5.2). As before, we include nearly black classes as the case $p = 0$. If the radius η is small, we would hope that the estimation error $R_q(\hat{\mu}, \mu)$ should be appropriately small. How small is benchmarked in terms of the minimax risk

$$R_{n,q}(\ell_p[\eta]) = \inf_{\hat{\mu}} \sup_{\mu \in \ell_p[\eta]} R_q(\hat{\mu}, \mu).$$

Suppose $\eta = \eta_n \to 0$ as $n \to \infty$ but that, in the case $q > p > 0$,

$$n^{-1/p} \eta^{-1} (\log \eta^{-p})^{\frac{1}{2}} \to 0, \tag{5.4}$$

which prevents η from becoming very small too quickly. (For $p = 0$ we require $n\eta \to \infty$.) Then (Donoho and Johnstone, 1994a, with slight modifications) we have the asymptotic relation

$$R_{n,q}(\ell_p[\eta_n]) \sim r_{p,q}(\eta_n) \qquad \text{as } n \to \infty, \tag{5.5}$$

where

$$r_{p,q}(\eta) = \begin{cases} \eta^q & 0 < q \le p \\ \eta^p (2 \log \eta^{-p})^{(q-p)/2} & 0 < p < q \\ \eta (2 \log \eta^{-1})^{q/2} & p = 0, q > 0. \end{cases} \tag{5.6}$$

The main theoretical result, stated in detail and proved in Johnstone and Silverman (2004a), gives comparable bounds on the risk function of the empirical Bayes thresholding procedure. Apart from an error of order $n^{-1}(\log n)^{2+(q-p\wedge 2)/2}$, the procedure uniformly attains the same error rate as the minimax estimator for all p in $[0, \infty]$ and q in $(0, 2]$. The result assumes that the prior function γ in the EbayesThresh procedure has tails that are at least as heavy as an exponential distribution and no heavier than Cauchy, so that both the Laplace and quasi-Cauchy estimators are included. There is considerable freedom about the choice of estimation rule. Apart from the posterior mean, all the rules mentioned above can be used and the result remains valid. (The posterior mean also works, but only for a restricted range for the parameter q.)

THEOREM 5.1. *Under appropriate assumptions on the estimation procedure set out in detail in Johnstone and Silverman (2004a), suppose that* $X \sim N_n(\mu, I)$, *that* $\delta(x,t)$ *is a thresholding rule with threshold* t *and that* $0 \leq p \leq \infty$ *and* $0 < q \leq 2$. *Let* $\hat{w}$ *be the weight chosen by marginal maximum likelihood within the EbayesThresh paradigm, and let* $\hat{t} = t(\hat{w})$, *where* $t(w)$ *denotes the threshold of the posterior median rule corresponding to the prior weight* w. *Then the estimator* $\hat{\mu}_i(x) = \delta(x_i, t(\hat{w}))$ *satisfies*

(a) (Uniformly bounded risk) There exists a constant $C_0(q)$ *such that*

$$\sup_{\mu} R_q(\hat{\mu}, \mu) \leq C_0.$$

(b) (Adaptivity) There exist constants $C_i(p,q)$ *such that for* $\eta \leq \eta_0(p,q)$ *and* $n \geq n_0(p,q)$

$$\sup_{\mu \in \ell_p[\eta]} R_q(\hat{\mu}, \mu) \leq C_1 r_{p,q}(\eta) + C_2 n^{-1} (\log n)^{2+(q-p\wedge 2)/2}. \tag{5.7}$$

When $q \in (1,2]$, *these results also hold for the posterior mean estimate* $\tilde{\mu}$.

We emphasize that it is not necessary that $\delta(x,t)$ be derived from the posterior median or mean rule. It might be hard or soft thresholding or some other nonlinearity with the stated properties. The point of the theorem is that empirical Bayes estimation of the threshold parameter suffices with all such methods to achieve both adaptivity and uniformly bounded risk.

From the theorem, it can be concluded that, for every p in $(0, \infty]$ and q in $(0,2]$, and for the nearly black case $p = 0$, our estimator attains the optimal q-norm risk (5.6), up to a constant multiplier, for all sufficiently large n and for η satisfying $n^{-1} \log^2 n \leq \eta^{p\wedge 2} \leq \eta_0^{p\wedge 2}$ if $p > 0$ and $n^{-1} \log^2 n \leq \eta \leq \eta_0$ if $p = 0$. In this sense it adapts automatically to the degree and character of sparsity of the signal in the optimum possible way over an extremely wide range of signal classes.

5.3. Wavelet estimators

The paper Johnstone and Silverman (2005b) explores in detail several theoretical aspects of the EbayesThresh approach to function estimation using wavelets. In this section, one of the key results is reviewed briefly; for full details see the original paper.

Suppose we have noisy observations of a function f at a grid of $N = 2^J$ points,

$$Z_i = f(i/N) + \epsilon_i \qquad \epsilon_i \text{ independent } N(0,1).$$

Assume that we construct an estimator by carrying out a discrete wavelet transform of the vector of observations Z_i, applying EbayesThresh to the vector of coefficients at each level, and then transforming back.

The paper finds overall bounds on the risk of the method subject to membership of the unknown function in one of a wide range of Besov classes, covering also the case of f of bounded variation. The rates obtained are optimal for any value of the parameter p in $(0, \infty]$, simultaneously for a wide range of loss functions, each dominating the L_q norm of the σth derivative, with $\sigma \geq 0$ and $0 < q \leq 2$.

The basic result of the paper is as follows. Let $d_{jk} = N^{\frac{1}{2}}\theta_{jk}$ be the coefficients of an orthogonal discrete wavelet transform of the sequence $f(t_i)$, and let d_j denote the vector with elements d_{jk} as k varies. For $0 < p \leq \infty$ and $\alpha > (1/p) - (1/2)$, let $a = \alpha - (1/p) + (1/2)$. Define the Besov sequence space $b^{\alpha}_{p,\infty}(C)$ to be the set of all coefficient arrays θ such that

$$\sum_k |\theta_{jk}|^p < C^p 2^{-apj} \text{ for all } j \text{ with } L - 1 \leq j < J. \tag{5.8}$$

Our theory shows that for some constant c, possibly depending on p and α but not on N or C,

$$\sup_{\theta \in b^{\alpha}_{p,\infty}(C)} N^{-1} E \sum_{i=1}^{N} \{\hat{f}(t_i) - f(t_i)\}^2 \leq c\{C^{2/(2\alpha+1)} N^{-2\alpha/(2\alpha+1)} + N^{-1}(\log N)^4\}. \tag{5.9}$$

For fixed C, the second term in the bound (5.9) is negligible, and the rate $O(N^{-2\alpha/(2\alpha+1)})$ of decay of the mean square error is the best that can be attained over the relevant function class. The result (5.9) thus shows that, apart from the $O(N^{-1} \log^4 N)$ term, our estimation method simultaneously attains the optimum rate over a wide range of function classes, thus automatically adapting to the regularity of the underlying function. Under appropriate conditions, the Besov sequence space norm used in (5.8) is equivalent to a Besov function space norm on f with the same parameters.

The main theorem of the paper goes considerably beyond (5.9), in the following respects:

- It demonstrates the optimal rate of convergence for mean q-norm errors for all $0 < q \leq 2$, not just the mean square error considered in (5.9).
- Beyond the posterior median, any thresholding method satisfying certain mild conditions can be used, and for $1 < q \leq 2$ the results also hold for the posterior mean.

- If an appropriate modified threshold method is used, the optimality also extends to the estimation of derivatives of f.

Overall, as for the single sequence case, the EbayesThresh estimates demonstrate remarkable adaptivity when applied in the wavelet context. The Besov scale of functions encompasses a very wide range of function behaviour, and what the EbayesThresh procedure does is to take advantage of inhomogeneity when it can. It would be natural to assume that the choice of a good estimator would depend strongly on the properties of the underlying function, but our theory demonstrates that the EbayesThresh estimator can automatically deal optimally with function classes of many kinds.

6. CONCLUDING REMARKS

The empirical Bayes method set out in this paper and implemented in the `EbayesThresh` package has wide potential applicability. There are increasingly many contexts where, implicitly or explicitly, one is estimating a large number of parameters and it is necessary or advisable to take advantage of possible sparsity in the parameter set. While it was the wavelet context that gave the authors the original motivation for investigating this methodology, both theoretical and practical considerations have already shown that it has much wider relevance, though of course there are many aspects that raise interesting topics for future research. I am extremely grateful to the Korean Statistical Society for inviting me to present this work.

REFERENCES

ANTONIADIS, A., JANSEN, M., JOHNSTONE, I. M. AND SILVERMAN, B. W. (2004). **EbayesThresh**: MATLAB software for Empirical Bayes thresholding, `http://www-lmc.imag.fr/lmc-sms/Anestis.Antoniadis/EBayesThresh`.

COIFMAN, R. R. AND DONOHO, D. L. (1995). *Translation-invariant de-noising*, (A. Antoniadis, ed), Wavelets and Statistics, Lecture Notes in Statistics, Springer.

DONOHO, D. L. AND JOHNSTONE, I. M. (1994a). "Minimax risk over ℓ_p-balls for ℓ_q-error", *Probability Theory and Related Fields*, **99**, 277–303.

DONOHO, D. L. AND JOHNSTONE, I. M. (1994b). "Idle spatial adaptation by wavelet shrinkage", *Biometrika*, **81**, 425–455.

JANSEN, M., NASON, G. P. AND SILVERMAN, B. W. (2004). "Multivariate nonparametric regression using lifting", *Technical Report 04:17*, Department of Mathematics, University of Bristol, `http://www.maths.bris.ac.uk/research/stats/pub/ResRept/2004.html`.

JOHNSTONE, I. M. AND SILVERMAN, B. W. (2004a). "Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequences", *The Annals of Statistics*, **32**, 1594–1649.

JOHNSTONE, I. M. AND SILVERMAN, B. W. (2005a). "EbayesThresh: R programs for empirical Bayes thresholding", *Journal of Statistical Software*, **12**, 1–38.

JOHNSTONE, I. M. AND SILVERMAN, B. W. (2005b). "Empirical Bayes selection of wavelet thresholds", *Annals of Statistics*, **33**, 1700–1752.

Johnstone, I. M. and Silverman, B. W. (2004b). "Boundary coiflets for wavelet shrinkage in function estimation", *Journal of Applied Probability*, **41A**, 81–98.

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.

NASON, G. P. (1996). "Wavelet shrinkage using cross-validation", *Journal of the Royal Statistical Society*, Ser. B, **58**, 463–479.

NASON, G. P. (1998). *WaveThresh3 Software*, Department of Mathematics, University of Bristol, Bristol, U.K.

SILVERMAN, B. W. (2005). **EbayesThresh**: Empirical Bayes thresholding and related methods, CRAN.