Wavelet Thresholding via MDL for Natural Images

Mark Hansen[1] and Bin Yu[1,2*]

Abstract

We study the application of Rissanen's Principle of Minimum Description Length (MDL) to the problem of wavelet denoising and compression for natural images. After making a connection between thresholding and model selection, we derive an MDL criterion based on a Laplacian model for noiseless wavelet coefficients. We find that this approach leads to an adaptive thresholding rule. While achieving mean squared error performance comparable with other popular thresholding schemes, the MDL procedure tends to keep far fewer coefficients. From this property, we demonstrate that our method is an excellent tool for simultaneous denoising and compression. We make this claim precise by analyzing MDL thresholding in two optimality frameworks; one in which we measure rate and distortion based on quantized coefficients and one in which we do not quantize, but instead record rate simply as the number of non-zero coefficients.

Index Terms – Compression, denoising, Laplacian distribution, MDL, natural images, statistical estimation, wavelet thresholding.

[*1] Bell Laboratories, Murray Hill, NJ and [2] University of California at Berkeley, Berkeley, CA

I Introduction

Donoho and Johnstone [13, 12, 14] proposed the use of wavelet *thresholding* for denoising 1-dimensional curves observed with additive, white noise. Their schemes are shown to be (essentially) minimax optimal in terms of mean squared error (MSE) over large classes of functions, most notably Besov spaces. Their soft-thresholding rules *VisuShrink* and *SureShrink* (which differ in the choice of threshold) remove noise from a signal by explicit setting "small" wavelet coefficients to zero; a form of high-level compression which we will call *analytical compression.* A single threshold parameter determines the behavior of these procedures, setting both the level below which coefficients are eliminated as well as determining how the remaining coefficients are to be estimated. Since the pioneering work of Donoho and Johnstone, many variants and improvements of their thresholding rules have appeared in the literature on statistical curve estimation (e.g. [6, 1, 7, 8, 28, 36]).

In this article we consider so-called *natural images.* Extensive empirical work has led to the characterization that the wavelet coefficients derived from noiseless natural signals approximately follow a Laplacian or generalized Gaussian distribution [19, 20, 29, 30]. For this class of signals, it is well known that the universal threshold $\sigma\sqrt{2\log n}$ used by *VisuShrink* eliminates too many coefficients, while a variant of the *SureShrink* procedure, known as *Sure*, works reasonably well. (We will describe *Sure* in Section II). Bayesian approaches that make use of the distributional characterization for natural images have yielded soft-thresholding rules that match the performance of *Sure* (cf.[3, 26, 24]). This similarity should be expected given that these schemes and *Sure* both attempt to minimize the same Bayes risk [3].

In general, wavelet thresholding can be thought of as a special case of statistical model selection where we have as many (orthogonal) predictor variables (corresponding to wavelet basis elements) as there are data points [22, 23, 32, 7, 8]. Viewing the problem in this way is attractive because it allows one to separate the "kill" action (setting coefficients to zero) from the "keep" action (estimating the remaining coefficients via shrinkage or some other procedure). In this article, we fold the prior distributional assumptions for natural images into a model selection framework for wavelet denoising via Rissanen's Principle of Minimum Description Length (MDL). Several applications of (two-stage) MDL to wavelet thresholding for images have appeared previously in the literature. Moulin [22, 23] and Saito [32], for example, derive thresholds similar in form to that employed by *VisuShrink*; while Moulin and Liu [24] employ Rissanen's universal prior on integers to construct an MDL criterion that has met with success in denoising natural images.

Our focus is on MDL thresholding rules for both *compressing and denoising* natural images. We introduce a criterion, $lMDL$, obtained from a Laplacian prior. We compare $lMDL$ with other thresholding rules based on two simultaneous denoising and compression optimality criteria; one in which we measure rate and distortion based on quantized coefficients and one in which we do not quantize, but instead record rate simply as the number of non-zero coefficients. We take as our benchmarks the *BayesShrink* approximate MSE-optimal, soft-threshold of Chang, Yu and Vetterli [3]; and the maximum a posteriori threshold based on a Laplacian prior [26, 24]. The former is known to give a slightly better performance than *Sure* and is much simpler to use. Our MDL procedures achieve comparable MSE performance, while keeping far fewer (non-zero) coefficients. All the MDL procedures and their comparative counterparts in this paper are based on the assumption that the wavelet coefficients from a given subband are a simple random sample from some distribution. Within the MDL paradigm more elaborate dependence structures (both within and between subband) could be incorporated.

The rest of the paper is organized as follows. In Section II we introduce the additive, white noise model in the wavelet domain and review two competing views of estimation under this framework. In Section III, we introduce $lMDL$ and present a small experiment to compare it with other thresholding rules. Section IV discusses issues of simultaneous denoising and compression under the two optimality frameworks mentioned above. In Section V we conclude and present open problems. The technical details and proofs of the results in Section III are collected in appendices at the end of the paper. Throughout our presentation, we will use standard terminology from wavelet image subband coding (cf. [35]).

II Denoising via Thresholding and Model Selection

Wavelet thresholding for signal denoising sets "small" coefficients to zero, yielding an analytical compression of the signal (cf. [13, 12, 14, 11, 5]). In a series of papers, Donoho and Johnstone [13, 12, 14] assume the following additive, white noise model in the wavelet domain:

$$y_i = \beta_i + \epsilon_i, \qquad i = 1, \dots, n \tag{1}$$

where the β_i are wavelet coefficients of the signal and the ϵ_i are iid $N(0, \sigma^2)$. The error criterion is taken as the average pixelwise MSE; that is, for any set of estimators $\hat{\beta}_i$,

$$MSE = \frac{1}{n} \sum_i (\beta_i - \hat{\beta}_i)^2. \tag{2}$$

Assuming a fixed, orthonormal wavelet transform (which we will use throughout this paper), the equivalent model in the spatial or time domain also involves additive, white Gaussian noise and the error measure remains MSE. Model (1) is applied separately to each subband, so that the "sample size" n refers to the number of pixels in a particular subband (usually a power of 2). All experimental results in this paper are based on a 3-level orthonormal wavelet decomposition using Daubechies' 8-symmelet (cf. [10, 20, 35]), and thresholding is applied only to the "detail" subbands HL, LH and HH at each level. Coefficients in LL(3) are estimated by their noisy counterparts. For simplicity, we assume that σ^2 is known and equal to one since it can be easily estimated from the finest or most detailed subband HH(1) (cf. [13, 14, 6, 36]). Once estimated, thresholding procedures can be applied to the standardized data subband-by-subband.

Recall the soft-thresholding function

$$f_s(y, T) \;:=\; \mathrm{sign}\,(y) \max\left(\,|y| - T, 0\,\right). \tag{3}$$

From this expression, it is clear that the performance of soft-thresholding is controlled by a single parameter T that simultaneously specifies the level below which coefficients are set to zero and the amount of shrinkage applied to those that are kept. Donoho and Johnstone's [13] procedure *VisuShrink* sets a soft-threshold at $\sigma\sqrt{2\log n}$ depending on the size of the subband n and is shown to be minimax optimal over Besov spaces. A more-data driven, soft-threshold, *SureShrink*, was also proposed and shown to be minimax optimal over signals with unknown smoothness (belonging to one of a range of Besov spaces). To be more specific, recall Stein's unbiased estimate of risk [33]

$$n - 2\sum_{i=1}^{n} I_{\{i:|y_i| \le t\}} + \sum_{i=1}^{n} (|y_i| \wedge t)^2. \tag{4}$$

Letting T denote the value that minimizes (4), Donoho and Johnstone [14] define the *SureShrink* soft-threshold T^* to be

$$T^* = \min(T, \sigma\sqrt{2\log n}). \tag{5}$$

The value T (without the $\sigma\sqrt{2\log n}$ bound) has also been used in wavelet denoising and is commonly referred to simply as the *Sure* threshold.

As mentioned in the introduction, the relationship in (1) describes a regression model with an orthogonal design matrix having the same number of regressors as the number of observations. Model selection in this context explicitly sets coefficients to zero and hence also produces an analytical compression of the observed signal. To fix notation, we introduce a binary vector $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_n) \in \{0,1\}^n$ as an index for the 2^n models, $\mathcal{M}_\gamma$. The elements of γ specify which variables are included in the model (i.e. have non-zero coefficients) so that (1) becomes

$$y_i = \begin{cases} \beta_i + \epsilon_i, & \gamma_i = 1 \\ \epsilon_i, & \gamma_i = 0 \end{cases} \tag{6}$$

In postulating these models, we follow the philosophy underlying Grenander's method of sieves sieves [16]; that is, models are viewed as (at best) approximations to the underlying data-generating mechanism and that these approximations can be chosen to possess a certain desirable property, which in our case is analytical compression. Under model $\mathcal{M}_\gamma$, those β_i for which $\gamma_i = 1$ are estimated from the noisy observations (6). By casting wavelet thresholding as a model selection problem, we separate the choice of which coefficients to discard from the estimation procedure applied to those that remain. This property will prove advantageous in terms of MSE performance.

Having made the connection between thresholding and model selection, we will apply MDL to derive effective denoising procedures for natural images in Section III. MDL provides a framework for statistical modeling in general and model selection in particular, and has been applied previously to the problem of wavelet thresholding. Moulin [22] and Saito [32] proposed (two-stage) MDL schemes that lead to thresholds $T = \sigma\sqrt{\log n}$ and $T = \sigma\sqrt{3\log n}$ respectively. They tend to aggressively select small models, a behavior similar to that observed with the universal threshold used in *VisuShrink*.

For natural images, it is well-known that *VisuShrink*, as well as the MDL schemes of Moulin and Saito, set far too many coefficients to zero (cf. the end of Section III) while the *Sure* threshold works well. Wavelet denoising for natural images is studied by [3, 24]. Both references take an empirical Bayes approach by assigning a data-driven prior to the coefficients in each subband. The former gives a simple closed-form approximation to the MSE-optimal, soft-threshold and explains the favorable performance of the *Sure* procedure when denoising natural images. The latter concentrates on the maximum *a posteriori* (MAP) method and presents a number of elaborate MDL (pixelwise) denoising schemes.

III An MDL Criterion for Wavelet Thresholding

In the previous section, we presented the connection between wavelet thresholding and variable selection for a normal linear regression model. Under this framework, we accomplish an analytical compression of the observed signal by positing descriptions (probability models) of the form (6) that explicitly set some subset of coefficients to zero. As a collection, the candidate model classes $\mathcal{M}_\gamma$ are viewed as (at best) approximations to the true data-generating distribution, each class having the desirable property that they afford us some degree of compression (depending on the number of nonzero entries in γ). We apply Rissanen's Principle of Minimum Description Length to choose among members of this collection. Before deriving the MDL selection criterion, we introduce a population model for the coefficients.

A *The Laplacian Population Model*

The family of generalized Gaussian distributions has been proposed and accepted for modeling noiseless subband data calculated from natural images (e.g. [19, 29, 30]). For analytical tractability, many authors have focused their attention on a single density in this family, the so-called Laplacian or double exponential:

$$\text{Lap} : \frac{\lambda}{2} e^{-\lambda|x|}, \quad x \in \mathbb{R}. \tag{7}$$

In image compression studies, this simplification does not usually produce a noticeable degradation in performance (e.g. [37]), and so we also adopt a Laplacian model. We assume that the coefficients for a given subband are an iid sample from the distribution (7), where λ is to be determined for each subband separately. This representation is only applied to the "detail" subbands HL, LH and HH at each level. We have chosen to treat the wavelet coefficients as iid samples mainly for computational reasons. Of course, under the MDL framework it is possible incorporate both within- and between-subband dependencies in γ. We comment on these more elaborate descriptions in Section V. Finally, recall that we will focus only on the case when $\sigma^2 = 1$. As noted earlier, σ^2 can be estimated from the HH(1) subband, at which point we can work with standardized data.

Proposition 1: Assume the wavelet coefficients computed for a given subband follow model (1) with $\sigma^2 = 1$. That is, the noisy wavelet coefficients are generated according to

$$y_i = \beta_i + \epsilon_i, \qquad i = 1, \dots, n$$

where the ϵ_i are iid $N(0,1)$. Also, assume that the coefficients β_i are iid according to (7). Let Φ and ϕ be the cumulative distribution and probability density functions of the standard normal distribution $N(0,1)$ respectively, and set

$$r_+(y;\lambda) = e^{(y-\lambda)^2/2}\Phi(y-\lambda) + e^{(y+\lambda)^2/2}\Phi(-y-\lambda)$$

and

$$r_-(y;\lambda) = e^{(y-\lambda)^2/2}\Phi(y-\lambda) - e^{(y+\lambda)^2/2}\Phi(-y-\lambda).$$

Then, the marginal density of y is

$$m_\lambda(y) = \frac{\lambda}{2}\sqrt{2\pi}\phi(y)r_+(y;\lambda).$$

Moreover, the conditional (posterior) mean of β given y is

$$\hat{\beta} = y - \lambda r_-(y;\lambda)/r_+(y;\lambda).$$

Proof: See Appendix A.

Under the assumption that the noiseless wavelet coefficients are iid samples from some distribution, we can replace the empirical MSE (2) with a population-level expression

$$MSE \;=\; E_{\beta,y}\,(\beta-\hat{\beta})^2\,.$$

where throughout this paper $\hat{\beta} = \hat{\beta}(y)$ is obtained by applying some type of thresholding operation to a noisy wavelet coefficient y. In the case of ordinary soft thresholding (3), for example, this becomes $\hat{\beta} = f_s(y,T)$ for some value of the threshold T. Based on numerical calculations with the generalized Gaussian distribution, Chang et al. [3] propose an approximate MSE-optimal soft-threshold *BayesShrink* $T_{bayes} = \sigma^2/\sigma_\beta$, where the signal power or second moment σ_β^2 is estimated by the truncated moment estimator $\max(0, \frac{1}{n}\sum_i y_i^2 - \sigma^2)$. In the Laplacian case, this becomes $T_{bayes} = \lambda/\sqrt{2}$ and we estimate it by plugging in the MLE of λ. In the next section, we will derive the MLE for λ based on the noisy subband data; see equation (11). On the other hand, it is known that the exact MAP soft-threshold under the Laplacian prior is $T_{map} = \lambda$ (cf. [26, 24]), which we call *MapShrink*. Using Stein's lemma and the derivations above, we can find the equation that the exact MSE-optimal threshold must satisfy.

Corollary 1: The exact MSE-optimal soft-threshold t satisfies the following equation

$$m_\lambda(t) = t\int_t^\infty m_\lambda(x)dx,$$

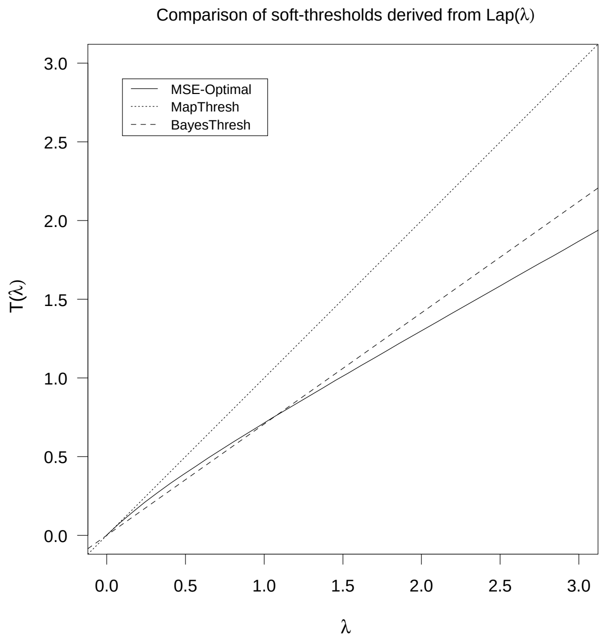

Figure 1: Comparing several soft-thresholds based on a Lap(λ) distribution for the noiseless wavelet coefficients. As $\lambda \to 0$, the MSE-optimal solution in Corollary 1 approaches T_{map}; while for large λ, the optimal solution is smaller than even T_{bayes}.

which simplifies to

$$\frac{t}{\lambda} = \frac{e^{\lambda t}\Phi(t-\lambda) + e^{\lambda t}\Phi(-t-\lambda)}{e^{-\lambda t}\Phi(t-\lambda) - e^{\lambda t}\Phi(-t-\lambda) + 2e^{\lambda^2/2}\Phi(-t)}.$$

Proof: See Appendix B.

This equation does not have a closed-form solution, but it is a simple non-linear equation which can easily be solved numerically. When λ tends to zero, or the signal-to-noise ratio gets high, t tends to zero, and can be shown to be of order $\lambda(1 + o(1))$ by Taylor expansions. In Fig. 1 we plot the optimal threshold as a function of λ and compare it to *BayesShrink* and *MapShrink*.

B *Model Classes and Analytical Compression*

As a principle, MDL suggests that we select a model or model class that yields the *shortest description* of a dataset. Two recent review articles are [2, 17]: the first geared toward an audience versed in information theory and the second written for the statistics community. The MDL philosophy is *descriptive* in the sense that models are viewed as a means of expressing properties evident in data, to paraphrase Rissanen (1989, p. 4). In our case, we choose a collection of models that set some (hopefully many) of the noisy wavelet coefficients to zero, ultimately achieving a simultaneous (analytical) compression and denoising. These models are not taken to be the true or data-generating distribution. For that we have constructed (7). As mentioned before, this use of models is similar to the ideas behind the method of sieves [16] (although popular applications of sieves have been mainly in the area of frequentist nonparametric function estimation, and our context is empirical Bayes).

In implementing an MDL procedure, we must specify a *description* or *code length* corresponding to each approximating model. To provide a valid selection criterion (i.e., one that yields a consistent or prediction-optimal procedure), we restrict our attention to optimal universal coding schemes based on model classes [27, 2, 17]. Distributional observations like the (approximate) Laplacian behavior of wavelet coefficients for natural signals can be easily incorporated into our MDL formulation.

As in the previous section, let $\beta_1, \ldots, \beta_n$ represent the noiseless wavelet coefficients for a given subband. The model classes $\mathcal{M}_\gamma$, indexed by the binary vector $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_n) \in \{0,1\}^n$, explicitly set some coefficients to zero, achieving an analytical compression. The density of each noiseless coefficient is then given by

$$f(\beta_i \,|\, \gamma_i, \lambda) \;=\; \begin{cases} \mathrm{Lap}(\lambda), & \gamma_i = 1 \\ \delta_0, & \gamma_i = 0 \end{cases} \tag{8}$$

where δ_0 is a point-mass at zero. In this expression, the scale parameter λ is common to each class $\mathcal{M}_\gamma$ and is taken from the population model given in (7). Now, consider the observed, noisy wavelet coefficients $y_1, \ldots, y_n$ generated by (1) given $\beta_1, \ldots, \beta_n$. Under $\mathcal{M}_\gamma$, each y_i is distributed according to

$$f(y_i \,|\, \gamma_i, \lambda) \;=\; \begin{cases} m_\lambda(y_i), & \gamma_i = 1 \\ \phi(y_i), & \gamma_i = 0 \end{cases} \tag{9}$$

where the marginal density m_λ is the convolution of $\mathrm{Lap}(\lambda)$ and ϕ, the standard normal density, and is computed in Proposition 1.

We now introduce an MDL criterion for evaluating the competing model classes $\mathcal{M}_\gamma$. For simplicity, we collect the noisy wavelet coefficients into a single vector $y = (y_1, \ldots, y_n)$. (Note that in the previous section we used y to denote a single coefficient, while in this section it will refer to the entire n-vector.) To derive an MDL criterion, we need to construct a code or description based on $\mathcal{M}_\gamma$ that can be used to (at least in principle) transmit the data vector y. Each model class is then compared via the code length function $L(y|\mathcal{M}_\gamma) = L(y|\gamma)$. We propose a multi-stage description of the observed subband data y under $\mathcal{M}_\gamma$ consisting of the following components:

1. First, we encode the index vector γ at a cost of $L(\gamma)$ bits.

2. Then, given γ we encode estimates of any hyperparameters $\hat{\theta}$ involved in defining $\mathcal{M}_\gamma$, and denote the cost by $L(\hat{\theta}|\gamma)$. Here, the vector θ includes the Laplacian scale parameter λ; if we had chosen the generalized Gaussian family, we would also have a shape parameter.

3. Finally, having encoded γ and the hyperparameter estimates $\hat{\theta}$, we describe the full data set y using a *mixture code* based on the marginal distribution defined in (9) corresponding to model class $\mathcal{M}_\gamma$, requiring $L(y|\gamma, \hat{\theta})$ bits.

The total description length of the data y based on $\mathcal{M}_\gamma$ is then simply the sum

$$L(y|\gamma) = L(\gamma) + L(\hat{\theta}|\gamma) + L(y|\gamma, \hat{\theta})\,. \tag{10}$$

We now consider constructing codes or descriptions for the information transmitted in each of the steps presented above.

We begin with Step 2. The Laplacian-based (mixture) model has only one hyperparameter to be estimated, so that $\theta = \lambda$. Chang et al. [3] derive a moment estimator for λ that seems to work well in this context. Maximum likelihood estimators (based on the marginal distribution of the data y) for similar hyperparameters have been applied in various Bayesian wavelet schemes [8, 1] yielding

$$\hat{\lambda} = \mathrm{argmax}_{\lambda>0} \prod_{i=1}^{n} m_\lambda(y_i)\,, \tag{11}$$

which we will apply in our application. Note that in both of these approaches we estimate λ under the *true model* $\gamma = (1, 1, \ldots, 1)$. Recall that in the approximating model classes, we take the Laplacian scale parameter to be the same as that for the truth (7). Because $\hat{\lambda}$ is the same for each $\mathcal{M}_\gamma$, the cost associated with its transmission $L(\hat{\lambda}|\gamma) = L(\hat{\lambda})$ is also the same for each class. Therefore, we can safely ignore this term when comparing $\mathcal{M}_\gamma$ for different γ. As for the description length of y from Step 3, we can now use the mixture distribution (9) so that each coefficient is associated with a cost

$$L(y_i|\gamma_i, \hat{\lambda}) = -\log f(y_i|\gamma_i, \hat{\lambda})\,, \quad i = 1, \ldots, n\,,$$

and $L(y|\gamma, \hat{\lambda}) = \sum_i L(y_i|\gamma_i, \hat{\lambda})$.

Finally, we take up the transmission of the model index γ implicit in Step 1. To build a code for γ, we assume that the individual γ_i are iid Bernoulli random variables with the probability of success (inclusion in the model) p. In early Bayesian wavelet applications, p was held fixed to express *prior beliefs* about the *sparsity* of significant coefficients in a wavelet transform [7]. More recent treatments have taken p to be another hyperparameter, similar in spirit to λ. Both p and λ affect the overall size of the model favored by our MDL criterion (10). Large λ produces small models, as do small values of p. In our setting, λ is common to all the approximating classes $\mathcal{M}_\gamma$ and is assumed to be the same as the true or population-level quantity. Similarly, we determine a data-dependent value for p that is fixed and is used to encode each γ. Before we go into the details of this method, we note that in general, the elimination of hyperparameters is still active research area in the MDL literature, and what we provide here is a method which yielded sensible results in this particular natural image application.

Under the iid Bernoulli assumption and if the γ_i were available, a good estimator of p would be the sample mean

$$\bar{p} = \frac{1}{n} \sum_i \gamma_i \,, \tag{12}$$

which would yield the usual two-stage coding scheme for γ. Since γ_i is not observable, we choose to seek a 0-1 valued estimator based on y_i, and such an estimator is equivalent to a hypothesis test. This leads to the search for an optimal test of the hypotheses

$$H_0 : \ \gamma_i = 0 \ \text{ vs. } \ H_1 : \ \gamma_i = 1$$

or equivalently

$$H_0 : \ \beta_i = 0 \ \text{ vs. } \ H_1 : \ \beta_i \neq 0.$$

Assuming (1) and (7), the distributions of y_i are $\phi(\cdot)$ and $m_\lambda(\cdot)$ respectively under the null and alternative hypotheses.

With a uniform prior on the two hypotheses, the optimal Bayesian procedure rejects the null model, or sets $\hat{\gamma}_i = 1$, if and only if

$$\frac{P(H_0|y_i)}{P(H_1|y_i)} = \frac{P(y_i|H_0)P(H_0)}{P(y_i|H_1)P(H_1)} = \frac{\phi(y_i)}{m_\lambda(y_i)} < 1.$$

This implies the following estimator for γ_i:

$$\hat{\gamma}_i = I_{\{m_{\hat{\lambda}}(y_i) > \phi(y_i)\}}.$$

Plugging $\hat{\gamma}_i$ in (12) and using the estimator $\hat{\lambda}$ in the place of λ, we obtain a data-driven estimator of the hyperparameter p used in coding γ:

$$\hat{p} = \frac{1}{n} \sum_i I_{\{m_{\hat{\lambda}}(y_i) > \phi(y_i)\}}. \tag{13}$$

This expression can be seen as a first-pass at optimizing the overall code length (10) componentwise: If $m_{\hat{\lambda}}(y_i)$ is larger than $\phi(y_i)$, then the MAP estimate for γ_i is 1. Once we have an estimate of p, we use a simple two-stage coding scheme for each value of γ so that

$$L(\gamma) \;=\; L(\hat{p}) + L(\gamma|\hat{p})\,.$$

Finally, collecting the various estimates and code lengths, we arrive at the MDL criterion

$$lMDL(y|\gamma) \;=\; L(y|\gamma) \;=\; L(\hat{p}) + L(\hat{\lambda}) + L(\gamma|\hat{p}) + L(y|\gamma, \hat{\lambda})\,, \tag{14}$$

where the "l" in $lMDL$ represents the use of the Laplacian distribution in (8). Again, observe that for the purpose of model comparison, the costs $L(\hat{p})$ and $L(\hat{\lambda})$ are fixed for each class $\mathcal{M}_\gamma$ and hence can be ignored. Minimizing the total description length $lMDL(y|\gamma)$ over γ results in an adaptive thresholding rule, as shown by the theorem below.

Theorem 1: Assume the set-up of Proposition 1. If we encode the estimated parameters $\hat{\lambda}, \hat{p}$ with a fixed precision (say, $1/\sqrt{n}$) and $L(\gamma|\hat{p})$ is taken as the Bernoulli coder with probability $\hat{p}$, then for any γ with k_γ 1's,

$$lMDL(y|\gamma) = L(\hat{\lambda}) + L(\hat{p}) - k_\gamma \log(\hat{p}) - (n - k_\gamma)\log(1-\hat{p}) - \sum_{i\in\gamma} \log m_{\hat{\lambda}}(y_i) - \sum_{i\notin\gamma} \log \phi(y_i).$$

Minimizing $lMDL$ is equivalent to a thresholding rule which sets to zero all β_i for which $|y_i|$ is less than

$$T_{lMDL} = h_{\hat{\lambda}}^{-1}\left(\frac{1-\hat{p}}{\hat{p}}\right), \tag{15}$$

where $h_{\hat{\lambda}}(y) = m_{\hat{\lambda}}(y)/\phi(y) = \frac{\lambda}{2}\sqrt{2\pi}r_+(y;\lambda)$.

Proof: See Appendix C.

As MDL is really concerned about model classes, we have considerable freedom to entertain various estimates of the coefficients β_i for which $\gamma_i = 1$. We have chosen to use the conditional (posterior) mean given y_i shown in Proposition 1, calculated under the model (6)–(9). With this choice of estimator, minimizing the $lMDL$ selection criterion (14) is equivalent to the following procedure:

$$\hat{\beta}_i = \begin{cases} y_i - \hat{\lambda} r_-(y_i; \hat{\lambda})/r_+(y_i; \hat{\lambda}) & \text{if } |y_i| \geq T_{lMDL} \\ 0 & \text{if } |y_i| < T_{lMDL}; \end{cases}, \tag{16}$$

where $\hat{\lambda}$ and $\hat{p}$ are set as in (11) and (13), respectively. As with the determination of how to handle hyperparameters, estimation after model selection within the MDL framework is also an area of active research.

Several comments are in order before we present experimental results. First, throughout our discussion, we have taken the noise variance σ^2 equal to one. In practice this is obviously another parameter that must be estimated before we can properly code the data y. As mentioned in part A of Section II, estimates of σ^2 can be obtained from coefficients in the most detailed subband HH(1) (cf. [13, 14, 6, 36]). Using this estimate, we can standardize the data from each subband and apply (16). Next, we have chosen to estimate λ and p sequentially. An alternative is to simultaneously optimize the code length function (10) over these parameters. This is equivalent to the marginal maximum likelihood procedures of Clyde and George [8]. Unfortunately, this approach tends to keep far too many coefficients, producing an inefficient analytical compression.

C *Experimental Results*

To evaluate the performance of the various thresholding schemes, we conducted an experiment involving seven standard images from the USC image library. The MSE results and the proportion of significant (non-zero) coefficients are given in Table 1. For each image, we examined three noise levels ($\sigma = 10, 20$ and 40). Our MSE results in the table are scaled by the standard deviation of the noise so as to be comparable between runs on the same image. Shrinkage was applied separately to the HL, LH and HH subbands for levels 1, 2 and 3. The smooth subband LL(3) was not subject to any shrinkage or estimation. The percent of significant coefficients kept by each method is presented in the row below the MSE figures. The coefficient count includes the 4096 coefficients from LL(3). To make things comparable, we restrict ourselves to the thresholding rules *BayesShrink* of Chang et al. [3], *MapShrink* of Nikolova [26] (cf. also [24]), and the MSE-optimal threshold derived in Corollary 1; although other less restrictive (and more expensive in terms of coefficient count) denoising techniques might give better MSE performance (cf. [9]).

The MSE's in this table are quite comparable for all four methods. The threshold T_{bayes} and T_{opt} perform the best in terms of MSE by at most 10% over T_{map} and $lMDL$, while $lMDL$ requires

Image		T_{map}	T_{bayes}	T_{opt}	$lMDL$
Baboon	*snr=4.2*	0.740 (0.637)	0.729 (0.734)	0.727 (0.728)	0.844 (0.359)
	snr=2.1	0.509 (0.312)	0.483 (0.431)	0.483 (0.450)	0.553 (0.169)
	snr=1.1	0.276 (0.093)	0.258 (0.159)	0.257 (0.180)	0.285 (0.056)
Barbara	*snr=5.5*	0.50 (0.405)	0.505 (0.523)	0.502 (0.532)	0.545 (0.175)
	snr=2.7	0.322 (0.196)	0.317 (0.279)	0.316 (0.294)	0.346 (0.088)
	snr=1.4	0.182 (0.055)	0.166 (0.104)	0.165 (0.121)	0.183 (0.035)
Boat	*snr=5.2*	0.380 (0.256)	0.376 (0.345)	0.374 (0.358)	0.410 (0.118)
	snr=2.6	0.221 (0.112)	0.214 (0.164)	0.213 (0.178)	0.234 (0.057)
	snr=1.3	0.108 (0.046)	0.105 (0.068)	0.104 (0.074)	0.112 (0.029)
Couple	*snr=4.4*	0.448 (0.320)	0.443 (0.425)	0.443 (0.436)	0.498 (0.151)
	snr=2.2	0.270 (0.125)	0.256 (0.191)	0.254 (0.213)	0.282 (0.068)
	snr=1.1	0.127 (0.050)	0.122 (0.074)	0.122 (0.080)	0.132 (0.031)
Goldhill	*snr=4.9*	0.433 (0.269)	0.421 (0.378)	0.422 (0.394)	0.476 (0.136)
	snr=2.5	0.236 (0.095)	0.223 (0.150)	0.222 (0.171)	0.246 (0.055)
	snr=1.2	0.106 (0.035)	0.101 (0.052)	0.10 (0.058)	0.108 (0.025)
Lena	*snr=4.2*	0.267 (0.170)	0.270 (0.246)	0.270 (0.222)	0.285 (0.073)
	snr=2.1	0.155 (0.070)	0.148 (0.109)	0.148 (0.122)	0.160 (0.038)
	snr=1.0	0.075 (0.031)	0.072 (0.044)	0.072 (0.049)	0.076 (0.022)
Tank	*snr=2.4*	0.536 (0.277)	0.501 (0.403)	0.500 (0.427)	0.569 (0.172)
	snr=1.2	0.254 (0.084)	0.242 (0.131)	0.241 (0.151)	0.262 (0.057)
	snr=0.6	0.098 (0.029)	0.095 (0.042)	0.094 (0.048)	0.10 (0.023)

Table 1: Comparing several thresholding schemes in terms of MSE and coefficient count on seven images from the literature. The numbers in the parentheses are the proportions of significant coefficients. The criterion $lMDL$ is at most 10% worse in terms of MSE, but can require as little as half of the coefficients.

as little as half the number of coefficients as T_{map}. This is why we say that $lMDL$ achieves a good trade-off between denoising and compression. In the next section, we assess this trade-off formally. In terms of visual quality, the reconstructed images produced by each of the competing schemes are also similar.

The proportion of coefficients kept by each method also gives us a sense of the adaptivity at work here. For example, the image *Baboon* consists mainly of separate regions of texture and exhibits very little in the way of smoothness. As a result, at the high SNR level (4.2) the proportion of significant coefficients varies from 36% for $lMDL$ to 73% for *BayesShrink*. By comparison, consider the *Lena* image which clearly exhibits large, smooth regions. At the high SNR level (4.2) the proportion of significant coefficients ranges from only 7% to 25%.

To provide one further benchmark for our MDL results, we also implemented the simple two-stage MDL thresholds suggested by Moulin [22, 23] and Saito [32] and tested them on the *Lena* and *Baboon* images. Saito's procedure produced spectacular decreases in the numbers of kept coefficients (sometimes throwing away up to 80% *more* than $lMDL$), with a considerable rise in MSE and a corresponding degradation in the visual quality of the reconstruction. The threshold of Moulin performs better, but the results depend strongly on the signal-to-noise ratio (the best results coming from high noise settings). At their most successful, these schemes pay 20% more in terms of MSE than $lMDL$, while 50% is typical.

IV Assessing Simultaneous Denoising and Compression

When both denoising and compression matter, the experimental results from the previous section indicate the superiority of $lMDL$ over its competitors such as *BayesShrink* (or *Sure*) and *MapShrink*. In this section, we formally address the problem of optimal simultaneous denoising and compression, and we distinguish between analytical and actual compression. For a selected subband of the Lena image, the performance of the different thresholding schemes is evaluated within each framework and compared to the optimal. For comparable results, we restrict ourselves to methods based on iid methods and scalar quantization based on iid observations since scalar quantization has been the most common in wavelet image compression.

Clearly, denoising and compression are related. In the work of Donoho and Johnstone, the minimax-optimal thresholding rules are designed to denoise, but simultaneously perform an analytical compression of the noisy signal by setting some fraction of the small coefficients to zero. To describe the effectiveness of such a scheme, we refer to the percentage of non-zero coefficients as the *analytical compression rate.* This quantity is proportional to the actual coding rate if a fixed quantizer is applied to the surviving coefficients. Natarajan [25] argues that a good compression algorithm should act as a denoiser, providing the distortion level matches the standard deviation of the noise. We follow Chang et al. [3] and consider quantizers having a *zero-zone* corresponding to the threshold level of the denoiser. A more sophisticated use of MDL as a complexity-regularized quantizer can be found in Liu and Moulin [24].

A *Optimal Simultaneous Denoising and Analytical Compression*

Suppose we are told that from a noisy wavelet decomposition we can only keep k coefficients in a given subband. To minimize MSE, we should choose to eliminate the smallest $n-k$. If in addition we know the empirical distribution of the true noiseless coefficients, then the best estimator for the k remaining coefficients is a Bayes posterior mean where we take the empirical truth as our prior. In this case, the Bayes risk is the empirical oracle Bayes risk since it is calculated using the actual noiseless wavelet coefficients. We denote this quantity by $E(p)$ (E for error), where $p = k/n$. A plot of p against $E(p)$ is termed the *empirical oracle R-E curve*, reminiscent of the rate-distortion (R-D) curve for actual compression. The R-E curve is the gold standard against which we should compare thresholding rules. We note that unlike traditional R-D relationships, the oracle performance for the R-E curve may not be achievable.

In practice, we must be content with model-based priors as in (10) which might require estimating one or more hyperparameters from the noisy data. For example, we estimate λ in our Laplacian (mixture) model by maximizing the (marginal) likelihood of the data, and apply an empirical Bayes estimator to the coefficients we decide to retain. In such cases, we can construct a curve that complements the R-E analysis above by simply replacing the empirical oracle Bayes risk with MSE = MSE(p). Clearly, we can draw such a curve for any thresholding scheme.

For the HL(1) subband of Lena, the lefthand panel in Fig. 2 gives the empirical oracle R-E curve together with the MSE curves for a Laplacian model-based scheme, and the soft-thresholding methods. The coordinates of the horizontal axis range from 0 to 0.1 since (as seen in Table 1), this subband seems to have relatively few significant coefficients. The points highlighted on these curves mark the performance of the different thresholding schemes under study. *MapShrink* of Nikolova[26] and Moulin and Liu [24] stops too short on the soft-threshold curve while Chang et al.'s *BayesShrink* [3] threshold comes close to achieving the minimum MSE performance among the class of soft-thresholding estimates for this subband. Not shown here is the fact that beyond the 0.1 level, the soft-thresholding curve continues to rise rather quickly. This behavior is in contrast with the Laplacian curve which flattens out (too early), but remains relatively constant for a large range of proportions. The same flattening out is observed with the empirical oracle curve (but at a lower value). Ideally, one would like to find the trade-off point from the empirical oracle curve to achieve a good MSE with a low rate. A point where the curve turns flat (or the curve's elbow) is desirable. If we map the proportions chosen by $lMDL$ and $BayesShrink$ to this oracle curve, we

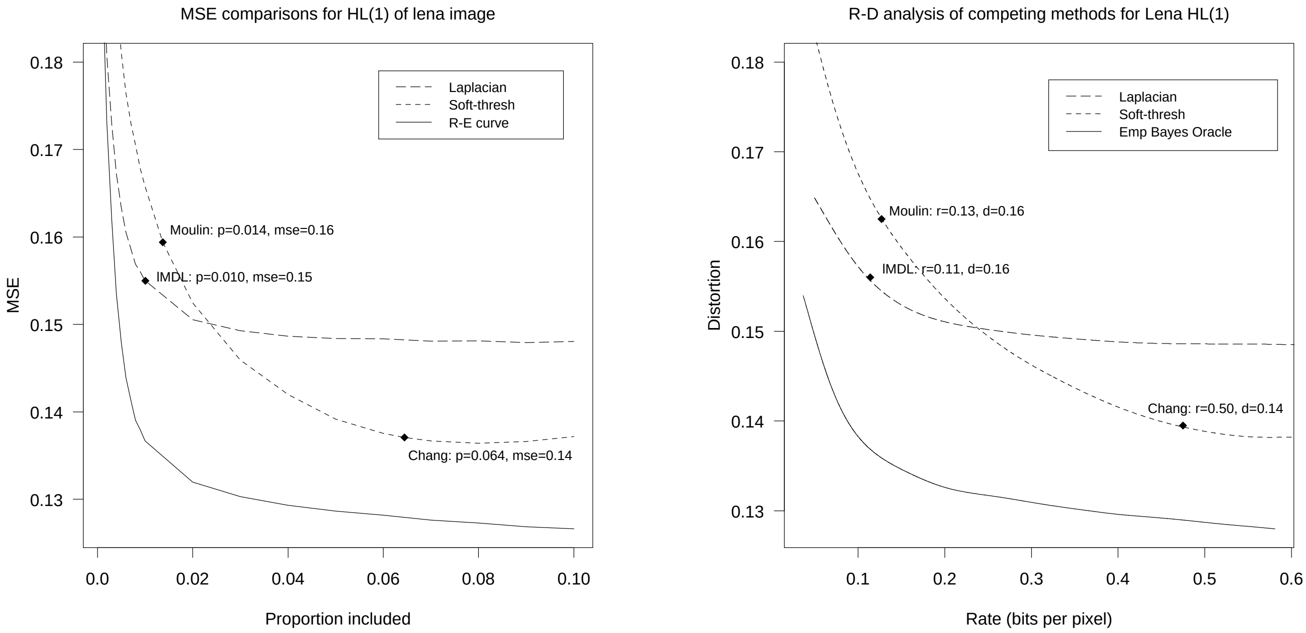

Figure 2: Left: MSE comparison of the soft-threshold and Lap-model methods against the the oracle R-E curve. The subband is HL(1) of Lena. Right: Rate-distortion comparison of the soft-threshold and Lap-model models against the oracle R-D curve. The subband is HL(1) of Lena.

see that the former is a bit short of the turning point or elbow, but the latter goes too far into the flat region. Relatively speaking, the $lMDL$ proportion is more sensible. However, the MSE of the $lMDL$ rule is noticeably above that from the oracle curve with the same proportion because the Laplacian R-E curve is almost a constant shift upwards from the orcale curve.

B *Optimal Simultaneous Denoising and (Actual) Compression*

After thresholding, we construct a (denoised) estimate of the signal using only the "large" coefficients. As mentioned above, this reduction of data represents an analytical compression of the original, noisy signal, in the sense that we require fewer parameters to form the reconstruction. For transmission or storage purposes, our interest is in actual compression; that is, reducing the number of *bits* required for a reconstruction. This involves a final quantization step, yielding the classical tradeoff between rate and distortion. Chang et al. [3] connect denoising via thresholding and compression by taking the threshold as the zero-zone in quantization and using MDL to select the quantization step size for significant coefficients. We take a similar approach here to carry out compression after thresholding and therefore compare the actual compression effects of different thresholding procedures. However, for a given threshold or zero-zone, roughly speaking, our compression rate is the entropy of the quantized y that gives the best rate and distortion trade-off among all possible quantizations that incorporate a zero-zone.

Again, we use an oracle simultaneous denoiser and compressor. Because of the quantization, an empirical oracle rate-distortion (R-D) curve is obtained based on scalar quantization of an iid model as the benchmark against which we compare the procedures. We now describe the steps to get points on this oracle R-D curve. The problem is that of optimal quantization based on noisy observations. That is, we need to find the optimal quantizer $Q^*(y)$ based on y to minimize the distortion with respect to β:

$$Q^* = \text{argmin}_Q E\, d\,(\beta, Q(y)) \tag{17}$$

for some distortion measure $d(\cdot, \cdot)$. The expectation in the expression above is with respect to the joint distribution of β and y. Ephraim and Gray [15] derive a simple solution to this problem by rewriting the expectation in (17) as

$$E\, d\,(\beta, Q(y)) = E\, d'\,(y, Q(y))\ , \quad \text{where} \quad d'\,(y, Q(y)) = E\,\{d\,(\beta, Q(y))\,|y\}$$

is a new distortion measure comparing y and $Q(y)$. Therefore, a solution Q^* of (17) can be obtained by directly quantizing the noisy data y using the new distortion measure $d'(\cdot, \cdot)$. Ephraim and Gray [15] provide conditions on the original function d under which the commonly applied Lloyd-Max algorithm is guaranteed to converge for the new problem based on d'.

Let $f(\beta, y) = f(\beta)f(y|\beta)$ denote the joint distribution of β and y. We can interpret $f(\beta)$ as either a prior for β in the usual Bayesian paradigm; or as a tool for constructing a mixture code for β as we have done in Section III. Adopting Bayesian terminology for the moment, we recognize d' as the expected distortion calculated under the posterior distribution of β, $f(\beta|y)$. In our analysis, we focus on squared-error as a distortion measure, so that

$$d(\beta, Q(y)) = (\beta - Q(y))^2,$$

whose expectation over the posterior distribution is minimized by the Bayes estimator or the posterior mean $\hat{\beta}_b$. Hence

$$d'\,(y, Q(y)) = (\beta - \hat{\beta}_b)^2 + E\{(\hat{\beta}_b - Q(y))^2|y\}.$$

Since the mapping of y to $\hat{\beta}_b$ is continuous and 1-1, all quantizers Q have their counterparts in the $\hat{\beta}_b$ domain and thus the problem of minimizing Q^* becomes that of finding the Lloyd-Max quantizer of the Bayes estimator of β.

We now consider designing a quantizer for y based on the modified distortion measure d'. Because we are studying threshold-estimates, our quantizer $Q(y)$ should explicitly incorporate a zero-zone $[-T, T]$ corresponding to a selected threshold value T. The interval $(-\infty, -T]$ is then divided into K (possibly unequally sized) bins, as is $[T, \infty)$. Let $\Delta = \{\Delta_k\}$ denote the resulting partition of the real line, $\cup_k \Delta_k = \mathbb{R}$. We do not insist that the two sets of bins on either side of the zero-zone be symmetric, although our results do not change dramatically if we impose this restriction. For a given partition Δ, the quantized variable $Q(y)$ has entropy

$$H^y(\Delta) = -\sum P_k \log_2 P_k\,, \qquad \text{where} \qquad P_k = \int_{\Delta_k} f(y)dy\,.$$

Then the bit rate per pixel (bpp)

$$R^y(\Delta) \approx H^y(\Delta) + \text{TotalOverhead}/n\,,$$

where n is the size of the subband and TotalOverhead accounts for the transmission cost for the quantizer.

For K bins on either side of the zero-zone, we can transmit the breakpoints of the partition Δ using $2K-1$ parameters. The cost associated with the reproduction values for the quantizer will depend on whether or not we are working with parametric models for β and y. If not, we require $2K$ values (one for each bin). By contrast, the Laplacian-Gaussian model described in Section III requires only 2 parameters, λ and σ. In either case, let L denote the number of continuous parameters needed to be encoded. Then,

$$\text{TotalOverhead} \approx L \log_2 10^j + \log_2 K\,,$$

where j controls the precision with which we transmit the L parameters, and K is the number of bins used on either side of the zero-zone. Denoting Q_j the discretized reproduction level at precision j, the MSE is

$$D(\Delta) = E\, d(\beta, Q_j(y)) = E(\beta - Q_j(y))^2\,.$$

Given a fixed model (either parametric or non-parametric) for the joint distribution of (β, y), and particular choices of threshold T and precision j, we vary k to obtain a curve denoted by R-D(T, j).

By changing T and j we obtain a collection of curves. We then take the (numerical) lower-convex hull of this collection as our R-D curve. We interpret the result as the best possible R-D trade-off given our (mild) restrictions on the type of quantizer. To a certain extent, this construction eliminates the possible suboptimality of any fixed quantizer, making possible the comparison of the different models or estimation schemes in Section III.

Because the soft-thresholding rules do not have a prior corresponding with it, in calculating the curves to come we used the L-M optimal quantizers as discussed above to the same single realization of the noise sequence. Under the white noise model, if we take $f(\beta)$ to be the empirical distribution of the true coefficients in a subband, the resulting joint distribution for (β, y) is non-parametric. With this choice, we obtain the so-called oracle R-D curve for the subband. Under a Laplacian specification for $f(\beta)$, our model is fully parametric requiring only estimates of λ and σ^2. Our particular threshold T_{lMDL} corresponds to a point on the resulting R-D curve. For the soft-thresholding R-D curve, including *BayesShrink* and *MapShrink*, we do not have a distribution $f(\beta)$ to work with, but we do have estimates $\hat{\beta}_i$ of β_i based on y_i. This R-D curve requires a non-parametric overhead calculation. As noted above, our distortion calculations involve only the posterior means, themselves estimates of β_i. Therefore, it is fair to compare soft-thresholding procedures with Bayesian estimators in this fashion. As with $lMDL$, *BayesShrink* and *MapShrink* correspond to two points on this soft-thresholding R-D curve.

The three R-D curves and related points of $lMDL$, *BayesShrink* and *MapShrink* are calculated for Lena HL(1) and shown in Fig. 2. They are very similar to the R-E curves we saw in the previous section. Hence we made the case that analytical compression relates to actual compression in an intimate way.

V Conclusions

Wavelet thresholding is a simple and effective pixel-based tool for image denoising and analytical compression. Application of this method depends only on determining the value of a single parameter, the threshold. Our approach based on statistical model selection, however, separates the "kill" and "keep" actions, resulting in a more flexible procedure that allows for improved estimation of the coefficients that are not set to zero. MDL as a general model selection principle works best when the underlying assumptions agree with observations about the data. Within the class of natural images, this is certainly the case. For each subband, the marginal distribution of wavelet coefficients is known to be approximately Laplacian or double exponential. Starting from this distributional characterization, we devise a coding scheme and carry out model selection using a multi-stage MDL criterion. The resultant procedure, $lMDL$, is an adaptive thresholding rule, and we estimate the significant coefficients via a conditional (posterior) mean. The $lMDL$ procedure is compared with existing successful thresholding rules for images, *BayesShrink* and *MapShrink*, and it is shown to be an excellent simultaneous denoiser and compressor. This property is then formally assessed in two optimality frameworks in Section IV.

These optimality analyses show that the thresholding rules based on the Laplacian model fall shy of attaining optimal performance by almost a constant amount. One possible explanation is that the distributional characterization motivating our approach is an imperfect approximation for some subbands. This can been seen via the quantile-quantile (QQ) plots of noiseless wavelet coefficients in Fig. 3 for HL(1) of the *Lena* image. Clearly, with one global parameter to control its shape, the Laplacian distribution is forced to strike a compromise between peak and tail. It appears from Fig. 3, however, that the larger family of generalized Gaussian distributions can also suffer from the same "constraint of form." Furthermore, the mathematical form of the generalized Gaussian density makes it difficult to work with in practice.

The simplicity of the Laplacian, however, has led to closed-form expressions for almost every quantity required in the denoising process. This suggests a two-piece Laplacian, also known in the statistics literature as a (linear) logspline distribution (see [34] for an overview and applications of this family). To generate Fig. 3, we fit a linear logspline model with two symmetrically placed knots or breakpoints:

$$\text{Logspline}: \ \frac{e^{-(a-b)|x|-b(|x|-\kappa)_+}}{2\left[e^{(b-a)\kappa}\left(\frac{1}{a}+\frac{1}{b-a}\right)+\frac{1}{a-b}\right]} \tag{18}$$

where $(\cdot)_+ = \max(0,\cdot)$, and $\kappa > 0$ denotes a knot contained within the range of the data. (This seemingly awkward parametrization is actually advantageous for estimation.) The logspline fits in Fig. 3 are produced based on estimated parameters a, b and κ by applying maximum likelihood to the noiseless data β_i. It is clear that this logspline model is able to better separate the peak and tail behavior for the coefficients in this subband, with one more parameter than the generalized Gaussian. Replacing the Laplacian distribution with a logspline model, we can derive results similar to Proposition 1 and Theorem 1, exhibiting closed-form expressions for the threshold and shrinkage. We are exploring fitting logspline models based on noisy data y_i, and will carry it out on test images. We believe by incorporating this two-piece Laplacian model in our approach we can close the gap of performance as seen in the R-E and R-D analyses. In contrast, when these models have been applied locally, to denoise a sample of pixels in a small neighborhood, there has been no observed improvement in performance [21]. Given the small sample sizes used in the local model of [21], it is difficult to discriminate between a logspline and Laplacian model.

Finally, the rules proposed in this paper are based on an assumption that coefficients are iid from some population. Improvements are expected if one takes into account the spatial dependence (cf. [31, 18, 4, 9, 37, 21]). This can be done in our MDL framework by coding the model index γ using a Markov Random Field (MRF) instead of a Bernoulli coder. Especially appealing are the MRFs such as Chien's model which give rise to edges and lines. These are the topics of current and future research.

Appendix A

Proof of Proposition 1

On a non-negative interval $[a, b]$, it is straightforward to verify the following two identities.

$$\begin{aligned} I(y; a, b, \lambda) &:= \int_a^b \frac{\lambda}{2} e^{-\lambda\beta} \phi(y - \beta) d\beta \\ &= \frac{\lambda}{2} \sqrt{2\pi} \phi(y) e^{(y-\lambda)^2/2} [\Phi(b - y + \lambda) - \Phi(a - y + \lambda)]; \end{aligned}$$

$$\begin{aligned} II(y; a, b, \lambda) &:= \int_a^b \frac{\lambda}{2} e^{-\lambda\beta} \beta \phi(y - \beta) d\beta \\ &= \frac{\lambda}{2} \sqrt{2\pi} \phi(y) e^{(y-\lambda)^2/2} [\frac{1}{\sqrt{2\pi}} e^{-(a-y+\lambda)^2/2} \\ &\quad - \frac{1}{\sqrt{2\pi}} e^{-(b-y+\lambda)^2/2} + (y - \lambda)(\Phi(b - y + \lambda) - \Phi(a - y + \lambda))]. \end{aligned}$$

Using this notation, we find

$$\begin{aligned} m_\lambda(y) &= \int_{-\infty}^{\infty} \frac{\lambda}{2} e^{-\lambda\beta} \phi(y - \beta) d\beta \\ &= I(y; 0, \infty, \lambda) + I(-y; 0, \infty, \lambda) \\ &= \frac{\lambda}{2} \sqrt{2\pi} \phi(y) [e^{(y-\lambda)^2/2} [1 - \Phi(-y + \lambda) + e^{(y+\lambda)^2/2} [1 - \Phi(y + \lambda)] \\ &= \frac{\lambda}{2} \sqrt{2\pi} \phi(y) [e^{(y-\lambda)^2/2} [\Phi(y - \lambda) + e^{(y+\lambda)^2/2} \Phi(-y - \lambda)] \\ &=: \frac{\lambda}{2} \sqrt{2\pi} \phi(y) r_+(y; \lambda), \end{aligned}$$

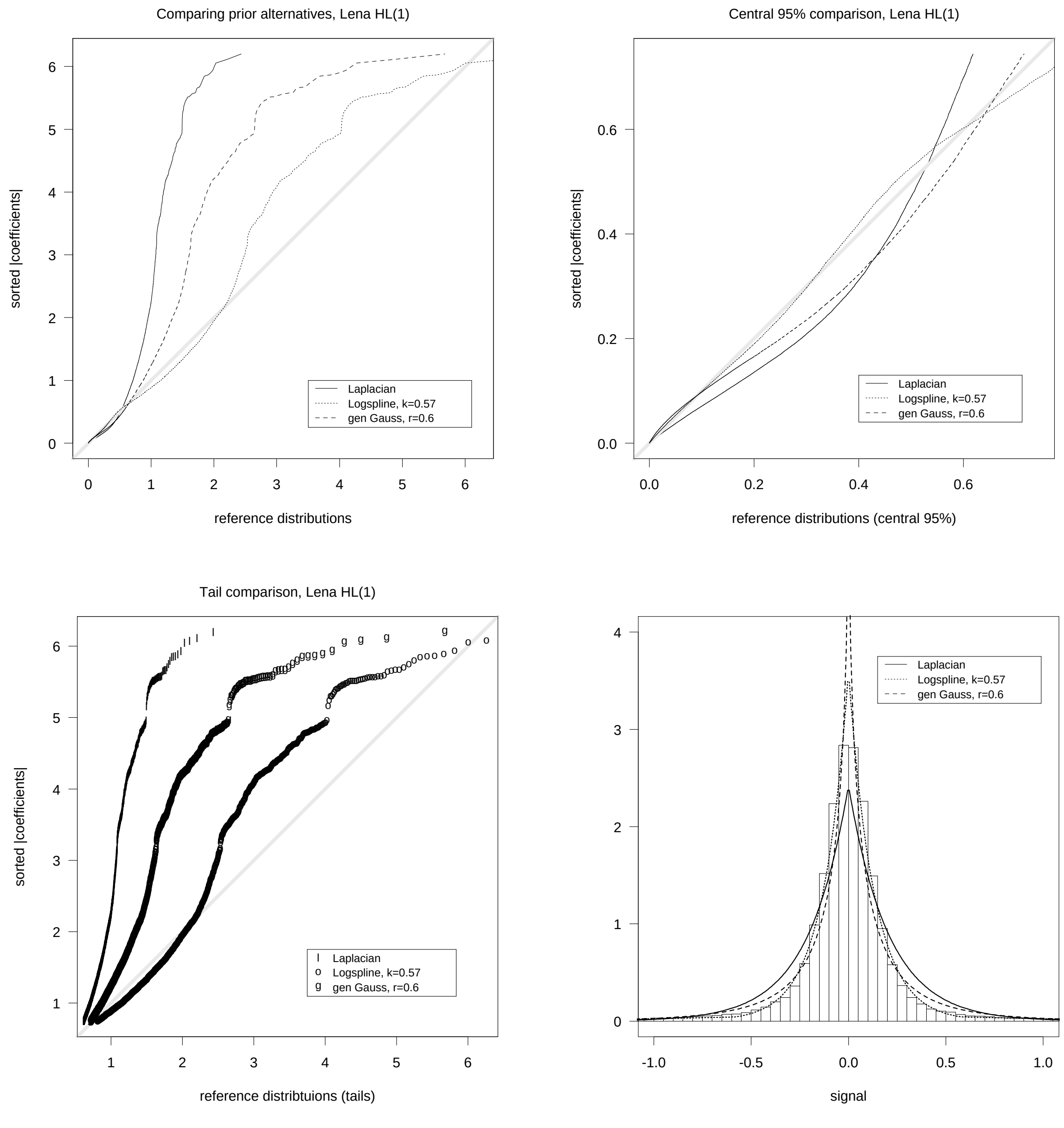

Figure 3: Comparing prior models for the (absolute value of the) noiseless coefficients of Lena HL(1). The Laplacian distribution makes a compromise between the tails and the peak at zero, while the generalized Gaussian and logspline are closer to the 1-1 line.

where

$$r_+(y;\lambda) = e^{(y-\lambda)^2/2}\Phi(y-\lambda) + e^{(y+\lambda)^2/2}\Phi(-y-\lambda).$$

For the posterior mean, let us first calculate

$$\begin{aligned}
& \int_{-\infty}^{\infty} \beta\frac{\lambda}{2}e^{-\lambda\beta}\phi(y-\beta)d\beta \\
= \ & II(y;0,\infty,\lambda) - II(-y;0,\infty,\lambda) \\
= \ & \frac{\lambda}{2}\sqrt{2\pi}\phi(y)[\frac{1}{\sqrt{2\pi}} + (y-\lambda)e^{(y-\lambda)^2/2}\Phi(y-\lambda) - \frac{1}{\sqrt{2\pi}} + (y+\lambda)e^{(y+\lambda)^2/2}\Phi(-y-\lambda)] \\
= \ & \frac{\lambda}{2}\sqrt{2\pi}\phi(y)[(y-\lambda)e^{(y-\lambda)^2/2}\Phi(y-\lambda) + (y+\lambda)e^{(y+\lambda)^2/2}\Phi(-y-\lambda)] \\
= \ & \frac{\lambda}{2}\sqrt{2\pi}\phi(y)[yr_+(y;\lambda) - \lambda r_-(y;\lambda)]
\end{aligned}$$

where

$$r_-(y;\lambda) = e^{(y-\lambda)^2/2}\Phi(y-\lambda) - e^{(y+\lambda)^2/2}\Phi(-y-\lambda) \ \ .$$

Thus the posterior mean

$$\begin{aligned}
& \int_{-\infty}^{\infty} \beta\frac{\lambda}{2}e^{-\lambda\beta}\phi(y-\beta)d\beta/m_\lambda(y) \\
= \ & \frac{\lambda}{2}\sqrt{2\pi}\phi(y)[yr_+(y;\lambda) - \lambda r_-(y;\lambda)]/[\frac{\lambda}{2}\sqrt{2\pi}\phi(y)r_+(y;\lambda)] \\
= \ & y - \lambda r_-(y;\lambda)/r_+(y,\lambda).
\end{aligned}$$

Appendix B

Proof of Corollary 1

Denote (4) by

$$R_{\text{sure}}(y,t) := n - 2\sum_{i=1}^{n} I_{\{i:|y_i|\le t\}} + \sum_{i=1}^{n}(|y_i|\wedge t)^2. \tag{19}$$

Conditioning on $\{\beta_i\}$ and for $Y_i \sim N(\beta_i, 1)$, Stein's lemma [33] gives:

$$E\sum_i (Y_i - f_s(Y_i;T))^2 = ER_{\text{sure}}(Y,T).$$

When $\{\beta_i\}$ are iid with a prior distribution w, it follows that

$$\begin{aligned}
& MSE(f_s(Y,T)) \\
= \ & E(Y - f_s(Y;T))^2 \\
= \ & ER_{\text{sure}}(Y,T)/n \\
= \ & 1 - 2P(|Y| \le T) + E((|Y| \wedge T)^2),
\end{aligned}$$

where the expectation is taken over the marginal distribution of Y. This distribution is $m_\lambda(y)$ when w is Laplacian (λ). Now the LHS of the above expression can be rewritten as

$$M(T) := MSE(f_s(Y,T)) = 1 - 4\int_0^T m_\lambda(y)dy + 2\int_0^T y^2 m_\lambda(y)dy + 2T^2\int_T^\infty m_\lambda(y)dy.$$

Taking the derivative of $M(T)$ with respect to T, we obtain the equation that the optimal MSE soft-threshold t must satisfy:

$$m_\lambda(t) = t\int_t^\infty m_\lambda(y)dy, \tag{20}$$

which actually holds for any prior w if we use the corresponding m.

Rewrite

$$m_\lambda(y) = \frac{\lambda}{2}e^{\lambda^2/2}q_+(y;\lambda),$$

with

$$q_+(y;\lambda) = e^{-\lambda y}\Phi(y-\lambda) + \exp^{\lambda y}\Phi(-y-\lambda).$$

By integration by parts, it can be calculated that

$$\begin{aligned}\int_t^\infty q_+(y;\lambda)dy \quad = \quad & \frac{1}{\lambda}e^{-\lambda t}\Phi(t-\lambda) + \frac{1}{\lambda}e^{\lambda^2/2}\Phi(-t) \\ & -\frac{1}{\lambda}e^{\lambda t}\Phi(-t-\lambda) + \frac{1}{\lambda}e^{\lambda^2/2}\Phi(-t).\end{aligned}$$

Hence equation (20) becomes

$$\frac{t}{\lambda} = \frac{e^{-\lambda t}\Phi(t-\lambda) + e^{\lambda t}\Phi(-t-\lambda)}{e^{-\lambda t}\Phi(t-\lambda) - e^{\lambda t}\Phi(-t-\lambda) + 2e^{\lambda^2/2}\Phi(-t)} \quad .$$

Appendix C

Proof of Theorem 1

Because h_λ is an increasing function of $|y|$, for fixed k the collection of indices γ_k^* minimizing $lMDL$ corresponds to the k largest $|y_i|$. $lMDL$ then selects from among the γ_k^* the largest model such that $h_{\hat\lambda}(z_i) > (1-\hat p)/\hat p$ for all $i \in \gamma_k^*$.

Acknowledgement

We would like to thank two anonymous referees and the Guest Associate Editor P. Moulin, whose constructive comments led to a much-improved paper.

References

[1] F. Abramovich, T. Sapatinas, and B. Silverman, "Wavelet Thresholding via a Bayesian Approach," *J. R. Statist. Soc. B*, vol. 60(4), pp. 715-749, 1998.

[2] A. Barron, J. Rissanen, and B. Yu, "The Minimum Description Length principle in coding and modeling," Invited paper for the 50th anniversary issue of *IEEE. Trans. Information Theory*, vol. 44, pp. 2743-2760, 1998.

[3] G. Chang, B. Yu, and M. Vetterli, "Wavelet Thresholding for Image Denoising and Compression," *IEEE Trans. Image Processing*, to appear.

[4] G. Chang, B. Yu, and M. Vetterli, "Spatially adaptive wavelet thresholding with context modeling for image denoising," *IEEE Trans. Image Processing*, to appear.

[5] A. Chambolle, R. A. DeVore, N. Lee, and B. J. Lucier. "Nonlinear wavelet image processing: variational problems, compression, and noise removal through wavelet shrinkage," *IEEE Trans. Image Processing*, vol. 7, pp. 319-335, 1998.

[6] H. Chipman, E. Kolaczyk, and R. McCulloch, "Adaptive Bayesian wavelet shrinkage," *Journal of the American Statistical Association*, vol. 92(440), pp. 1413-1421, 1997.

[7] M. Clyde, G. Parmigiani, and B. Vidakovic, "Multiple shrinkage and subset selection in wavelets," *Biometrika*, vol. 85, pp. 391-402, 1998.

[8] M. Clyde and E. I. George. "Empirical Bayes Estimation in Wavelet Nonparametric Regression," *Bayesian Inference in Wavelet Based Models*, eds P. Müller and B. Vidakovic, Springer-Verlag, 1999.

[9] M. S. Crouse, R. D. Nowak, R. G. Baraniuk. "Wavelet-based statistical signal processing using hidden Markov models," *IEEE Trans. Signal Processing*, vol. 46(4), pp. 886-902, 1998.

[10] I. Daubechies, *Ten Lectures on Wavelets*, vol. 61 of *CBMS-NSF Regional Conference Series in Applied Mathematics*, SIAM, Philadelphia, 1992.

[11] R. A. DeVore and B. J. Lucier, "Fast wavelet techniques for near-optimal image processing", *IEEE Military Communications Conference Record*, San Diego, Oct. 11-14, IEEE, Piscataway, NJ, pp. 1129-1135, 1992.

[12] D.L. Donoho, "De-noising by soft-thresholding," *IEEE Trans. Information Theory*, vol. 41, no. 3, pp. 613-627, 1995.

[13] D. L. Donoho and I.M. Johnstone, "Ideal spatial adaptation via wavelet shrinkage," *Biometrika*, vol. 81, pp. 425-455, 1994.

[14] D.L. Donoho and I.M. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *Journal of the American Statistical Association*, vol. 90, no. 432, pp. 1200-1224, 1995.

[15] Y. Ephraim and R. M. Gray, "A unified approach for encoding clean and noisy sources by means of waveform and autoregressive model vector quantization," *IEEE Trans. Information Theory*, vol. 34, no. 4, pp. 826-833, 1988.

[16] U. Grenander. *Abstract inference.* Wiley. 1981.

[17] M. Hansen and B. Yu, "Model selection and Minimum Description Length principle," *Journal of the American Statistical Association*, submitted, 1998. (http://cm.bell-labs.com/who/cocteau/papers/)

[18] S. M. LoPresto, K. Ramchandran, and M. T. Orchard, "Image coding based on mixture modeling of wavelet coefficients and a fast estimation-quantization framework," *Proc. of Data Compression Conference,* (Snowbird, Utah), pp. 221-230, 1997.

[19] S. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE Trans. PAMI*, **11**, pp. 674-692, 1989.

[20] S. Mallat, *A Wavelet Tour of Singal Processing*, Academic Press, 1998.

[21] M. K. Mıhçak, I. Kozintsev, K. Ramchandran and P. Moulin, "Low complexity image denoising based on statistical modeling of wavelet coefficients," *IEEE Signal Processing Letters*, vol. 6, pp. 300-303, 1999.

[22] P. Moulin, "A wavelet regularization method for diffuse radar-target imaging and speckle-noise reduction," *J. Math. Imaging and Vision*, vol. 3, pp. 123-134, 1993.

[23] P. Moulin, "Wavelet thresholding techniques for power spectrum estimation," *IEEE Trans. Signal Processing*, vol. 42, pp. 3126-3136, 1994.

[24] P. Moulin and J. Liu, "Analysis of Multiresolution Image Denoising Schemes Using Generalized-Gaussian and Complexity Priors," IEEE Trans. Information Theory, vol. 45, pp. 909-919, 1999.

[25] B.K. Natarajan, "Filtering random noise from deterministic signals via data compression," *IEEE Trans. on Signal Processing*, vol. 43, pp. 2595-2605, 1995.

[26] M. Nikolova, "Estimées locales fortement homogènes," *Comptes Rendus Ac. Sci. Paris, Series I*, vol. 325, pp. 665-670, 1997.

[27] J. Rissanen, *Stochastic Complexity in Statistical Inquiry*, World Scientific, 1989.

[28] F. Ruggeri and B. Vidakovic, "A Bayesian decision theoretic approach to wavelet thresholding," *J. Amer. Statist. Assoc.*, vol 93(441), pp. 173-179, 1998.

[29] E. Simoncelli and E. Adelson, "Noise removal via Bayesian wavelet coring," In *Proc. IEEE Int. Conf. Image Processing,* vol. I, pp. 379-382, September 1996.

[30] E. Simoncelli, "Bayesian denoising of visual images in the wavelet domain," *Bayesian Inference in Wavelet Based Models*, eds P. Müller and B. Vidakovic, Springer-Verlag, 1999.

[31] J. Shapiro, "Embedded Image Coding Using Zerotrees of Wavelet Coefficients," *IEEE Trans. on Signal Processing,* vol. 41, no. 12, pp. 3445-3462, 1993.

[32] N. Saito, "Simultaneous noise suppression and signal compression using a library of orthonormal bases and the Minimum Description Length criterion," *Wavelets in Geophysics* (E. Foufoula-Georgiou & P. Kumar, eds.), pp. 299-324, 1994.

[33] C. Stein, "Estimation of the Mean of a Multivariate Normal Distribution," *The Annals of Statistics*, vol. 9, 1981, pp. 1135-1151.

[34] C. J. Stone and Mark Hansen and C. Kooperberg and Young K. Truong, "Polynomial splines and their tensor products in extended linear modeling (with discussion)," *The Annals of Statistics,* vol. 25, pp. 1371-1470, 1998.

[35] M. Vetterli and J. Kovačević, *Wavelets and Subband Coding*, Prentice Hall, Englewood Cliffs, NJ, 1995.

[36] B. Vidakovic, "Nonlinear wavelet shrinkage with Bayes rules and Bayes factors," *Journal of the American Statistical Association*, vol. 93, pp. 173-179, 1998.

[37] Y. Yoo, A. Ortega, and B. Yu, "Image subband coding using context based classification and adaptive quantization," *IEEE Trans. Image Processing*, vol. 8, pp. 1702-1215, 1999.