Article type: Advanced Review

Bayesian Wavelet Shrinkage

Gabriel Huerta

Department of Mathematics and Statistics, University of New Mexico, Albuquerque, NM 87131, USA

Keywords

Bayes shrinkage, Data Denoising, Discrete Wavelet Transformation, Smooth Shrinkage, Multivariate Bayes Shrinkage

Abstract

Bayesian wavelet shrinkage methods are defined through a prior distribution on the space of wavelet coefficients after a Discrete Wavelet Transformation has been applied to the data. Posterior summaries of the wavelet coefficients establish a Bayes shrinkage rule. After the Bayes shrinkage is performed, an Inverse Discrete Wavelet Transformation can be used to recover the signal that generated the observations. This article reviews some of the main approaches for Bayesian wavelet shrinkage that span both smooth and multivariate types of shrinkage.

Background

Thresholding rules became of considerable interest when De Vore and Lucier (1992) and Donoho and Kerkyacharian (1995) applied them in the wavelet shrinkage context. Wavelet shrinkage refers to a process of transforming the data or a signal with a Discrete Wavelet Transformation (DWT), implementing some type of reduction to the wavelet coefficients, and then applying the Inverse Discrete Wavelet Transformation (IDWT) to reconstruct the signal. Hard/Soft thresholding is a type of shrinkage in which those coefficients whose absolute value is smaller than a certain bound, are replaced by zero. Analytically simple, these rules are very efficient in data denoising and especially data compression problems.

However, shrinkage by thresholding poorly accounts for any prior information available about the structure of the data, the signal and the noise. From a Bayesian perspective, methods of wavelet reduction have been developed to incorporate prior knowledge on the parameters that define a model for the signal and noise. Bayesian approaches for choosing a shrinkage method have shown to be effective. In general, Bayes rules are ”shrinkers” with the desirable property that they heavily shrink small wavelet coefficients and only slightly shrink large coefficients.

Discrete Wavelet Transformation

Basics on wavelets can be found in many different texts, monographs and papers at many different levels of exposition, for example, the reader could consider Daubechies (1992), Vidakovic (1999) and Efromovich (1999) among others. A brief review of the DWT used in this article is now presented.

Let y be a data vector of dimension (size) N, where $N = 2^k$ for some positive integer k and suppose that the DWT is applied to the vector y and transformed into a vector d, i.e., $d = Wy$. This transformation of the data is linear and is represented by an orthogonal matrix W of dimensions $N \times N$. In practice, one performs the DWT without explicitly exhibiting the matrix W and by using fast filtering algorithms based on the so-called quadrature mirror filters that uniquely correspond to the wavelet of choice. More precisely, the wavelet decomposition of the vector y is a vector $d = (Gy, GHy, GH^2y, \ldots, GH^{n-1}y, H^ny)$.

The operators G and H act on sequences and are defined via high-pass and low-pass quadrature mirror filters corresponding to a particular wavelet basis. The elements of d are the wavelet coefficients where its sub-vectors represent different levels in the pyramid indexing of the wavelet coefficients. For instance, the vector Gy contains $N/2$ coefficients representing the level of finest detail. These elements are represented by $d_1 = (d_{1,0}, d_{1,1}, \ldots, d_{1,N/2-1})$.

In general, the level j of the wavelet decomposition of y is a vector that contains $N/2^j$ elements and represented by $d_j = GH^{j-1}y = (d_{j,0}, d_{j,1}, \ldots, d_{j,N/2^j-1})$. When the coefficients correspond to a "smooth" level rather than a "coarse" level of details, then the elements are typically denoted by $s_{j,k}$. For simplicity, in this paper we denote any type of wavelet coefficients as d_{jk}. The main strength of the DWT in statistics is that it induces local-in-time and time-space plane divisions that form unconditional basis for a range of function spaces.

Data Denoising and Bayes Inference

Let $y = (y_1, y_2, \ldots, y_N)$ be a vector of equally spaced observations of size N whose elements satisfy that

$$y_i = f_i + \varepsilon_i;\ i = 1, 2, \ldots, N.$$

f_i is the underlying signal generating the observed process and ε_i forms a sequence of independent and identically distributed errors with constant variance σ^2. This model can be rewritten in vector form as

$$y = f + \varepsilon$$

where $f = (f_1, f_2, \ldots, f_N)$ and $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N)$.

From a statistical point of view, the problem of *data denoising* is addressed as the problem of estimating the unknown vector $f = (f_1, f_2, \ldots, f_N)$ given the observed data y. After applying a DWT to y, we obtain the following model for the wavelet coefficients

$$d = \theta + \varepsilon'$$

where $d = Wy$, $\theta = Wf$, and $\varepsilon' = W\varepsilon$ with $\varepsilon' = (\varepsilon'_1, \varepsilon'_2, \ldots, \varepsilon'_N)$. If $\varepsilon_i \sim N(0, \sigma^2)$ and since W is an orthogonal matrix, the probability distribution for ε' corresponds to a multivariate Normal with a zero mean vector and covariance matrix $\sigma^2 I_{N\times N}$, i.e., $\varepsilon' \sim MVN(0, \sigma^2 I_{N\times N})$ where $I_{N\times N}$ denotes an identity matrix of dimension N. This implies that the probability distribution of d given θ and σ^2, $f(d|\theta, \sigma^2)$, is a multivariate Normal with mean vector θ and covariance matrix $\sigma^2 I_{N\times N}$. On the other hand, for every DWT there is an Inverse Discrete Wavelet Transformation (IDWT) that permits the reconstruction of the signal through the expression $f = W^t\theta$, where W^t is the transpose matrix of W. In consequence, inferences about θ based on the wavelet coefficients d automatically produce inferences on the signal of interest f.

The problem of estimating θ from a Bayesian perspective requires a prior probability distribution $p(\theta, \sigma^2)$. The joint posterior distribution for (θ, σ^2) given d is defined by

$$p(\theta, \sigma^2|d) \propto f(d|\theta, \sigma^2)p(\theta, \sigma^2).$$

This joint posterior distribution can be marginalized with respect to σ^2 to obtain the marginal posterior distribution of θ given d,

$$p(\theta|d) = \int_0^\infty p(\theta, \sigma^2|d)d\sigma^2.$$

A summary from this posterior distribution is typically used to define a *Bayes shrinkage*. For example, the posterior expectation $E(\theta|d)$ or the posterior median can be used as a point estimator of θ and then transformed back via the IDWT to the original signal-noise model. More formal theoretic decision approaches can define shrinkage rules based on loss function optimality criteria. In fact, the established Bayesian framework is quite general and does not rely on the assumption of Normal errors. However, the convenience of a Normal distribution assumption generally simplifies the calculations of the Bayes shrinkage.

Bayes Smooth Shrinkage

Different Bayes shrinkage methods depend on particular prior specifications. For example, Vidakovic (1998) proposed a Double Exponential prior distribution for each wavelet coefficient independently. Furthermore, the shrinkage problem is formulated as a decision theoretic problem and the goal is that the resulting optimal actions will mimic ”desirable thresholding rules”. In the case of this Double Exponential prior distribution, the shrinkage rule corresponds to a posterior expectation in closed form

that can be expressed in terms of a Laplace transformation. For the Adaptive Bayesian Wavelet Shrinkage (ABWS) proposed by Chipman and McCulloch (1997), a mixture of normals with different variances is used as a prior distribution for each wavelet coefficients. One of the variances is chosen to be near zero which approximates the situation of a point mass prior. The main advantage of this approach is that it provides closed expressions for the posterior distribution of interest and so the computations can be done quickly. The wavelet coefficients d_{jk} are modeled independently within and across levels. Therefore, Chipman and McCulloch (1997) start with the model $p(d_{jk}|\theta_{jk}, \sigma^2) = N(\theta_{jk}, \sigma^2)$. The prior on θ_{jk} is defined as a mixture of two Normal disributions where

$$\begin{aligned} p(\theta_{jk}|\gamma_j) &= \gamma_j N(0, (c_j\tau_j)^2) + (1-\gamma_j)N(0, \tau_j^2) \\ p(\gamma_j) &= p_j^{\gamma_j}(1-p_j)^{1-\gamma_j}; \gamma_j = 0, 1 \end{aligned}$$

so the γ_j's are independently distributed Bernoulli(p_j) variables. Since the hyperparameters p_j, c_j and τ_j depend on the level j to which the corresponding θ_{jk} (or d_{jk}) belongs and can be level-wise different, the method is adaptive. The Bayes rule under squared error loss for θ_{jk} has an explicit form

$$\hat{\theta}(d)_{jk} = \left[P(\gamma_j = 1|d_j)\frac{(c_j\tau_j)^2}{\sigma^2 + (c_j\tau_j)^2} + (1 - P(\gamma_j = 1|d_j))\frac{\tau_j^2}{\sigma^2 + \tau + j^2} \right] d_{jk}$$

where $P(\gamma_j = 1|d_j)$ is the posterior probability that $\gamma_j = 1$ and d_j is the vector of level j wavelet coefficients. A sophisticated empirical Bayes argument is used for tuning the hyperparameters level-wise.

The approach used by Clyde and Vidakovic (1998) is based on a mixture prior which allows for each wavelet coefficient to be zero with a positive probability (prior point mass) or to follow a Normal distribution,

$$p(\theta_{jk}|\gamma_j, \sigma^2) = N(0, (1-\gamma_j) + \gamma_j c_j \sigma^2)$$

for non-zero cases. The indicator (zero-one) variables, γ_j, define which basis element, i.e., column of W should be selected. As before the subscript j indicates the level to which θ_{jk} belongs. The prior distribution on σ^2 is an inverse χ^2 and the vector γ is formed by all the γ_j elements. The posterior mean for the vector of all the θ_{jk}'s is obtained by averaging over all $P(\gamma|d)$, the posterior probabilities of each γ given all the wavelet coefficients. Therefore,

$$E(\theta|d) = \sum_{\gamma} P(\gamma|d)E(\theta|d, \gamma)$$

where $E(\theta|d, \gamma)$ is the posterior expectation of θ given d and a specific γ. However, calculating the posterior probabilities of γ and the mixture estimate for the posterior mean implies summing over all 2^N values of γ. The calculations for such mixing is prohibitive even for problems of moderate size, and either approximations or stochastic methods for selecting subsets of γ that have a high posterior probability, must be used.

In a related paper, Clyde and George (2000) proposed a shrinkage approach that uses a hierarchical model with heavy-tailed error distributions that are a scale mixtures of normals. The prior specifications for some of the parameters in their model are difficult, hence, an *empirical Bayes* procedure is used to estimate these hyperparameters. These authors suggested various choices of priors and the posterior mean or median was used as an estimator of the wavelet coefficients. Furthermore, their methods allow to obtain threshold shrinkage estimators based on Bayesian model selection and multiple shrinkage estimators based on model averaging. An amazing volume of simulations were performed to justify the various error distributions and the various Bayesian models. All of these Bayesian models produce estimators as efficient (or better) than the ones based on traditional thresholding rules no matter whether the errors are Normal or not. To the best of the author's knowledge, the papers by Vidakovic (1998) and Clyde and George (2000) are the only ones that deal with non-normal errors for shrinkage problems in a Bayesian framework.

Multivariate Bayes Shrinkage

The main purpose in a *multivariate* Bayes shrinkage is to introduce a prior distribution that relaxes the unrealistic assumption of independence among wavelet coefficients. In particular, one can place a multivariate Normal-Inverse Gamma distribution for $p(\theta, \sigma^2)$ with $p(\theta|\sigma^2) = MVN(0, \sigma^2\Sigma)$ and $p(\sigma^2) = IG(\alpha, \delta)$. This model can incorporate dependence of the wavelet coefficients via an appropriate specification of the $N \times N$ prior covariance matrix Σ. In this case, the posterior distribution for (θ, σ^2) is also a Normal-Inverse Gamma distribution with hyperparameters m^*, Σ^*, α^* and δ^* where

$$\begin{aligned}
\Sigma^* &= (I_{N\times N} + \Sigma^{-1})^{-1}, \\
m^* &= \Sigma^* d, \\
\alpha^* &= \alpha + ||d|| + (m^*)^t(\Sigma^*)^{-1}(m^*), \\
\delta^* &= \delta + N.
\end{aligned}$$

Therefore a Bayes estimator for θ is m^*.

Building on this structure, Huerta (2005) proposed the following model for a multivariate Bayes shrinkage,

$$\begin{aligned}
p(d|\theta, \sigma^2) &= MVN(\theta, \sigma^2 I_{N\times N}), \\
p(\theta|\tau^2) &= MVN(0, \tau^2\Sigma),
\end{aligned}$$

where σ^2 and τ^2 are scale parameters and Σ is again a N dimensional matrix that induces prior correlations among the wavelet coefficients. In this model it is also assumed that $\sigma^2 \sim IG(\alpha_1, \delta_1)$ and $\tau^2 \sim IG(\alpha_2, \delta_2)$. The additional scale parameter τ^2 may induce some extra shrinkage to the wavelet coefficients leading to a more flexible Bayes wavelet shrinkage. This hierarchical prior implies a marginal multivariate t prior on θ, however there is no useful closed form expression for $E(\theta|d)$. Therefore

Huerta (2005) introduced a Gibbs sampler to generate posterior samples of $(\theta, \sigma^2, \tau^2)$ to approximate the posterior mean of θ. Specifically, the full conditional distribution for θ given τ^2, σ^2 and d, is a multivariate Normal distribution with vector mean m^* and covariance matrix Σ^*,

$$p(\theta|\tau^2, \sigma^2, d) \quad = \quad N(m^*, \Sigma^*)$$

where $m^* = \Sigma^* \sigma^{-2} d$ and $\Sigma^* = \left(\tau^{-2}\Sigma^{-1} + \sigma^{-2} I_{N\times N}\right)^{-1}$. The full conditional distribution of σ^2 given θ, τ^2 and d is an Inverse Gamma distribution with parameters α_1^* and δ_1^*,

$$p(\sigma^2|\theta, \tau^2, d) \quad = \quad IG(\alpha_1^*, \delta_1^*),$$

with $\alpha_1^* = N/2 + \alpha_1$ and $\delta_1^* = (\theta - d)^t(\theta - d)/2 + \delta_1$. Finally, the full conditional distribution for τ^2 given θ, σ^2 and d is an Inverse Gamma distribution with parameters α_2^* and δ_2^*,

$$p(\tau^2|\theta, \sigma^2, d) \quad = \quad IG(\alpha_2^*, \delta_2^*),$$

with $\alpha_2^* = N/2 + \alpha_2$ and ${\delta_2}^* = [\theta^t \Sigma^{-1} \theta]/2 + \delta_2$.

This method has shown to be robust for large errors in the observations. In particular, the multiresolution analysis of a sequence of measurements in Atomic Force Microscopy (AFM) and the reconstructed signal using the multivariate Bayes shrinkage appears in Figure 1. The rows indexed $d_1 - d_6$ are the values of the coefficients for the details, d_1 representing the finest level and d_6 the coarsest level. The row indexed by s_6 includes the coefficients of the smooth level. The Bayes shrinkage reduces the values for the fine levels of details towards zero while keeping almost intact coefficients corresponding to the smooth part or high levels of details s_6, d_6. Notice that the reconstructed signal is quite smooth while other shrinkage approaches such as *Sure Shrink* provide less smooth signal estimates with noisy artifacts. Similar data was analyzed via a Γ-Minimax wavelet shrinkage approach in Angelini and Vidakovic (2004) with comparable results to those shown here.

Hyperparameters and Laurent submatrices

For the case, $p(\theta|\sigma^2) = MVN(\theta, \sigma^2\Sigma)$, a reduction on the number of hyperparameters and imposing a hierarchical prior structure on some of the remaining hyperparameters was suggested by Vannucci and Corradi (1999). These authors applied their method, which they call *Bayes Shrink*, to density estimation and regression problems. Assuming that θ corresponds to an autoregressive process in time,Vannucci and Corradi (1999) demonstrate that the matrix Σ depends on only two hyperparameters, λ and ρ. The parameter ρ is the "autocovariance index" and λ is the precision parameter. The covariance matrix $\Sigma(\lambda, \rho) = \lambda\Sigma(\rho)$ has an interesting "finger like" structure. These authors suggest that,

$$\begin{aligned} p(\lambda) &= IG(p/2, q/2) \\ p(\rho) &\propto (C-\rho)^{r_1-1}(C+\rho)^{r_2-1}, |\rho| < C. \end{aligned}$$

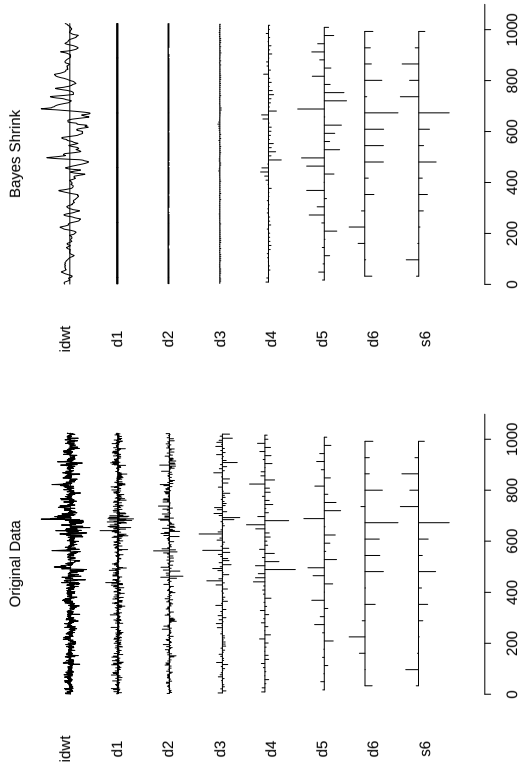

Figure 1: *DWT* of a Atomic Force Microscopy data and its Bayes Wavelet Shrinkage estimate

An alternative specification for the Σ matrix that also produces interesting shrinkage results is to fix it as a block diagonal matrix with each block of the form $\lambda_j \Sigma_j$. Each of the terms Σ_j defines a correlation structure inside the j-th level of coefficients for the wavelet decomposition. The values of λ_j are intended to tune the amount of shrinkage at level j. The block diagonal assumption for Σ establishes no correlation between coefficients at different levels of the wavelet decomposition of the data. Furthermore, Σ_j can be defined as a matrix with entries $\sigma_{i,s} = \rho^{|i-s|}$, i.e., the largest the difference between the sub-indexes i and s, the smaller the correlation between coefficients. ρ is a scalar quantity in $(0, 1)$. This type of specification was extensively applied in the modeling framework of Huerta (2005) and used to obtain Figure 1. Also, Vidakovic (1999) presents an example of "affine" or linear Bayes shrinkage via Laurent submatrices for the well known *galaxy velocities* data set.

A Full Bayes Model

The methods in previous sections are similar to the the traditional wavelet-shrinkage paradigm but with the wavelet coefficient being shrunk in a Bayesian fashion. Additionally Mueller and Vidakovic (1995) proposed a full Bayesian model for the wavelet coefficients targeted for density estimation problems. Their prior model explicitly defines geometrically decreasing prior probabilities for non-zero coefficients at higher levels of detail. An indicator variable performs the modeled induced thresholding and if a wavelet coefficient is not included, a *pseudo-prior* is assumed to avoid parameter spaces of varying dimension. MCMC techniques with *Metropolis-Hastings* updating steps can be used to sample the posterior distribution defined on the wavelet parameter space.

Conclusion

Bayes wavelet shrinkage methods provide powerful tools for data denoising problems. This paper reviews some of the most important Bayes wavelet shrinkage approaches and in particular, a multivariate Bayes shrinkage that relies on MCMC methods. Other Bayes shrinkage methods can be computationally less or more demanding, like empirical-Bayes approaches or Full Bayes methods, and more formally based on decision theoretic aspects.

References

Angelini, C. and Vidakovic, B. (2004). Gamma-minimax wavelet shrinkage: A robust incorporation of information about energy of a signal in denoising applications. *Statistica Sinica* **14**, 103–125.

Chipman, H. Kolaczyk, E. and McCulloch, R. (1997). Adaptive Bayesian wavelet shrinkage. *J. Amer. Statist. Assoc.* **92**, 1413–1421.

Clyde, M. Parmigiani, G. and Vidakovic, B. (1998). Multiple shrinkage and subset selection in wavelets. *Biometrika* **85**, 391–402.

Clyde, M. and George, E. (2000). Flexible empirical Bayes estimation for wavelets. *J. Roy. Statist.Soc. B* **62**, 681–698.

Daubechies, I. (1992). *Ten Lectures on Wavelets*. CBMS-NSF Regional Conference Series in Applied Mathematics. 3600 University City Science Center, Philadelphia, Pennsylvania 19104-2688: Society for Industrial and Applied Mathematics.

De Vore, R.A. Jawerth, B. and Lucier, B. J. (1992). Image compression through wavelet transform coding. *IEEE Transactions on Information Theory* **38**(2), 719–746.

Donoho, D. Johnstone, I. and Kerkyacharian, G. (1995). Wavelet shrinkage:asymptotia? *J. Roy. Statist. Soc. B* **57**(2), 301–369.

Efromovich, S. (1999). *Nonparametric Curve Estimation*. Springer Verlag, New York.

Huerta, G. (2005). Multivariate Bayes wavelet shrinkage and applications. *J. Appl. Statist.* **32**(5), 529–542.

Mueller, P. and Vidakovic, B. (1995). Bayesian inference with wavelets: Density estimation. *J.Comput. Graph. Statist.* **7**(4), 456–468.

Vannucci, M. and Corradi, F. (1999). Covariance structure on wavelet coefficients:theory and models in a Bayesian perspective. *J. Roy. Statist. Soc. B.* **61**(4), 971–986.

Vidakovic, B. (1998). Nonlinear wavelet shrinkage with Bayes rules and Bayes factors. *J. Amer. Statist. Assoc.* **93**, 173–179.

Vidakovic, B. (1999). *Statistical Modeling by Wavelets*. John Wiley and Sons, Inc., New York.