Curve Estimation using Wavelets and MDL

Thomas C. M. Lee[1]
The University of Chicago, Department of Statistics
5734 S. University Avenue, Chicago, IL 60637, USA
tlee@galton.uchicago.edu

This note presents a wavelet regression procedure for recovering functions corrupted by *correlated* noise. Two special features of the procedure are that, it imposes a so–called *tree constraint* on the wavelet coefficients, and that it uses the minimum description length (MDL) principle to define its "best" estimate. The full version of this note can be found in Lee (1999).

Background

Suppose we observe n equidistant noisy observations:

$$y_i = f_i + e_i, \quad f_i = f(\frac{i}{n-1}), \quad i = 0, \ldots, n-1,$$

where f is an unknown function, and the e_i's are noise. Our goal is to recover f using wavelet methods. We shall assume that $n = 2^{J+1}$ is an integer power of 2, and consider both the cases of independent and correlated noise.
Let $\boldsymbol{f} = (f_0, \ldots, f_{n-1})^T$, $\boldsymbol{y} = (y_0, \ldots, y_{n-1})^T$ and $\boldsymbol{W}$ be the discrete wavelet transform (DWT) matrix (e.g., see Donoho & Johnstone 1994). A wavelet–based estimate $\hat{\boldsymbol{f}}$ of $\boldsymbol{f}$ can be obtained by the following three steps:

1. compute the *empirical* wavelet coefficient vector $\boldsymbol{w} = \boldsymbol{W}\boldsymbol{y}$,
2. apply a filtering operation (e.g., thresholding) to $\boldsymbol{w}$ and obtain an *estimated* wavelet coefficient vector $\hat{\boldsymbol{w}}$ (usually some elements of $\hat{\boldsymbol{w}}$ are thresholded to zero),
3. apply the inverse DWT to $\hat{\boldsymbol{w}}$ and obtain $\hat{\boldsymbol{f}} = \boldsymbol{W}^T\hat{\boldsymbol{w}}$.

In the above $\boldsymbol{w}$ (and $\hat{\boldsymbol{w}}$) is a column vector with length n. We shall follow Donoho & Johnstone (1994) to use a double indexing scheme to label its elements:

$$\boldsymbol{w} = (\underbrace{w_{-1,0}}, \overbrace{w_{0,0}}, \underbrace{w_{1,0}, w_{1,1}}, \overbrace{w_{2,0}, w_{2,1}, w_{2,2}, w_{2,3}}, \ldots, w_{j,k}, \ldots, \underbrace{w_{J,0}, \ldots, w_{J,2^J-1}})^T.$$

Notice that there is a one–to–one correspondence between $\hat{\boldsymbol{w}}$ and $\hat{\boldsymbol{f}}$.

Using MDL for Wavelet Regression

In short the MDL principle *defines* the "best" model as the one that enables the best encoding (or compression) of the data; e.g., see Rissanen (1989). Here our data is $\boldsymbol{y}$. One way to encode $\boldsymbol{y}$ is to first "split" $\boldsymbol{y}$ into two components and then encode one component at a time: split $\boldsymbol{y}$ into a fitted model $\hat{\boldsymbol{w}}$ plus the corresponding residuals $\hat{\boldsymbol{e}} = \boldsymbol{y} - \hat{\boldsymbol{f}}$. If we use $L(a)$ to denote the *codelength* (amount of hardware memory required to store a) for the object a, we have

$$L(\boldsymbol{y}) = L(\hat{\boldsymbol{w}}) + L(\hat{\boldsymbol{e}}|\hat{\boldsymbol{w}}). \tag{1}$$

[1]Present address: Department of Statistics, Colorado State University, Fort Collins, CO 80523-1877, USA

Thus for our problem the MDL principle defines the best $\hat{\boldsymbol{w}}$ (and hence $\hat{\boldsymbol{f}}$) as the one that gives the smallest $L(\boldsymbol{y})$. In the above expression we have stressed that $\hat{\boldsymbol{e}}$ is "conditional" on $\hat{\boldsymbol{w}}$. Now we need expressions for $L(\hat{\boldsymbol{w}})$ and $L(\hat{\boldsymbol{e}}|\hat{\boldsymbol{w}})$.

Tree Constraint

To encode $\hat{\boldsymbol{w}}$ we need to encode (i) the indices (i.e., the locations) and (ii) the actual estimated values of those non–zero coefficients $\hat{w}_{j,k}$ in $\hat{\boldsymbol{w}}$. In this work we suggest using a tree to represent the indices of those non–zero elements of $\hat{\boldsymbol{w}}$. One advantage of using such a tree representation is that it agrees with the intuition that wavelet coefficients at a finer level (or resolution) should have a higher chance of being killed when comparing to those coarser level wavelet coefficients at the same relative location.
Define $\hat{w}_{j+1,2k}$ and $\hat{w}_{j+1,2k+1}$ as the *children* of $\hat{w}_{j,k}$, $0 \le j < J$. We also call $\hat{w}_{j,k}$ the *parent* of $\hat{w}_{j+1,2k}$ and $\hat{w}_{j+1,2k+1}$. We shall impose the constraint that, if $\hat{w}_{j,k}$ is killed, none of its children can survive. That is,

$$\hat{w}_{j,k} = 0, \;\; 0 \le j < J \quad \Rightarrow \quad \hat{w}_{j+1,2k} = 0 \;\text{ and }\; \hat{w}_{j+1,2k+1} = 0.$$

This sort of constraint, sometimes known as *tree constraint*, is just a severe executioner of the intuition mentioned above that wavelet coefficients at finer levels should have higher chances of being killed. See Fig. 1 for an example of tree representation.

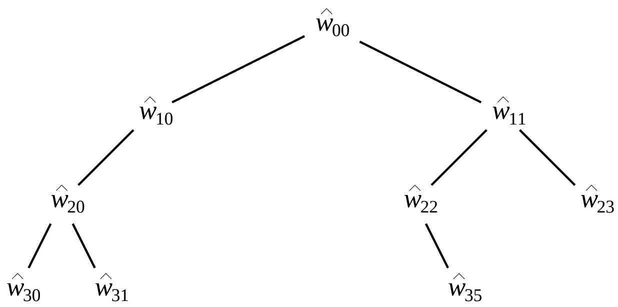

Figure 1: Tree Representation.

MDL–Based Wavelet Regression for Independent Noise

For comparison we first present an asymptotic expression for $L(\boldsymbol{y})$ when the noise e_i's is *independent* Gaussian:

$$\mathrm{MDL}_{\mathrm{IND}}(\hat{\boldsymbol{f}}) = \frac{\hat{m}}{2}\log_2 n + \frac{n}{2}\log_2 \frac{\|\hat{\boldsymbol{e}}\|^2}{n} \tag{2}$$

with $\hat{m}$ as the number of non–zero $\hat{w}_{j,k}$'s. When the noise is known to be independent Gaussian, we propose to estimate $\boldsymbol{f}$ by the minimizer of $\mathrm{MDL}_{\mathrm{IND}}(\hat{\boldsymbol{f}})$, subject to the condition that the tree constraint is satisfied. The derivation of $\mathrm{MDL}_{\mathrm{IND}}(\hat{\boldsymbol{f}})$ is reported in Lee (1999).

MDL–Based Wavelet Regression for Correlated Noise

Life is usually more complicated. Suppose that the noise e_i is correlated, and can be adequately modelled by an AR series of an unknown order p:

$$e_i = a_1 e_{i-1} + a_2 e_{i-2} + \ldots + a_p e_{i-p} + \tau_i,$$

where $\boldsymbol{a} = \{a_1, \ldots, a_p\}$ are unknown AR parameters and τ_i is a Gaussian innovation. Note that it reduces to the independent noise case when $p = 0$.
It is shown in Lee (1999) that, when the noise is autoregressively correlated, $L(\boldsymbol{y})$ can be approximated by

$$\mathrm{MDL}_{\mathrm{COR}}(\hat{\boldsymbol{f}}) = \frac{\hat{m} + \hat{p}}{2} \log_2 n + \frac{n}{2} \log_2 \frac{\|\hat{\boldsymbol{\tau}}\|^2}{n}, \tag{3}$$

where $\hat{p}$ and $\hat{\boldsymbol{\tau}}$ are estimates of p and $\boldsymbol{\tau} = (\tau_0, \ldots, \tau_{n-1})^T$ respectively.
When the noise is correlated, we propose to estimate $\boldsymbol{f}$ (and p and $\boldsymbol{\tau}$) by the minimizer of $\mathrm{MDL}_{\mathrm{COR}}(\hat{\boldsymbol{f}})$, subject to the tree constraint.
Observe that $\mathrm{MDL}_{\mathrm{COR}}(\hat{\boldsymbol{f}})$ generalizes $\mathrm{MDL}_{\mathrm{IND}}(\hat{\boldsymbol{f}})$, as $\mathrm{MDL}_{\mathrm{COR}}(\hat{\boldsymbol{f}})$ reduces to $\mathrm{MDL}_{\mathrm{IND}}(\hat{\boldsymbol{f}})$ when $\hat{p} = 0$. For this reason, if one is not sure whether the noise is independent or correlated, we recommend using $\mathrm{MDL}_{\mathrm{COR}}(\hat{\boldsymbol{f}})$ as the target.
In Lee (1999) a tree–growing algorithm for approximating the minimizer of $\mathrm{MDL}_{\mathrm{COR}}(\hat{\boldsymbol{f}})$ (and $\mathrm{MDL}_{\mathrm{IND}}(\hat{\boldsymbol{f}})$) is presented.

Simulation Results: Independent Noise

Here we report some simulation results regarding the relative practical performances of four wavelet regression methods when the noise is *independent*. These four methods were:

1. the SURE thresholding procedure of Donoho & Johnstone (1995),
2. the Bayesian thresholding procedure of Abramovich, Sapatinas & Silverman (1998),
3. the MDL–based thresholding (MDL–Thresh) procedure of Saito (1994), and
4. the proposed tree-growing procedure (MDL–Tree) which aims to minimize $\mathrm{MDL}_{\mathrm{COR}}(\hat{\boldsymbol{f}})$ (*not* the independent criterion $\mathrm{MDL}_{\mathrm{IND}}(\hat{\boldsymbol{f}})$).

Five test functions were used: the four test functions of Donoho & Johnstone (1994), and the broken polynomial test function of Nason & Silverman (1994). Also, throughout all our numerical experiments, Daubechies' order 4 "least asymmetric" wavelet was used.
We define signal–to–noise ratios (snrs) as: snr $= \|\boldsymbol{f}\|/\sigma$, and used three levels: high snr=9, medium snr=7 and low snr=5. For each combination of test function and snr, 50 sets of noisy observations were simulated, and the number of data points n for each data set was 512. We used one n: it is because n and snr are interchangeable, as the asymptotic performance depends on σ^2/n.
For each simulated data set, we applied the four wavelet methods listed above to estimate the test function. The numerical measure that we used to evaluate the quality of an $\hat{\boldsymbol{f}}$ was $\mathrm{MISE}(\hat{\boldsymbol{f}}) = \|\boldsymbol{f} - \hat{\boldsymbol{f}}\|^2/n$.
We performed paired Wilcoxon tests to test if the difference between the median $\mathrm{MISE}(\hat{\boldsymbol{f}})$ values of two wavelet methods is significant or not; the significance level used was 1.25%. We then ranked the four methods using these test results. Of course, ranking the methods in this manner is not perfectly legitimate, but it does provide an indicator about the relative merits of the methods (see Wand 1997).
The average paired Wilcoxon test rankings for the SURE, Bayes, MDL–Thresh and MDL–Tree procedures are 3.03, 1.77, 3.47 and 1.73 respectively (with 1 being the best). Judging by this measure, one may conclude that the Bayes and MDL–Tree procedures performed equally

well, and are superior to the other two procedures. Also, the SURE procedure seems to have performed better than the MDL–Thresh procedure.

Simulation Results: Correlated Noise

Here we are interested in the performances of the SURE procedure and the proposed MDL–Tree procedure (which aims to minimize $\mathrm{MDL}_{\mathrm{COR}}(\hat{\boldsymbol{f}})$) under the situations that the noise is correlated. We note that Johnstone & Silverman (1997) show that the SURE procedure is also applicable when the noise is correlated.

The setup for this correlated noise experiment was essentially the same as the independent noise experiment, with the exception that the noise $\boldsymbol{e} = (e_1, \ldots, e_n)^T$ was generated from the ARMA (p, q) model: $e_i = a_1 e_{i-1} + \ldots + a_p e_{i-p} + \tau_i + b_1 \tau_{i-1} + \ldots + b_q \tau_{i-q}$, where τ_i is a Gaussian innovation. Three different types of ARMA noise were considered: AR (1) with $a_1 = -0.8$, AR (2) with $a_1 = 4/3$ and $a_2 = -8/9$, and ARMA (1,1) with $a_1 = 0.2$ and $b_1 = -0.9$. Throughout the whole the experiment, the noise was always linearly stretched so that $\{\max(\boldsymbol{f}) - \min(\boldsymbol{f})\}/\{\max(\boldsymbol{e}) - \min(\boldsymbol{e})\} = 5$. Note that Johnstone & Silverman (1997) also use the same AR (2) noise in their numerical examples, and that the ARMA (1,1) case does not satisfy the assumption made by $\mathrm{MDL}_{\mathrm{COR}}(\hat{\boldsymbol{f}})$.

The average paired Wilcoxon test rankings for the SURE and MDL–Tree procedures are 1.77 and 1.23 respectively: some evidence suggesting that MDL–Tree is superior.

References

Abramovich, F., Sapatinas, T. & Silverman, B. W. (1998), 'Wavelet thresholding via a Bayesian approach', Journal of the Royal Statistical Society Series B **60**, *725–749.*

Donoho, D. L. & Johnstone, I. M. (1994), 'Ideal spatial adaptation by wavelet shrinkage', Biometrika **81**, *425–455.*

Donoho, D. L. & Johnstone, I. M. (1995), 'Adapting to unknown smoothness via wavelet shrinkage', Journal of the American Statistical Association **90**, *1200–1224.*

Johnstone, I. M. & Silverman, B. W. (1997), 'Wavelet threshold estimators for data with correlated noise', Journal of the Royal Statistical Society Series B **59**, *319–351.*

Lee, T. C. M. (1999), 'Tree–based wavelet regression for correlated data using the minimum description length principle'. Unpublished manuscript.

Nason, G. P. & Silverman, B. W. (1994), 'The discrete wavelet transform', Journal of Computational and Graphical Statistics **3**, *163–191.*

Rissanen, J. (1989), Stochastic Complexity in Statistical Inquiry, *World Scientific, Singapore.*

Saito, N. (1994), Simultaneous noise suppression and signal compression using a library of orthonormal bases and the minimum description length criterion, in *E. Foufoula-Georgiou & P. Kumar, eds, 'Wavelets in Geophysics', New York: Academic Press, pp. 299–324.*

Wand, M. P. (1997), 'A comparison of regression spline smoothing procedures'. Unpublished Manuscript.