ON CHOOSING A NON-INTEGER RESOLUTION LEVEL WHEN USING WAVELET METHODS

Peter Hall[1], Guy P. Nason[1,2]

ABSTRACT. In curve estimation using wavelet methods it is common to select the resolution level to be an integer, so as to exploit the computational advantages of the pyramid or cascade algorithm. This choice, however, can produce a noticeable amount of either oversmoothing or undersmoothing. Its analogue for estimation by kernel methods is to restrict the bandwidth to be an integer power of $\frac{1}{2}$, which would seldom be acceptable. In this note we quantify the advantages of non-integer resolution levels.

KEYWORDS. Bandwidth, curve estimation, density estimation, dyadic expansion, mean squared error, kernel estimator, nonparametric regression.

SHORT TITLE. Smoothing wavelet estimators.

AMS (1991) SUBJECT CLASSIFICATION. Primary 62G07, Secondary 62G30.

[1] Centre for Mathematics and its Applications, Australian National University, Canberra, ACT 0200, Australia.

[2] Department of Mathematics, University of Bristol, Bristol BS8 1TW, U.K.

1. INTRODUCTION

Wavelet methods offer excellent adaptivity and computational efficiency in a variety of applications to nonparametric curve estimation. For example, traditional techniques for estimating piecewise-smooth functions with jump discontinuities typically require first an algorithm for identifying the location, and perhaps also the sizes, of jumps; and then a smoother for estimating the function between jumps. Nonlinear wavelet methods, on the other hand, accomplish the entire operation in a single step, and with greater computational ease (see e.g. Donoho, Johnstone, Kerkyacharian and Picard 1995). In this note we argue that the degree of adaptivity that is often forced on the wavelet smoothing parameter by the method of computation does not allow as much adaptivity to variation in a *smooth* target function as is permitted by kernel methods. It may be likened to demanding that the bandwidth of a kernel estimator be taken equal to 2^{-j} for an integer j.

We quantify the penalty that is paid for restricting smoothing to a dyadic grid. It is convenient to focus attention on the case of density estimation, as it involves few structural assumptions, but our conclusions in the case of nonparametric regression are identical. Let $X_1, \ldots, X_n$ denote a random sample from the distribution with density f, and suppose f admits the wavelet expansion (Meyer, 1992)

$$f = \sum_{-\infty<j<\infty} b_j\, \phi_j + \sum_{i=0}^{\infty} \sum_{-\infty<j<\infty} b_{ij}\, \psi_{ij}\,,$$

where $\phi_j(x) = p^{1/2}\, \phi(px+j)$, $\psi_{ij}(x) = p_i^{1/2}\, \psi(p_i x+j)$, ϕ and ψ are the "father" and "mother" wavelet functions (assumed orthonormal), $\log_2 p = \log_2 p(n)$ is the most coarse resolution level of the fitted wavelet, $p_i = 2^i\, p$, $b_j = \int f\, \phi_j$ and $b_{ij} = \int f\, \psi_{ij}$. (The logarithm function to base 2 is denoted by $\log_2$.) Unbiased estimators of b_j and b_{ij} are given by

$$\hat{b}_j = n^{-1} \sum_{m=1}^{n} \phi_j(X_m) \quad \text{and} \quad \hat{b}_{ij} = n^{-1} \sum_{m=1}^{n} \psi_{ij}(X_m)\,.$$

Substituting them into the theoretical expansion of f, truncating the sum over i so as to ensure convergence, and applying a threshold to the coefficients $\hat{b}_{ij}$ so as to exclude those that are principally noise, we obtain an empirical wavelet expansion,

$$\hat{f} = \sum_{-\infty<j<\infty} \hat{b}_j \, \phi_j + \sum_{i=0}^{q-1} \sum_{-\infty<j<\infty} \hat{b}_{ij} \, I(|\hat{b}_{ij}| \geq \delta) \, \psi_{ij} \, . \tag{1.1}$$

Here, $\delta = \text{const.} \, (n^{-1} \log n)^{1/2}$ denotes the threshold, and q is an integer that is typically chosen so that $p \, 2^q$ is close to n.

Critically, p plays the role of the inverse of bandwidth in determining the performance of $\hat{f}$. In particular, if f has r derivatives in a piecewise sense, if the wavelet ψ is of order r (that is $\int x^m \psi(x) \, dx = 0$ for $m = 0, \ldots r-1$), and if the threshold δ and truncation parameter q are chosen appropriately, then the estimator defined at (1.1) enjoys the following expansion of mean integrated squared error:

$$\int E(\hat{f} - f)^2 = n^{-1} \, p + C \, p^{-2r} + o\big(n^{-1} \, p + p^{-2r}\big) \tag{1.2}$$

as $n, p \to \infty$. Here, the constant C depends on only f and ψ, and the first two terms on the right-hand side represent the main contributions from variance and squared bias, respectively. See for example Hall and Patil (1995), and Hall and Patil (1996) for the counterpart in the case of nonparametric regression. One might possibly expect something other than (1.2), because of the nonlinear nature of a wavelet estimator. That this classical decomposition into variance and squared bias is valid in the nonlinear case reflects the fact that the choice of threshold is relatively conservative in places where f is smooth, and so has negligible effect there. To first order its influence is felt only in places where f has a discontinuity in a derivative of lower order than that of the wavelet. There the influence is very local, acting to adjust the bias contribution so that the estimate correctly tracks the true curve, and preserving the classical variance–bias decomposition of mean integrated squared error.

Result (1.2) is a direct analogue of the mean integrated squared error expansion of an r'th order kernel estimator with bandwidth h, where

$$\int E(\hat{f} - f)^2 = C_1 \, (nh)^{-1} + C_2 \, h^{2r} + o\big\{(nh)^{-1} + \, h^{2r}\big\} \tag{1.3}$$

as $n, h^{-1} \to \infty$, and C_1, C_2 depend only on f and the kernel. See for example Wand and Jones (1995, p. 21). (On the occasion of (1.3) it is assumed that f is r times differentiable, not just in a piecewise sense.)

It is common to take $\log_2 p$ to be an integer, so as to make use of Mallat's pyramid algorithm for calculating $\hat{f}$. A comparison of (1.2) and (1.3) reveals, however, that this dyadic choice is tantamount to insisting that in kernel smoothing, h be taken equal to 2^{-j}. Such a restriction would seldom be acceptable; indeed, some contemporary bandwidth selectors produce empirical choices of h that are root-n consistent for the optimal bandwidth, in relative terms. We argue that, insofar as one is interested in accurate estimation of the smooth part of f, dyadic choice of p is often not acceptable for wavelet estimators either. The computational labour of smoothing in the continuum, or at least on a fine grid, is usually insignificant for practical density estimation. It only becomes an issue when one is conducting a simulation study with many replications.

2. PENALTY FOR DISCRETE CHOICE OF SMOOTHING PARAMETER

2.1. Optimal choice of smoothing parameter. Consider a density estimation problem where the asymptotic mean squared error, or mean integrated squared error, may be written as

$$M_n(h) = c_1(nh)^{-1} + c_2 h^{2r} ,$$

in which n denotes sample size, c_1 and c_2 are positive constants, and $r \geq 1$ is a fixed integer representing the "order" of the method. Since we are interested only in proportionate changes to M_n then we may work with M_n/c_2 instead of M_n. Thus, defining c by $c_1 = 2rcc_2$, we may suppose without loss of generality that

$$M_n(h) = 2rc\,(nh)^{-1} + h^{2r} . \tag{2.1}$$

In this case the minimum value of M_n is achieved with $h(n) = (c/n)^{1/(2r+1)}$, and equals $M^0 = (2r+1)\,(c/n)^{2r/(2r+1)}$.

In the wavelet case, h does not always vary in the continuum. For reasons of computational efficiency, in particular to utilise the pyramid or cascade algorithm of Mallat, h is often restricted to a geometric sequence, typically to $\{2^j, -\infty < j < \infty\}$ although more generally to $\mathcal{S}_d = \{d^j, -\infty < j < \infty\}$, where $d > 1$; see for example Auscher (1989, 1992). Let $v_d \in (0, 1/2)$ denote the solution of

$$d^{(2r+1)v_d}\left(d^{2r} - 1\right) = 2rd^{2r}\,(d-1)\,,$$

let $j_d(n)$ be the integer nearest to $-\log_d h(n)$ (the smaller of the two integers if there is a tie), and put $u = u(n) = j_d(n) + \log_d h(n) \in (-1/2, 1/2]$. Here, $\log_d$ denotes the logarithm function to base d.

Proposition 2.1. *For each $d > 1$, the value of h that minimizes M_n (defined at (2.1)) when the argument is constrained to be in $\mathcal{S}_d$, equals $d^{-k_d(n)}$ where*

$$k_d(n) = \begin{cases} j_d(n) & \textit{if } u > -v_d \\ j_d(n)+1 & \textit{if } u < -v_d. \end{cases}$$

When $u = -v_d$ the choices $k_d(n) = j_d(n)$ and $k_d(n) = j_d(n) + 1$ are equally appropriate, since there, $M_n(d^{-j_d(n)}) = M_n(d^{-j_d(n)-1}) < M_n(d^{-j})$ for all $j \notin \{j_d(n), j_d(n)+1\}$.

To derive the proposition, observe that $M_n(d^{-j_d(n)}) = M^0\, a(u)$, where $a(u) = (2r+1)^{-1}(2rd^u + d^{-2ru})$. The function a is strictly increasing on $(0, \infty)$ and strictly decreasing on $(-\infty, 0)$. Furthermore, $a(-t) > a(t)$ for all $t > 0$, and $a(u) > a(u+1)$ [respectively, $a(u) = a(u+1)$] if and only if $u < -v_d$ [$u = -v_d$]. The proposition follows from these properties.

2.2. The penalty, as a function of n, for smoothing discretely. Let

$$\rho(n) = M_n(d^{-k_d(n)})/M^0$$

denote the penalty for smoothing on the grid $\mathcal{S}_d$ rather than in the continuum. To appreciate the size of $\rho(n)$ we examine the worst and average cases, respectively.

Now,

$$\begin{aligned}\rho(n) \leq \rho_{\max} &= a(1-v_d)\\ &= (2r+1)^{-1}\Big(2rd\left[\left(d^{2r}-1\right)/\left\{2rd^{2r}\,(d-1)\right\}\right]^{1/(2r+1)}\\ &\qquad + d^{-2r}\left\{2rd^{2r}\,(d-1)/\left(d^{2r}-1\right)\right\}^{2r/(2r+1)}\Big)\,, \qquad (2.2)\end{aligned}$$

and $\rho(n)$ can be rendered arbitrarily close to $\rho_{\max}$ along a sequence of values of n diverging to infinity. To appreciate the average size of $\rho(n)$, write

$$n = c\,h(n)^{-(2r+1)} = cd^{(2r+1)j_d(n)}\,W\,,$$

where $W = d^{-(2r+1)u}$. Since integers n are equally spaced then we shall consider W to be uniformly distributed on its range, i.e. on $(d^{-(2r+1)/2}, d^{(2r+1)/2})$. (Even in the context of regression, wavelet methods may be used for arbitrary values of n; interpolation techniques may be employed to overcome the traditional need for a dyadic sample size.) Let $V = -(\log_d W)/(2r+1)$ and $U = V\,I(V > -v_d) + (V + 1)\,I(V \leq -v_d)$, in which notation $\rho(n) = a(U)$. Expressing $\rho(n)$ in terms of W we obtain:

$$\begin{aligned}\rho(n) = (2r+1)^{-1}\Big\{&\left(2r\,W^{-1/(2r+1)} + W^{2r/(2r+1)}\right) I\left(W < d^{v_d(2r+1)}\right)\\ &+ \left(2rd\,W^{-1/(2r+1)} + d^{-2r}\,W^{2r/(2r+1)}\right) I\left(W \geq d^{v_d(2r+1)}\right)\Big\}\,.\end{aligned}$$

Direct calculation shows that the expected value of the right-hand side equals

$$\begin{aligned}\rho_{\rm av} = \left(d^{(2r+1)/2} - d^{-(2r+1)/2}\right)^{-1}\Big[&d^{2rv_d} - d^{-r} + d\left(d^{r} - d^{2rv_d}\right)\\ &+ (4r+1)^{-1}\Big\{d^{(4r+1)v_d} - d^{-(4r+1)/2}\\ &+ d^{-2r}\left(d^{(4r+1)/2} - d^{(4r+1)v_d}\right)\Big\}\Big]\,, \qquad (2.3)\end{aligned}$$

representing the average value of $\rho(n)$.

r	ρ_{av}	$\rho_{\max}$	v_2
2	1.118	1.237	0.419
4	1.269	1.413	0.334
6	1.358	1.525	0.276
$r \to \infty$	$2^{1/2}$	2	$(\log_2 r)/(2r)$

Table 1: Values of average (ρ_{av}) and largest ($\rho_{\max}$) factors by which $M_n(d^{-j_d(n)})$ exceeds M^0, and values of v_2, for $r = 2, 4, 6$ and in the case of a dyadic smoothing parameter ($d = 2$). The last row provides asymptotic formulae as $r \to \infty$.

Table 1 gives the values of ρ_{av}, $\rho_{\max}$ and v_d in the cases $r = 2, 4, 6$, and for $d = 2$. Note that for large r, v_2 is close to zero, indicating that k_2 is equal to $j_2 + 1$ almost as often as it is to j_2.

2.3. Alternative expressions for the penalty. The penalty, ρ, may be regarded as a function of the error variance, in the context of nonparametric regression; or of the scale of the true density, in the case of density estimation; or of other variable parameters. Such views of the problem do not alter the value of $\rho_{\max}$, but they may affect ρ_{av}, depending on how "average" is defined.

For example, in the case of nonparametric regression when the error variance σ^2 is considered a variable parameter, the variance contribution to mean squared error is proportional to $(nh)^{-1}\,\sigma^2$ rather than simply $(nh)^{-1}$. See Hall and Patil (1996). All the arguments given earlier remain valid if we replace n by n/σ^2 throughout. In particular, the expression for $\rho_{\max}$ at (2.2) is still correct. So also is that for ρ_{av} at (2.3), provided the average is taken uniformly over values of σ^{-2} within an interval $(\sigma_0^{-2}, d\,\sigma_0^{-2})$ for some $\sigma_0 > 0$. Alternative definitions of ρ_{av}, of which there are of course a great many, will produce different values of ρ_{av}, all of them lying in the interval $(1, \rho_{\max})$.

In the case of nonparametric density estimation where the density $f = f_\sigma$ has variance σ^2, the bias contribution to mean squared error is proportional to $\sigma^{2r+1}h^{2r}$ rather than h^{2r}. See for example Hall and Patil (1995). The arguments in Sections 2.1 and 2.2 continue to hold if we replace n by $n\sigma^{2r+1}$. Thus, $\rho_{\max}$ is again given by (2.2), and ρ_{av} is given by (2.3) if the average is taken uniformly over values of σ^{2r+1} in an interval $(\sigma_0^{2r+1}, d\,\sigma_0^{2r+1})$ for some $\sigma_0 > 0$. This definition of "average" is contrived, but versions of (2.3) using more natural definitions are readily constructed.

To elucidate the case of density estimation we conducted a simulation study where the sampling density f_σ was Normal with mean zero and variance σ^2. Sample size was fixed at $n = 256$ throughout; we varied σ. The threshold δ in the wavelet estimator at (1.1) was taken to be $(Cn^{-1}\log n)^{1/2}$, where $C = 2\,\mathrm{sup} f_\sigma = (2/\pi)^{1/2}\sigma^{-1}$. The truncation point q was set equal to the integer part of the base 2 logarithm of n/p. Mean integrated squared error, rather than pointwise mean squared error, was taken as the measure of risk, and was evaluated by averaging 500 simulated density estimators calculated at 1000 grid points spread evenly over their finite support. All wavelet computations were conducted using *WaveThresh* Version 3 (Nason, 1995). In particular, this meant that the wavelets used enjoyed $r = 6$, and that $d = 2$.

The optimal smoothing parameter $p_{\text{opt}} = h_{\text{opt}}^{-1}$ was calculated by minimising this mean squared error approximation using the `FMIN` procedure from the netlib `GO` package (Brent, 1973). The value of mean integrated squared error computed at the optimal smoothing parameter was taken as the numerical approximation, $\bar{M}^0$, to M^0. The minimum mean integrated squared error for a dyadic choice of the smoothing parameter was calculated by inspecting the relatively small range of possibilities. The ratio of this to $\bar{M}^0$ was our numerical approximation to the penalty, ρ, for the given values of n and σ.

Table 2 lists values of ρ for different σ's, indicating that in this rather different setting the range of values of the penalty is nevertheless similar to that suggested by Table 1. The values of ρ listed in Table 2 were each calculated as the mean of at least 10 different independent trials of the procedure described in the previous two paragraphs.

σ	ρ
0.6	1.647 (0.043)
1.0	1.250 (0.077)
1.5	1.019 (0.059)
2.0	1.218 (0.048)
3.0	1.017 (0.025)
4.0	1.211 (0.042)
5.0	1.457 (0.034)

Table 2: Numerical approximations to ρ in the case $n = 256$, for various values of σ. Standard deviations are given in parentheses.

Figure 1: Graphs of true $\mathrm{N}(0, (32/3)^2)$ density (thin unbroken line), optimally smoothed wavelet estimator (thick unbroken line), and density estimators for the nearest dyadic resolution levels below (broken curve with peak above true density) and above (broken curve with peak below true density) the optimum. Sample size is $n = 256$.

Next, again in the case of density estimation when the true density is Normal $\mathrm{N}(0, \sigma^2)$, with sample size 256, we illustrate a typical sample. We took $\sigma = 32/3$. Figure 1 depicts the true density, indicated by the thin unbroken line, and the wavelet density estimator with optimal smoothing parameter $p_{\mathrm{opt}} = 0.094$, indicated by the thick unbroken line. The nearest dyadic values are $1/8 = 0.125$ and $1/16 = 0.0625$, and the corresponding density estimators are indicated in the figure by broken lines, that for $p = 1/8$ having the lower peak. Compared with the optimally smoothed estimator, these "dyadic estimators" show a marked inability to correctly resolve the true peak. Thus, restriction to a dyadic smoothing parameter significantly affects visual as well as mathematical performance.

2.4. Empirical choice of primary resolution level. The close parallel between mean integrated squared error formulae for wavelet and kernel methods (see (1.2) and (1.3)) suggests that relatively classical techniques, such as those based on cross-validation, the bootstrap or plug-in rules, may be employed to select an empirical version of the appropriate smoothing parameter, p. Indeed, theoretical arguments that are entirely similar to those in classical curve estimation problems may be used to prove that this is the case for smooth targets. Plug-in rules are arguably less attractive for relatively unsmooth (e.g. discontinuous) targets, since the constant C in (1.2) depends on piecewise integrals of $(f^{(r)})^2$ over sets where it is well-defined; see Hall and Patil (1995). However, cross-validation and bootstrap methods are relatively attractive there, the latter having important similarities to their counterparts in kernel-based nonparametric regression (e.g. Faraway and Jhun 1990).

Acknowledgement. The helpful comments of a reviewer are gratefully acknowledged.

REFERENCES

AUSCHER, P. (1989). *Ondelettes Fractales et Applications.* PhD Thesis, Université de Paris, Dauphine, Paris, France.

AUSCHER, P. (1992). Wavelet bases for $L^2(\mathbb{R})$, with rational dilation factor. In: *Wavelets and their Applications*, eds M.B. Ruskai, G. Beylkin, R. Coifman, I. Daubechies, S. Mallat, and L. Raphael, pp. 439–452. Jones & Bartlett, Boston.

BRENT, R.P. (1973). *Algorithms for Minimization Without Derivatives.* Prentice Hall, Englewood Cliffs.

DONOHO, D.L., JOHNSTONE, I.M., KERKYACHARIAN, G. AND PICARD, D. (1995). Wavelet shrinkage: asymptopia? (With discussion.) *J. Roy. Statist. Soc. Ser. B* **57**, 301–369.

FARAWAY, J.J. AND JHUN, M. (1990). Bootstrap choice of bandwidth for density estimation. *J. Amer. Statist. Assoc.* **85**, 1119–1122.

HALL, P. AND PATIL, P. (1995). Formulae for mean integrated squared error of nonlinear wavelet-based density estimators. *Ann. Statist.* **23**, 905–928.

HALL, P. AND PATIL, P. (1996). On the choice of smoothing parameter, threshold and truncation in nonparametric regression by nonlinear wavelet methods. *J. Roy. Statist. Soc. Ser. B* **58**, 361–377.

MEYER, Y. (1992). *Wavelets and Operators.* Cambridge University Press, Cambridge.

NASON, G.P. (1995). WaveThresh Version 3. *Technical Report, Department of Mathematics, University of Bristol.*

WAND, M.P. AND JONES, M.C. (1995). *Kernel Smoothing.* Chapman & Hall, London.