Statistica Sinica **11**(2001), 691-704

THRESHOLDING FOR WEIGHTED χ^2

Iain Johnstone

Stanford University

Abstract: Given data from a spherical Gaussian distribution with unknown mean vector θ, estimates of quadratic functionals are constructed by thresholding. Mean squared error bounds are derived via a comparison with those already available for a suitable noncentral χ^2 variate. By way of illustration, the resulting inequalities are used to yield an optimal rate adaptivity result for estimation of integrated squared derivatives in the white noise model of nonparametric function estimation.

Key words and phrases: Adaptive estimation, integrated squared derivative, noncentral χ^2, quadratic functional.

1. Introduction

Consider estimation of a quadratic functional $\rho_\alpha = \sum_{k=1}^d \alpha_k \theta_k^2, \alpha_k \geq 0$ on the basis of independent Gaussian data $y_k \sim N(\theta_k, \epsilon^2)$, $k = 1, \ldots, d$, with unknown mean θ and known variance ϵ^2. Such quadratic functionals occur throughout parametric and non-parametric statistics, for example in the analysis of variance (power analysis and variance components), spectral estimation and bandwidth selection. For examples, see Johnson and Kotz (1970, Chapter 29) and Section 4.

The standard unbiased estimate is $R_\alpha = \sum_1^d \alpha_k(y_k^2 - \epsilon^2)$. If ρ_α is thought to be small, this is unattractive due to high variability and significant probability of a negative estimate. Thresholding provides a simple remedy for both these defects: set $x_+ = \max(x, 0)$ and

$$\hat{\rho}(R_\alpha; t) = (R_\alpha - t\epsilon^2)_+. \tag{1.1}$$

With appropriate choice of threshold $t > 0$, one may hope for much better estimation for small ρ_α combined with acceptable properties in the event that ρ_α is not small. The purpose of this note is to give mean squared error bounds of this flavor, so long as the weights $\{\alpha_k\}$ are comparable in magnitude.

In the simplest, symmetric, case $\rho = \sum_1^d \theta_k^2$ and the unbiased estimate $R = \sum_1^d (y_k^2 - \epsilon^2)$ is, up to a constant, a non-central χ^2 variate. An earlier paper, Johnstone (2000) ([J] below) derived distributional and mean-squared error bounds for this setting.

For the more general functional ρ_α, the estimate R_α is now essentially a weighted combination of non-central χ^2 variates. This note develops a comparison with thresholding of an appropriate single non-central χ^2 variate so that the bounds of [J] may be applied. The resulting comparison bounds are insensitive to d, ϵ and θ and depend on the weights α_k only through the imbalance $\bar{\alpha} = \max \alpha_k / \min \alpha_k$. Suppose, then, that $\min \alpha_k = 1$ and

$$1 \leq \alpha_k \leq \bar{\alpha}. \tag{1.2}$$

Section 1 derives the comparison bound for mean qth power error. Section 2 recalls necessary results on the balanced case and derives illustrative consequences of the bound obtained in Section 1. Section 3 gives an application to estimation of nonparametric functionals involving derivatives, $\int (D^l f)^2$.

2. Comparison Bound for MSE

The main tool in fact holds for all L_q error measures, $q > 0$.

Theorem 2.1. *With the preceding notations, there exists an absolute constant γ such that for all $d, \alpha, \theta, \epsilon^2$ and t,*

$$E|\hat{\rho}(R_\alpha; \bar{\alpha} t) - \rho_\alpha|^q \leq \gamma \bar{\alpha}^q E|\hat{\rho}(R; t) - \rho|^q. \tag{2.1}$$

The proof of Theorem 2.1 uses a risk comparison based on a tail domination condition.

Proposition 2.2. *Let U and U' be random variables with means $0 \leq \mu \leq \mu'$ and distribution functions F and F'. Set $\tilde{F}(u) = 1 - F(u)$ and $\tilde{F}'(u) = 1 - F'(u)$. Assume that $F(-u)$ and $\tilde{F}(u) = o(u^{-q})$ as $u \to \infty$, and similarly for F'. Suppose there exists $\gamma > 0$ such that*

$$F(\mu + s) \leq \gamma F'(\mu' + s), \qquad \tilde{F}(\mu + s) \leq \gamma \tilde{F}'(\mu' + s), \qquad s \in \mathbb{R}. \tag{2.2}$$

Let $\delta(u) = (u - \tau)_+$ denote soft thresholding. Then

$$E|\delta(U) - \mu|^q \leq \gamma E|\delta(U') - \mu'|^q. \tag{2.3}$$

Proof of Proposition 2.2. Integration by parts gives

$$\begin{aligned}
&E|\delta(U) - \mu|^q \\
&= q \int_{-\infty}^{\tau+\mu} |\delta(u) - \mu|^{q-1} \delta'(u) F(u) du + q \int_{\tau+\mu}^{\infty} |\delta(u) - \mu|^{q-1} \delta'(u) \tilde{F}(u) du \\
&= q \int_{\tau}^{\tau+\mu} |u - \tau - \mu|^{q-1} F(u) du + q \int_{\tau+\mu}^{\infty} |u - \tau - \mu|^{q-1} \tilde{F}(u) du \\
&= q \int_0^{\mu} |w|^{q-1} F(\mu + \tau - w) dw + q \int_0^{\infty} |w|^{q-1} \tilde{F}(\mu + \tau + w) dw.
\end{aligned}$$

The result follows by inserting inequalities (2.2) and then retracing steps.

To obtain Theorem 2.1, we set $U = R_\alpha$, $U' = \bar{\alpha} R$ so that $\mu = \rho_\alpha$ and $\mu' = \bar{\alpha}\rho$. If $\tau = \bar{\alpha} t \epsilon^2$, then $\delta(U) = \hat{\rho}(R_\alpha, \bar{\alpha} t)$ and $\delta(U') = \bar{\alpha}\hat{\rho}(R; t)$, and (2.1) follows directly from (2.3) once conditions (2.2) have been verified. Define $e_k = y_k^2 - \theta_k^2 - \epsilon^2$ and introduce variables

$$V_1 = \sum \alpha_k e_k, \qquad V_2 = \sum (\bar{\alpha} - \alpha_k) e_k. \tag{2.4}$$

Recall that random variables $Y = (Y_1, \ldots, Y_k)$ are said to be *associated* if $\mathrm{Cov}\,(g(Y), h(Y)) \geq 0$ for all functions g, h monotonically increasing in each argument. (Tong (1980, Ch. 5.2) collects the properties we use.) Then, since ϵ_k are independent and $\alpha_k \geq 0$, V_1 and V_2 are associated and so $U - \mu = V_1$ and $U' - \mu' = V_1 + V_2$. To verify the tail domination conditions (2.2), the following lemma is now convenient.

Lemma 2.3. *If random variables V_1 and V_2 are associated, then for all $s \in \mathbb{R}$,*

$$P(V_1 \leq s) \leq \frac{P(V_1 + V_2 \leq s)}{P(V_2 \leq 0)}, \tag{2.5}$$

$$P(V_1 \geq s) \leq \frac{P(V_1 + V_2 \geq s)}{P(V_2 \geq 0)}. \tag{2.6}$$

Proof. We have

$$\begin{aligned} P(V_1 \leq s) &= P(V_1 \leq s, V_2 > 0) + P(V_1 \leq s, V_2 \leq 0) \\ &\leq P(V_1 \leq s, V_2 > 0) + P(V_1 + V_2 \leq s). \end{aligned}$$

The association implies that $P(V_1 \leq s, V_2 > 0) \leq P(V_1 \leq s) P(V_2 > 0)$. Now (2.5) follows by rearrangement, and the proof of (2.6) is analogous.

It remains, therefore, to exhibit a lower bound for $P(V_2 \geq 0)$ and $P(V_2 \leq 0)$ when V_2 has the form (2.4). To this we now turn.

2.1. Asymmetry of weighted χ^2 variables

Definition. X is a *weighted non-central χ^2 variable* if there exist independent random variables $y_k \sim N(\theta_k, \epsilon^2)$, $k = 1, \ldots, d$ and constants $\alpha_k \geq 0$, $k = 1, \ldots d$ such that

$$X = \sum_{k=1}^{d} \alpha_k y_k^2. \tag{2.7}$$

Proposition 2.4. *Let X be a weighted non-central χ^2 variable with mean μ. There exists an absolute constant $\gamma > 0$, not depending on $d, \alpha, \theta, \epsilon^2$, such that*

$$\min\{P(X \leq \mu), P(X \geq \mu)\} \geq \gamma^{-1}.$$

The proof is obtained in two steps. First, existence of an upper bound on the kurtosis of a random variable implies that it cannot be too asymmetric about its mean, (Lemma 2.5 and Proposition 2.6). Second, such a kurtosis bound is relatively easy to compute for weighted chi-square variables (Lemma 2.7). (We remark that this method does not, alas, yield anything like a sharp bound for γ^{-1}, which we conjecture is given by $P(\chi^2_{(1)} \le 1) = 2[\Phi(1) - 1/2] \doteq 0.36$.)

Lemma 2.5. *Let W be a positive random variable with $EW^2 \ge 1$. If $EW \le \frac{1}{2}$, then*

$$EW^4 \ge \frac{1}{3(EW)^2}. \tag{2.8}$$

Proof. It is easily checked by rescaling that it suffices to take $EW^2 = 1$. Suppose then that $EW = \epsilon$. The set $\mathcal{F}$ of probability measures F supported on $[0, \infty)$ and satisfying $\int x dF(x) = \epsilon$, $\int x^2 dF(x) = 1$ is convex. Since $EW^4 = \int x^4 dF(x)$ is linear in F, it is enough to establish a lower bound for the extreme points of $\mathcal{F}$, which have the form

$$F = p\nu_a + q\nu_b \qquad 0 \le a \le \epsilon \le b, \qquad p + q = 1, \tag{2.9}$$

where ν_a denotes a unit point mass at x. Such F satisfy

$$ap + bq = \epsilon \qquad a^2p + b^2q = 1. \tag{2.10}$$

From (2.9) we have $a^2p \le \epsilon^2$ and from (2.10) $b^2q \ge 1 - \epsilon^2$. Since $bq \le \epsilon$, it follows that $b \ge \epsilon^{-1}(1 - \epsilon^2)$ and hence

$$EW^4 \ge b^4q \ge \epsilon^{-2}(1 - \epsilon^2)^3.$$

The latter quantity certainly exceeds $1/(3\epsilon^2)$ if $\epsilon \le 1/2$.

Let X be a random variable with mean μ. Kurtosis is often measured by the ratio

$$\beta_2(X) = \frac{E(X - \mu)^4}{[E(X - \mu)^2]^2}.$$

Proposition 2.6. *There exists an absolute constant $C \ge 1/12$ such that*

$$\min\{P(X \le \mu), P(X \ge \mu)\} \ge \frac{C}{\beta_2(X)}. \tag{2.11}$$

Proof. Without loss of generality, after centering, scaling and sign change, we may assume that $\mu = EX = 0$, $EX^2 = 1$, and confine attention to $p = P(X \le 0)$. We aim to show

$$EX^4 \ge Cp^{-1}. \tag{2.12}$$

Let X_+ (and X_-) be random variables having the distribution of X conditioned to be strictly positive (and negative, respectively). Set $p = P(X \leq 0)$ and $q = P(X \geq 0)$; then

$$1 = EX^2 = pEX_-^2 + qEX_+^2.$$

Suppose first $pEX_-^2 \geq 1/2$. Then, by Hölder's inequality

$$EX^4 \geq pEX_-^4 \geq p(EX_-^2)^2 \geq (4p)^{-1}.$$

Now assume the contrary, that $pEX_-^2 < 1/2$. Then $qEX_+^2 \geq 1/2$, and we will apply Lemma 2.5 to $W = \sqrt{2q}X_+$. First note that $EX = 0$ implies

$$qEX_+ = p|EX_-| \leq p(EX_-^2)^{1/2} \leq \sqrt{p/2},$$

so that $EW \leq (p/q)^{1/2}$.

If $p > 1/5$, then (2.12) holds trivially with $C = 1/5$. On the other hand $p \leq 1/5$ implies $EW \leq 1/2$ and (2.8) then entails $4q^2EX_+^4 \geq q/3p$. Hence

$$EX^4 \geq qEX_+^4 \geq (12p)^{-1}.$$

Remark. Although harder to exploit for our purposes, an easier inequality like (2.11) is

$$P(X \geq \mu) \geq \frac{1}{4}\frac{[E|X-\mu|]^2}{E(X-\mu)^2},$$

which may be simply proved starting from the Cauchy-Schwartz bound

$$[E(X-\mu)I\{X \geq \mu\}]^2 \leq E(X-\mu)^2 P(X \geq \mu).$$

Lemma 2.7. *Let X be a weighted non-central χ^2 variable and denote its mean by μ. Then*

$$\beta_2(X) \leq 15, \tag{2.13}$$

with the equality attained, for example, by $X = \chi^2_{(1)}$.

Proof. Use the representation (2.7) to write $X - \mu = \sum_1^d Y_k$, with $Y_k = \alpha_k(y_k^2 - Ey_k^2)$ being independent with mean zero. Thus

$$E(\sum Y_k)^2 = \sum_k \operatorname{Var} Y_k, \qquad E(\sum Y_k)^4 = \sum_k \operatorname{Var} Y_k^2 + [\sum_k \operatorname{Var} Y_k]^2.$$

Consequently

$$\beta_2(X) \leq \frac{\sum_k \operatorname{Var} Y_k^2}{[\sum_k \operatorname{Var} Y_k]^2} + 1.$$

Thus (2.13) will follow from the bound $\operatorname{Var} W^2(\mu) \leq 14[\operatorname{Var} W(\mu)]^2$ where $W(\mu) = (Z+\mu)^2 - (1+\mu^2)$ and $Z \sim N(0,1)$. Calculations using $EZ^{2k} = \pi^{-1/2}2^k\Gamma(k+1/2)$ show that

$$\frac{\operatorname{Var} W^2(\mu)}{[\operatorname{Var} W(\mu)]^2} = 2\frac{(4\mu^2+14)^2 - 168}{(4\mu^2+2)^2},$$

which is decreasing in μ^2, and yields the desired bound at $\mu = 0$.

3. MSE Bounds for Chi-square Distributions and Consequences

We recall some results from [J] in the special case $\alpha_k = 1$ and $\epsilon^2 = 1$. For this we adopt the notation $W_d \sim \chi^2_d(\xi)$ and set

$$\sigma^2(\xi) = \operatorname{Var} W_d = 2d + 4\xi. \tag{3.1}$$

Let $\tilde{F}_d(w) = P(\chi^2_d \geq w)$ denote the survivor function of *central* χ^2. A threshold estimator with threshold t will be denoted $\hat{\xi}_t(w) = (w - d - t)_+$ and we write $r(\xi, t; d) = E(\hat{\xi}_t(W_d) - \xi)^2$ for its mean squared error. Define auxiliary constants

$$\eta_1 = 2\tilde{F}_{d+2}(d+t), \qquad \eta_2 = \eta_1 + t/d,$$

which will be small for d large and $t = o(d)$ large.

Proposition 3.1.([J]) *For all* $d \geq 1, \xi \geq 0$ *and* $t \geq 2$,

$$r(\xi,t) \leq \sigma^2(\xi) + t^2, \tag{3.2}$$

$$r(\xi,t) \leq r(0,t) + \eta_1 + (1+\eta_2)\xi^2, \tag{3.3}$$

$$r(0,t) \leq 8\Big(\frac{t+d}{t+2}\Big)^2\tilde{F}_d(d+t), \tag{3.4}$$

$$\frac{\partial^2 r}{\partial \xi^2}(\xi,t) \leq 2(1+t/d). \tag{3.5}$$

Bound (3.2) has a variance character and is useful for large ξ. Bound (3.3) has a 'bias' flavor and is effective for small ξ. Bound (3.4) shows that the risk at 0 is small for suitably large t, while bound (3.5) is a global curvature estimate. While the inequalities hold for all degrees of freedom $d \geq 1$, in the large d limit they transform to the appropriate bounds for a Gaussian shift problem - see [J] for details.

Bounds (3.2)-(3.5) are intended to be reasonably sharp and, when combined with the comparison bound (2.3), are the basic tools for deriving MSE bounds for weighted chi-square situations. These bounds may be less sharp, but more convenient, for particular applications. We give an example here, to be used in Section 3.

Corollary 3.2. *Let* $t_{\beta d} = \sqrt{2d}\sqrt{2\beta \log d}$ *for* $\beta \geq 1$. *There exists a constant* $c = c_\beta$ *such that for* $d \geq 16$,

$$r(\xi, t_{\beta d}) \leq cd^{1-\beta} + \min\{2\xi^2, \sigma^2(\xi) + t_{\beta d}^2\}. \tag{3.6}$$

Proof. Bound (3.2) yields immediately $r(\xi, t_{\beta d}) \leq \sigma^2(\xi) + t_{\beta d}^2$, whereas (3.3) and (3.4) together yield, for $t = t_{\beta d}$

$$r(\xi, t_{\beta d}) \leq 8\Big(\frac{t+d}{t+2}\Big)^2 \tilde{F}_d(d+t) + 2\tilde{F}_{d+2}(d+t) + (1+\eta_2)\xi^2.$$

As a corollary of a more refined bound, Lemma 6.1 of [J] shows that for $d \geq 16$ and $s \leq d^{1/6}$, we have $\tilde{F}_d(d + s\sigma_d) \leq s^{-1}e^{-s^2/2}$. From this follows, after some simple algebra, the bound $r(\xi, t_{\beta d}) \leq cd^{1-\beta} + 2\xi^2$ for $d \geq 16$ and suitable $c = c_\beta$.

Using the comparison Theorem 2.1, this goes over the the weighted case.

Corollary 3.3. *Suppose that* $y_k \sim N(\theta_k, \epsilon^2), k = 1, \ldots, d$ *are independent. Let* $t_{\beta d} = \sqrt{2d}\sqrt{2\beta \log d}$ *for* $\beta \geq 1$. *There exists* $c = c_\beta$ *such that for* $d \geq 16$,

$$E[\hat{\rho}_\alpha(R_\alpha; \bar{\alpha} t_{\beta d}) - \rho_\alpha]^2 \leq \gamma\bar{\alpha}^2[cd^{1-\beta}\epsilon^4 + \min\{2\rho_\alpha^2, \sigma^2(\rho_\alpha) + t_{\beta d}^2\epsilon^4\}]. \tag{3.7}$$

Proof. Direct application of Theorem 2.1 and then (3.6) bounds the left side of (3.7) by

$$\gamma\bar{\alpha}^2 E[\hat{\rho}(R; t) - \rho]^2 = \gamma\bar{\alpha}^2[cd^{1-\beta}\epsilon^4 + \min\{2\rho^2, \sigma^2(\rho) + t_{\beta d}^2\epsilon^4\}].$$

Since $\rho = \sum \theta_k^2 \leq \sum \alpha_k \theta_k^2 = \rho_\alpha$, (3.7) follows.

4. Illustration: estimation of $\int (D^l f)^2$

Assume observations from the white noise model $Y_t = \int_0^t f(s)ds + \epsilon W_t$, $0 \leq t \leq 1$ with $\{W_t\}$ standard Brownian motion, ϵ known, and $f \in L_{per}^2[0,1]$ periodic and unknown. We seek to estimate $Qf = \int_0^1 (D^l f)^2$, where $l \in \mathbb{N}$. In various contexts of regression, density estimation and the present white noise model, this question has received considerable recent attention, in particular because of applications to bandwidth selection. A selection of recent references, in addition to those mentioned below, includes Hall and Marron (1987), Donoho and Nussbaum (1990), Hall and Johnstone (1992), Birgé and Massart (1995), Laurent (1996) and Cheng (1997).

Bickel and Ritov (1988), and later independently Fan (1991), found an interesting dichotomy in this problem. To recall this, assume that f has σ mean-square derivatives satisfying the Sobolev condition

$$\int_0^1 f^2 + (D^\sigma f)^2 \leq B^2. \tag{4.1}$$

In smooth cases, $\sigma > 2l + 1/4$, efficient estimation of Qf is possible at rate ϵ^2 as $\epsilon \to 0$. But, in less smooth settings, $l < \sigma \leq 2l + 1/4$, the minimax rate of convergence is just ϵ^{2r}, where

$$r = \frac{8(\sigma - l)}{4\sigma + 1}. \tag{4.2}$$

When σ is unknown, Efromovich and Low (1996) showed, in a density estimation setting, that adaptation to the optimal rate (4.2) was not possible for $Qf = \int (D^l f)^2$: an extra logarithmic term is necessary in the non-parametric zone. In what follows, we construct an estimator that adapts to varying levels of smoothness at the best possible rate allowed by the result of Efromovich and Low (1996) when σ is unknown. Begin with the Fourier basis,

$$\phi_0(t) = 1, \quad \phi_{2k-1}(t) = \sqrt{2} \sin 2\pi k t, \quad \phi_{2k}(t) = \sqrt{2} \cos 2\pi k t, \qquad k \geq 1,$$

which in the white noise model, as is well known, has a sequence representation

$$\begin{aligned} y_k &= \theta_k + \epsilon z_k \qquad z_k \overset{i.i.d}{\sim} N(0,1), \quad k \in \mathbb{N} \\ \theta_k &= \int_0^1 \phi_k f. \end{aligned} \tag{4.3}$$

The functional Qf takes a weighted quadratic form

$$\int (D^l f)^2 = \sum_0^\infty \lambda_k^l \theta_k^2 \qquad \lambda_0 = 0, \lambda_{2k-1} = \lambda_{2k} = (2\pi k)^2, \qquad k \geq 1. \tag{4.4}$$

To prepare to apply the weighted chi-square risk inequality, divide the sum into dyadic blocks over which the polynomially growing weights λ_k^l are balanced. Specifically, blocks B_j with associated weights κ_j are defined by setting $B_0 = \{0, 1, 2\}$ and

$$\begin{aligned} B_j &= \{d_j + 1, \ldots, 2d_j\}, \qquad d_j = 2^j, \ j \geq 1, \\ \kappa_j &= (\pi d_j)^{2l} = \pi^{2l} 2^{2jl}, \qquad j \geq 0. \end{aligned} \tag{4.5}$$

Given $k \in \mathbb{N}$, let $j(k)$ be the index of the block B_j in which k lies, and set

$$\alpha_k = \frac{\lambda_k^l}{\kappa_{j(k)}} \in [1, \bar{\alpha}], \qquad \bar{\alpha} = 2^{2l}. \tag{4.6}$$

In terms of the blocks B_j, Qf may be rewritten as

$$\int (D^l f)^2 = \sum_{j \geq 0} \kappa_j \rho_j, \qquad \rho_j = \sum_{k \in B_j} \alpha_k \theta_k^2, \tag{4.7}$$

and, of course, the aim is to apply the results of the previous section to estimation of ρ_j.

Smoothness condition (4.1) implies (proof in appendix) the uniform decay condition

$$\rho_j = \sum_{B_j} \alpha_k \theta_k^2 \ \leq \ C^2 2^{-2\sigma j} =: \bar{\rho}_j, \qquad\qquad j \geq 1, \tag{4.8}$$

if we set $C = \pi^{-\sigma} B$. Denote by $\Theta^\sigma(C)$ the collection of sequences θ satisfying this decay condition – it is in fact a norm ball in the periodic Besov space $B^\sigma_{2,\infty}$ on $[0,1]$ which contains the periodic Sobolev space $B^\sigma_{2,2}$ whose norm balls appear in (4.1). We use $\Theta^\sigma(C)$ below simply for convenience. See also Remark 1 below.

In smoother cases the low frequencies are most important whereas, in rough settings, higher frequencies are critical. The estimate therefore combines unbiased estimation at lower frequencies (where efficiency is the goal)

$$\hat{Q}_e = \sum_{k=1}^{k_0} \lambda_k^l (y_k^2 - \epsilon^2), \qquad k_0 = 2^{j_0}, \quad 2^{-(4l+1)j_0} = \epsilon^2 \sqrt{\log_2 \epsilon^{-2}}, \tag{4.9}$$

with thresholding at higher frequencies

$$\hat{Q}_t = \sum_{j=j_0+1}^{j_1} \kappa_j \hat{\rho}_j, \qquad\qquad 2^{-j_1} = \epsilon^4. \tag{4.10}$$

Here $\hat{\rho}_j$ is a threshold estimate of type (1.1) studied in previous sections: starting from the unbiased estimate $R_j = \sum_{k \in B_j} \alpha_k (y_k^2 - \epsilon^2)$ of ρ_j, and recalling $d_j = 2^j$,

$$\hat{\rho}_j = (R_j - \bar{\alpha} t_j \epsilon^2)_+, \qquad\qquad t_j = \sqrt{2d_j}\sqrt{2(4l+1)\log d_j}. \tag{4.11}$$

The thresholds t_j correspond to setting $\beta = 4l + 1$ in Corollary 3.3. These higher, and hence more conservative, thresholds are used because (4.7) shows that higher frequency blocks B_j receive larger weights κ_j, and the higher thresholds combat the "noise amplification" induced in the estimate (4.10). Abramovich and Silverman (1998) first used this device.

Of course j_0 and j_1 as just defined need not be integer valued. We agree that a sum $\sum_{j=a}^{b}$ is taken to run over $j = \lfloor a \rfloor = \text{floor}(a)$ to $j = \lceil b \rceil = \text{ceiling}(b)$. Below, c denotes a constant depending at most on l and σ, and not necessarily the same at each appearance.

Theorem 4.1. *Let $\hat{Q} = \hat{Q}_e + \hat{Q}_t$. Then the following hold.*
(i) *If $\sigma > 2l + 1/4$, let $Rf = \int (D^{2l} f)^2$ to find*

$$\sup_{f \in \Theta^\sigma(C)} |E(\hat{Q} - Qf)^2 - 4\epsilon^2 Rf| = o(\epsilon^2). \tag{4.12}$$

(ii) *If* $l < \sigma \le 2l + 1/4$,

$$\sup_{f \in \Theta^\sigma(C)} E(\hat{Q} - Qf)^2 \le cC^{2(2-r)}(\epsilon^2 \sqrt{\log(C\epsilon^{-1})}^r (1 + o(1)). \tag{4.13}$$

Thus, in the "parametric zone", $\sigma > 2l + 1/4$, $\hat{Q}$ is an efficient estimator of Qf – indeed $\hat{Q}_e$ is essentially the efficient estimator of Ibragimov and Has'minskii (1977) (see also Hall and Johnstone (1992, Proposition 1)) and $\hat{Q}_t$ is negligible. In the "non-parametric" zone, $l < \sigma \le 2l + 1/4$, the rate bound in (4.13) is the best allowed by the lower bounds of Efromovich and Low (1996), and comes from $\hat{Q}_t$, with $\hat{Q}_e$ being negligible here.

Proof. In the series expression (4.4), decompose $Qf = Q_e f + Q_t f + Q_r f$ where the ranges of summation match those of $\hat{Q}_e$ and $\hat{Q}_t$ in (4.9) and (4.10). Using the triangle inequality for $\|\delta\| = \sqrt{E\delta^2}$,

$$\sqrt{E(\hat{Q} - Qf)^2} \le \sqrt{E(\hat{Q}_e - Q_e f)^2} + \sqrt{E(\hat{Q}_t - Q_t f)^2} + Q_r f. \tag{4.14}$$

1°. *Tail Bound.* This is negligible in all cases: using (4.6), (4.8) and (4.10),

$$Q_r f \le \sum_{j_1}^{\infty} \kappa_j \rho_j \le cC^2 \sum_{j_1}^{\infty} 2^{2jl - 2j\sigma} \le cC^2 2^{-2(\sigma - l)j_1} = cC^2 \epsilon^{8(\sigma - l)} = o(\epsilon^r). \tag{4.15}$$

In particular, in the efficient case when $\sigma > 2l + 1/4$, $8(\sigma - l) > 8l + 2 \ge 2$, so that $Q_r f = o(\epsilon^2)$.

2°. *Efficient Term.* Since $\hat{Q}_e$ is unbiased, we have, using (4.9) and (3.1),

$$E(\hat{Q}_e - Q_e f)^2 = \operatorname{Var} \hat{Q}_e = 4\epsilon^2 \sum_1^{k_0} \lambda_k^{2l} \theta_k^2 + 2\epsilon^4 \sum_1^{k_0} \lambda_k^{2l}. \tag{4.16}$$

The second term is always negligible: from (4.4) and (4.9),

$$\epsilon^4 \sum_1^{k_0} \lambda_k^{2l} \le c\epsilon^4 \sum_1^{k_0} k^{4l} \le c\epsilon^4 k_0^{4l+1} = \frac{c\epsilon^2}{\sqrt{\log \epsilon^{-2}}} = o(\epsilon^2).$$

In the parametric zone, $\sigma > 2l + 1/4$, using (4.6), (4.6) and (4.8), uniformly on $\Theta^\sigma(C)$,

$$Rf - \sum_1^{k_0} \lambda_k^{2l} \theta_k^2 = \sum_{k_0+1}^{\infty} \lambda_k^{2l} \theta_k^2 \le c\bar{\alpha} \sum_{k_0+1}^{\infty} \kappa_j^2 \rho_j \le cC^2 \sum_{j_0+1}^{\infty} 2^{4lj - 2\sigma j} \le cC^2 2^{-j_0/2} = o(1). \tag{4.17}$$

Combining the three previous displays, in the parametric case

$$\sup_{\Theta^\sigma(C)} |E(\hat{Q}_e - Q_e f)^2 - 4\epsilon^2 Rf| = o(\epsilon^2).$$

To verify that the efficient term is negligible in the non-parametric zone $l < \sigma < 2l + 1/4$, one checks that the first term on the right side of (4.16) is $o(\epsilon^{2r})$. Indeed, as for (4.17),

$$\epsilon^2 \sum_1^{k_0} \lambda_k^{2l} \theta_k^2 \le c\epsilon^2 \sum_1^{j_0} 2^{(4l-2\sigma)j}.$$

If $2l \le \sigma \le 2l + 1/4$, this is obviously $O(\epsilon^2) = o(\epsilon^{2r})$. If $l < \sigma < 2l$, the right side is bounded by $c\epsilon^2 2^{2(2l-\sigma)j_0}$ and is seen to be $o(\epsilon^{2r})$ after substituting (4.9) and a short calculation.

3°. *Thresholding term.* The rest of the proof is concerned with bounding

$$\sqrt{E(\hat{Q}_t - Q_t f)^2} \le \sum_{j_0}^{j_1} \sqrt{\kappa_j^2 E(\hat{\rho}_j - \rho_j)^2}. \tag{4.18}$$

For the threshold estimate $\hat{\rho}_j$ defined at (4.11), the risk inequality Corollary 2.3 yields

$$E(\hat{\rho}_j - \rho_j)^2 \le \gamma\bar{\alpha}^2[c2^{-4lj}\epsilon^4 + 2\min\{\rho_j^2, \sigma^2(\rho_j) + t_j^2\epsilon^4\}],$$

where $\sigma^2(\rho_j) = 2d_j\epsilon^4 + 4\rho_j\epsilon^2$. Since $d_j = 2^j \le t_j^2$ (compare (4.11), we may bound

$$\begin{aligned}\kappa_j^2 E(\hat{\rho}_j - \rho_j)^2 &\le c\epsilon^4\kappa_j^2 2^{-4lj} + c\kappa_j^2 \min\{\bar{\rho}_j^2, \bar{\rho}_j\epsilon^2\} + c\kappa_j^2 \min\{\bar{\rho}_j^2, t_j^2\epsilon^4\} \\ &= T_1(j) + T_2(j) + T_3(j).\end{aligned}$$

We verify, with exceptions as noted, that the maps $j \to T_m(j)$, which with slight abuse of notation will be regarded as functions of real arguments, have unique maxima j_m and have geometric decay away from these maxima, so that the maximum term in (4.18) determines the rate of convergence.

Consider first the main term T_3. Since, for $l < \sigma$, $j \to \kappa_j^2\bar{\rho}_j^2$ is geometrically decreasing, and $j \to \kappa_j^2 t_j^2 \epsilon^4$ is geometrically increasing, the point of maximum j_3 occurs at the solution of $\bar{\rho}_j^2 = t_j^2\epsilon^4$. Now $j_3 = j_+ + O(1)$ as $\epsilon \to 0$ (details are in the appendix of [J]), where

$$2^{-(2\sigma+1/2)j_+} = \frac{\epsilon^2}{C^2}\sqrt{\log\frac{C}{\epsilon}},$$

and

$$T_3(j_+) = c\kappa_{j_+}^2 C^4 2^{-4\sigma j_+} = cC^4 2^{-4(\sigma-l)j_+} = cC^4\Big(\frac{\epsilon^2}{C^2}\sqrt{\log\frac{C}{\epsilon}}\Big)^r.$$

Since $T_3(j_3) \leq cT_3(j_+)$, this yields the upper bound (4.13). (Note that in the parametric zone $\sigma > 2l + \frac{1}{4}$, $r > 1$, and so $T_3(j_3) = o(\epsilon^2)$ is negligible.)

The T_1 term is easily handled: $T_1(j) \leq c\epsilon^4$ and so $\sum_1^{j_1} \sqrt{T_1(j)} \leq j_1\epsilon^2 = o(\epsilon)$. For T_2, first define j_4 as the solution of $\bar{\rho}_j\epsilon^2 = t_j^2\epsilon^4$: since $j \to \bar{\rho}_j\epsilon^2$ is decreasing, it follows that $T_2(j) \leq cT_3(j)$ for all $j \geq j_4$. It remains, then, to consider $j \in [j_0, j_4]$. Calculation shows

$$(1+2\sigma)j_4 = \log_2 C^2\epsilon^{-2} - \log_2\log_2 C\epsilon^{-1} + O(1).$$

If $\sigma \geq 2l$, note that $j_0 = (4l+1)^{-1}(\log_2 \epsilon^{-2} - \frac{1}{2}\log_2\log_2 \epsilon^{-1})$ satisfies $j_0 - j_4 \to \infty$ as $\epsilon \to 0$, and so the range $[j_0, j_4]$ is ultimately empty. Turning to $l < \sigma < 2l$, observe for $j \leq j_4$ that $\epsilon^2 \leq \bar{\rho}_j$ (since the point of equality occurs at $j_5 = (2\sigma)^{-1}\log_2 C^2\epsilon^{-2} > j_4$ for small ϵ). Hence $T_2(j) \leq c\kappa_j^2\bar{\rho}_j\epsilon^2 \leq c\epsilon^2 2^{2(2l-\sigma)j}$ grows exponentially, but nevertheless

$$T_2(j) \leq c\kappa_{j_5}^2\bar{\rho}_{j_5}\epsilon^2 = c\kappa_{j_5}^2\bar{\rho}_{j_5}^2 = cC^2 2^{4(l-\sigma)j_5} = O((\epsilon^2)^{4(\sigma-l)/2\sigma}) = o(\epsilon^{2r}).$$

Thus T_1 and T_2 are both negligible and this completes the analysis of the thresholding term, and hence the proof.

Remarks. 1. It would be possible to extend the dyadic block methods of this section to other Besov spaces $B^\sigma_{p,q}$, and to dispense with the periodicity assumption, by using appropriate wavelet bases. Since differentiation is not *exactly* diagonalized in wavelet bases, some extra technical bookkeeping would be needed.

2. The blocking technique clearly extends to more general quadratic functionals $\sum \lambda_j\theta_j^2$, (cf. e.g., Donoho and Nussbaum (1990), Fan (1991) so long as the coefficients λ_j satisfy the balancing relation (4.6). In particular, the λ_j should grow at most at a polynomial rate.

3. In the same vein, extensions are possible to estimation of certain quadratic functions in inverse problems with random noise. At least formally, consider a model of the form $Y = Kf + \epsilon W$ where $K : H_1 \to H_2$ is a bounded linear operator between Hilbert spaces, and W is an appropriate white noise process. Suppose that K has singular value decomposition $Ku_k = b_kv_k$ in terms of orthonormal bases $\{u_k\}, \{v_k\}$ for H_1, H_2 and singular values $b_k > 0$. Writing $f = \sum_k f_ku_k$, then the observed data takes the form (4.3) with $y_k = \langle Y, v_k\rangle$, $\theta_k = \langle Kf, v_k\rangle = b_kf_k$ and $z_k = \langle W, v_k\rangle$. A quadratic functional Qf of the form $\sum \mu_kf_k^2$ satisfies $Qf = \sum \lambda_k\theta_k^2$ with $\lambda_k = \mu_k/b_k^2$. From the previous remark, the methods of this paper will apply if the coefficients λ_k have at worst polynomial growth.

4. Finally, suppose that f is piecewise continuously differentiable, in the sense that Df has a finite number of jump discontinuities. Estimation of the sum of squares of the jumps of f, $\gamma(f)$ say, has recently been considered, along with

applications to growth models, by Müller and Stadtmüller (1999). A theorem of Wiener (Katznelson (1968, p.42) and Zygmund (1959, p.40)) asserts that $\gamma(f)$ is given by the large N limit of the normalized sum of squares of Fourier-Stieltjes coefficients

$$\frac{1}{2N+1}\sum_{0}^{2N}\Big(\int \phi_k df\Big)^2 = \frac{1}{2N+1}\sum_{0}^{2N}\lambda_k\theta_k^2$$

in the notation of (4.4). Thus, at least in principle, an estimator of $\gamma(f)$ could be built and studied using the methods of this section.

5. The thresholds in (4.11) are chosen high, with logarithmic terms, in order to achieve the adaptivity claimed in (4.13) as the smoothness σ varies. One could achieve smaller mean squared error in many settings with smaller (and even data determined) thresholds, but adaptivity would necessarily be sacrificed.

Acknowledgements

The author is grateful for the hospitality of the Australian National University, where much of this research was carried out, and for the helpful comments of the referee and associate editor. This research was also supported in part by NSF DMS 9505151 and DMS 0072661.

Appendix

Proof of (4.8). In analogy with (4.4), we have $\int(D^\sigma f)^2 = \sum_{k=0}^{\infty}\lambda_k^\sigma\theta_k^2$, (this is often taken as a definition of the left side for $\sigma \notin \mathbb{N}$). Condition (4.1) is then equivalent to

$$\sum_k (1+\lambda_k^\sigma)\theta_k^2 \le B^2.$$

Since $\lambda_k^\sigma = \alpha_k^{\sigma/l}(\pi 2^j)^{2\sigma}$, with $\sigma/l \ge 1$ and $\alpha_k \ge 1$, we recover (4.8):

$$B^2 \ge \sum_{k\in B_j}\lambda_k^\sigma\theta_k^2 \ge \pi^{2\sigma}2^{2j\sigma}\sum_{B_j}\alpha_k\theta_k^2.$$

References

Abramovich, F. and Silverman, B. W. (1998). Wavelet decomposition approaches to statistical inverse problems. *Biometrika* **85**,115-129.

Bickel, P. and Ritov, Y. (1988). Estimating integrated squared density derivatives: sharp best order of convergence estimates. *Sankhyā Ser. A* **50**, 381-393.

Birgé, L. and Massart, P. (1995). Estimation of integral functionals of a density. *Ann. Statist.* **23**, 11-29.

Cheng, M. Y. (1997). Boundary aware estimators of integrated density derivative products. *J. Roy. Statist. Soc. Ser. B* **59**, 191-203.

Donoho, D. L. and Nussbaum, M. (1990). Minimax quadratic estimation of a quadratic functional. *J. Complexity* **6**, 290-323.

Efromovich, S. and Low, M. (1996). On Bickel and Ritov's conjecture about adaptive estimation of the integral of the square of density derivative. *Ann. Statist.* **24**, 682-686.

Fan, J. (1991). On estimation of quadratic functionals. *Ann. Statist.* **19**, 1273-1294.

Hall, P. G. and Marron, J. S. (1987). Estimation of integrated squared density derivatives. *Statist. Probab. Lett.* **6**, 109-115.

Hall, P. and Johnstone, I. (1992). Empirical functionals and efficient smoothing parameter selection, with discussion. *J. Roy. Statist. Soc. Ser. B* **54**, 475-530.

Ibragimov, I. A. and Has'minskii, R. Z. (1977). A problem of statistical estimation in Gaussian white noise. *Soviet Mathematics Doklady* **18**, 1351-1354.

Johnson, N. L. and Kotz, S. (1970). *Distributions in Statistics: Continuous Univariate Distributions - 2.* Wiley, New York.

Johnstone, I. M. (2001). *Chi Square Oracle Inequalities.* In *State of the Art in Probability and Statistics*, Festchrift for Willem R. van Zwet, M. de Gunst, C. Klaassen and A. van der Waart, editors. *IMS Lecture Notes - Monographs* **36**, 399-418.

Katznelson, Y. (1968). *An Introduction to Harmonic Analysis.* Dover.

Laurent, B. (1996). Efficient estimation of integral functionals of a density. *Ann. Statist.* **24**, 659-681.

Müller, H.-G. and Stadtmüller, U. (1999). Discontinuous versus smooth regression. *Ann. Statist.* **27**, 299-337.

Tong, Y. L. 1980 . *Probability Inequalities in Multivariate Distributions.* Academic Press, New York.

Zygmund, A. (1959). *Trigonometric Series, Volume I.* Cambridge University Press, Cambridge.

Department of Statistics, Sequoia Hall, Stanford University, Stanford, CA 94305, U.S.A.

E-mail: imj@stat.stanford.edu

(Received February 2000; accepted February 2001)