CHI-SQUARE ORACLE INEQUALITIES

IAIN M. JOHNSTONE[1]

Stanford University

We study soft threshold estimates of the non-centrality parameter ξ of a non-central $\chi^2_d(\xi)$ distribution, of interest, for example, in estimation of the squared length of the mean of a Gaussian vector. Mean squared error and oracle bounds, both upper and lower, are derived for *all* degrees of freedom d. These bounds are remarkably similar to those in the limiting Gaussian shift case. In nonparametric estimation of $\int f^2$, a dyadic block implementation of these ideas leads to an alternate proof of the optimal adaptivity result of Efromovich and Low.

AMS subject classifications: 62E17, 62F11, 62G07, 62G05.
Keywords and phrases: Quadratic Functionals, Adaptive Estimation, Gaussian sequence model, Efficient estimation, non-central chi-square.

1 Introduction

The aim of this paper is to develop thresholding tools for estimation of certain quadratic functionals. We begin in a finite dimensional setting, with the estimation of the squared length of the mean of a Gaussian vector with spherical covariance. The transition from linear to quadratic functionals of the data entails a shift from Gaussian to (non-central) chi-squared distributions $\chi^2_d(\xi)$ and it is the non-centrality parameter ξ that we now seek to estimate. It turns out that (soft) threshold estimators of the noncentrality parameter have mean squared error properties which, after appropriate scaling, very closely match those of the Gaussian shift model. This might be expected for large d, but this is not solely an asymptotic phenomenon – the detailed structure of the chi-squared distribution family allows relatively sharp bounds to be established for the full range of degrees of freedom d.

We develop oracle inequalities which show that thresholding of the natural unbiased estimator of ξ at $\sqrt{2\log d}$ standard deviations (according to *central* χ^2_d) leads to an estimator of the non-centrality parameter that is within a multiplicative factor $2\log d + \epsilon_d$ of an 'ideal' estimator that can use knowledge of ξ to choose between an unbiased rule or simply estimating zero. These results are outlined in Section 2.

Section 3 shows that the multiplicative $2\log d$ penalty is sharp for large degrees of freedom d, essentially by reduction to a limiting Gaussian shift problem.

[1]This research was supported by NSF DMS 9505151 and ANU.

Section 4 illustrates thresholding in a well-studied nonparametric setting, namely estimation of $\int f^2$, which figures in the asymptotic properties (variance, efficiency) of rank based tests and estimates. We apply the oracle inequalities in the now classical model in which a signal f is observed in Gaussian white noise of scale ϵ. When this model is expressed in a Haar wavelet basis, the sum of squares of the empirical coefficients at a resolution level j has a $\epsilon^2\chi^2_{2^j}(\rho_j/\epsilon^2)$ distribution with parameter ρ_j equal to the sum of squares of the corresponding theoretical coefficients. Thus $\int f^2 = \sum \rho_j$ and this leads to use of the oracle inequalities on each separate level j.

Section 5 contains some remarks on the extension of our thresholding results to weighted combinations of chi-squared variates. In addition to proof details, the final section collects some useful identities for central and non-central χ^2, as well as a moderate deviations bound for central χ^2, Lemma 6.1, in the style of the Mill's ratio bound for Gaussian variates.

2 Estimating the norm of a Gaussian Vector

Suppose we observe $y = (y_i) \in \mathbb{R}^d$, where $y \sim N_d(\theta, \epsilon^2 I)$. We wish to estimate $\rho = \|\theta\|_2^2$. A natural unbiased estimate is $U = |y|^2 - d\epsilon^2 = \sum(y_i^2 - \epsilon^2)$. We propose to study the shrunken estimate

$$\hat{\rho}_t = \hat{\rho}(U;t) = (U - t\epsilon^2)_+. \tag{1}$$

This estimate is always non-negative, and like similar shrunken estimators we have studied elsewhere, enjoys risk benefits over U when ρ is zero or near zero. We will be particularly interested in $t = t_d = \sigma_d\sqrt{2\log d}$, where $\sigma_d = \sqrt{2d}$ is the variance of χ^2_d, the distribution of $|y|^2/\epsilon^2$ when $\theta = 0$. [The *positive part* estimator, corresponding to $t = 0$, has already been studied, for example by Saxena and Alam (1982); Chow (1987).]

The estimator $\hat{\rho}_t$ may be motivated as follows. Let

$$\sigma^2(\rho) = \mathrm{Var}(U) = 2\epsilon^4 d + 4\epsilon^2\rho. \tag{2}$$

An "ideal" but non-measurable estimate of ρ would estimate by 0 if $\rho \le \sigma(\rho)$ and by U if $\rho > \sigma(\rho)$. This rule improves on U when the parameter ρ is so small that the bias incurred by estimating 0 is less than the variance incurred by using estimator U. Hence, this ideal strategy would have risk $\min\{\rho^2, \sigma^2(\rho)\}$.

Of course, no *statistic* can be found which achieves this ideal, because the data cannot tell us whether $\rho \le \sigma(\rho)$ for certain. However, we show that $\hat{\rho}_t$ comes as close to this ideal as can be hoped for.

To formulate the main results, it is convenient to rescale to noise level $\epsilon = 1$, and to change notation to avoid confusion.

Thus, let $W_d \sim \chi^2_d(\xi)$ – we seek to estimate ξ using a threshold estimator

$$\hat{\xi}_t(w) = (w - d - t)_+.$$

Write $\sigma^2(\xi) = 2d + 4\xi$ for the variance of W_d, and $\tilde{F}_d(w) = P(\chi^2_d \geq w)$ for the survivor function of the corresponding *central* χ^2 distribution. Introduce two auxiliary constants (which are small for d large and $t = o(d)$ large):

$$\eta_1 = 2\tilde{F}_{d+2}(d+t), \qquad \eta_2 = \eta_1 + t/d, \qquad t \geq 2. \tag{3}$$

Let D_ξ and D^2_ξ denote partial derivatives with respect to ξ.

Theorem 2.1 *With these definitions, the mean squared error* $r(\xi, t) = E(\hat{\xi}_t(W_d) - \xi)^2$ *satisfies, for all* $d \geq 1, \xi \geq 0$ *and* $t \geq 2$,

$$r(\xi, t) \leq \sigma^2(\xi) + t^2, \tag{4}$$

$$r(\xi, t) \leq r(0, t) + \eta_1 + (1 + \eta_2)\xi^2, \tag{5}$$

$$r(0, t) \leq 8\Big(\frac{t+d}{t+2}\Big)^2 \tilde{F}_d(d+t), \tag{6}$$

$$D^2_\xi r(\xi, t) \leq 2(1 + t/d). \tag{7}$$

Bound (4) has a "variance" character and is useful for large ξ, while (5) has a "bias" flavour and is effective for small ξ. Bound (6) shows that the larger the threshold t, the smaller is the risk at 0, while (7) is a global curvature estimate.

Remark. These inequalities are valid for all degrees of freedom $d \geq 1$. However, since W_d is asymptotically Gaussian for d large, it is also informative to rescale these by defining

$$X_d = \frac{W_d - d}{\sqrt{2d}}, \quad \theta = \frac{\xi}{\sqrt{2d}}, \quad \lambda = \frac{t}{\sqrt{2d}}, \quad \hat{\theta}_\lambda(x) = (x - \lambda)_+,$$

and

$$\rho(\theta, \lambda) = E(\hat{\theta}_\lambda(X_d) - \theta)^2 = r(\xi, t)/2d.$$

Thus X is approximately distributed as $N(\theta, 1 + \theta\sqrt{8/d})$ for large d. If we also introduce $\tilde{\Phi}_d(z) = P\{X_d > z\}$, $\epsilon_1 = \eta_1/2d$ and $\epsilon_2 = \eta_2$ then inequalities (4) - (7) become

$$\begin{aligned}
\rho(\theta, \lambda) &\leq 1 + \lambda^2 + \theta\sqrt{8/d}, \\
\rho(\theta, \lambda) &\leq \rho(0, \lambda) + \epsilon_1 + (1 + \epsilon_2)\theta^2, \\
\rho(0, \lambda) &\leq 2\lambda^{-2}(1 + \lambda\sqrt{2/d})^2 \tilde{\Phi}_d(\lambda), \\
D^2_\theta \rho(\theta, \lambda) &\leq 2 + \lambda\sqrt{8/d}.
\end{aligned}$$

Aside from terms that are $O(d^{-1/2})$ or smaller, these inequalities are essentially identical to those for the Gaussian shift problem in which soft thresholding at λ is applied to $X \sim N(\theta, 1)$ (compare Donoho and Johnstone (1994, Appendix 2)).

Proof Missing details and basic facts about (non-) central χ^2 are collected in the Appendix. First define $t_1 = t+d$ and write $f_{\xi,d}$ for the density function of $\chi^2_d(\xi)$. The "variance" bound (4) is easy: since $\xi \geq 0$,

$$r(\xi, t) = E[(W_d - t_1)_+ - \xi]^2 \leq E[W_d - t_1 - \xi]^2 = \text{Var } W_d + t^2.$$

Partial integration and formula (49) lead to useful expressions for the risk function and its derivatives (details in Appendix): Let $p_\lambda(x) = e^{-\lambda}\lambda^x/\Gamma(x+1)$ denote the Poisson p.d.f. with mean λ: $p_\lambda(x)$ is also well defined for half-integer x.

$$r(\xi, t) = \xi^2 \int_0^{t_1} f_{\xi,d} + \int_{t_1}^{\infty} (w - t_1 - \xi)^2 f_{\xi,d}(w)dw, \tag{8}$$

$$r(0, t) = (t^2 + 2d)\tilde{F}_d(d + t) - d(t - 2)p_{t_1/2}(d/2), \tag{9}$$

$$D_\xi r(\xi, t) = 2\xi \int_0^{t_1} f_{\xi,d+2} + 4 \int_{t_1}^{\infty} f_{\xi,d+2}, \tag{10}$$

$$D^2_\xi r(\xi, t) = 2 \int_0^{t_1} f_{\xi,d+2} + (4 - 2\xi) f_{\xi,d+4}(t_1). \tag{11}$$

Some fairly crude bounds in (11) (see Appendix) then yield (7).

For (5), substitute (10) and (7) into the Taylor expansion

$$\begin{aligned} r(\xi, t) &= r(0, t) + \xi D_\xi r(0, t) + \int_0^\xi ds \int_0^\xi du\, D^2_\xi r(u, t) \\ &\leq r(0, t) + 2\xi\eta_{1t} + (1 + t/d)\xi^2. \end{aligned}$$

Replacing $2\xi \leq 1 + \xi^2$ leads to (5). Finally, formula (6) is derived from (8) in the Appendix. ∎

2.1 Numerical Illustration

Formulas (9) and (10) enable a straightforward numerical evalution of the risk of thresholding. Figure 1 compares the mean squared error (MSE) of thresholding at $t = 0, 1$ or $\sqrt{2 \log d}$ standard deviations σ_d for $d = 8$ and 16. [Numerical integration in $r(\xi, t) = r(0, t) + \int_0^\xi D_\xi r(u, t)du$ was performed using the routine `integrate` in `S-PLUS`.] The positive part rule (`REFER TO THIS`) ($t = 0$), namely $(w - d)_+$ yields up to 50% MSE savings at $\xi = 0$. However, to obtain smaller risks at 0 necessarily entails larger MSE at

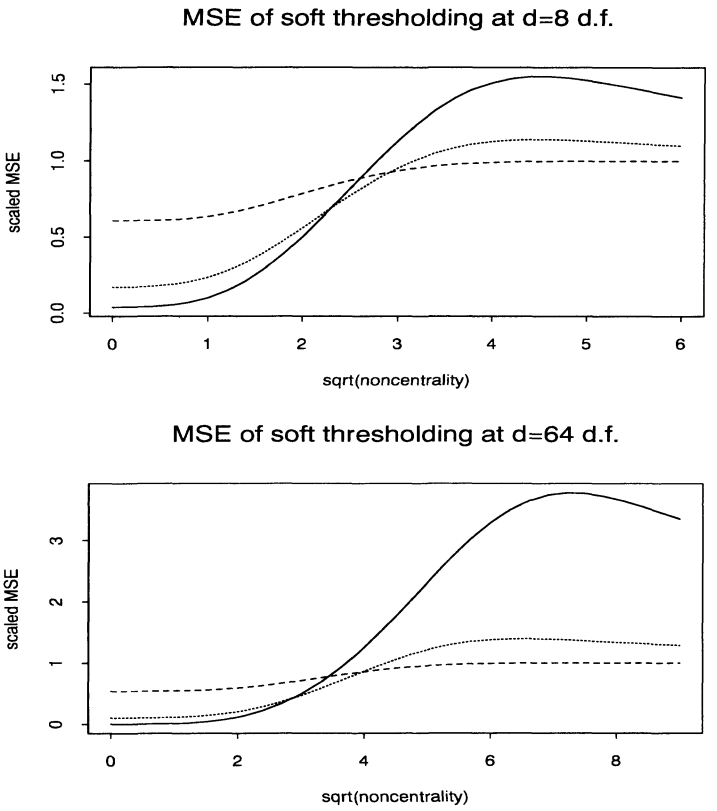

Figure 1. Mean squared error (MSE) of thresholding rules, calculated from formulas (9) and (10) . Horizontal axis is root noncentrality $\sqrt{\xi}$, vertical axis is scaled MSE $r(\xi,t)/\sigma^2(\xi)$ for thresholds $t = 0$ (dashed line), $t = \sigma_d$ (dotted line) and $t = \sigma_d\sqrt{2\log d}$ (solid line).

values of ξ near and beyond the threshold t, as is evident in the figure. The graphs show the qualitative features captured in the inequalities (4) - (7). Quantitatively, at $d = 64$, the variance bound (4) for $t = \sigma_d\sqrt{2\log d}$ gives scaled MSE bound $(\sigma^2(\xi) + t^2)/\sigma^2(\xi) = 4.25$ at $\xi = 50 = (7.07)^2$ compared with the actual scaled value $r(\xi,t)/\sigma^2(\xi) \doteq 3.75$ shown in the figure.

2.2 Oracle Inequalities

For applications of these bounds, in analogy with the Gaussian case, we set $t = t_d = \sigma_d\sqrt{2\log d}$, where as before $\sigma_d^2 = \text{Var}\,\chi_d^2 = 2d$. This choice might be motivated by the inequality

$$P\{\chi_d^2 - d \geq \sigma_d\sqrt{2\log d}\} \leq 1/(2d), \tag{12}$$

which shows that if $\xi = 0$, then $\hat{\xi}_{t_d} = 0$ with probability approaching 1 as $d \to \infty$. Thus, there is a vanishing chance that $\hat{\xi}_{t_d}$ will spuriously assert the presence of structure when ξ is actually 0. Formula (12) follows from Lemma 6.1 for large d ($d \geq 72$ will do), while for smaller d, (12) may be verified numerically.

We give two inequalities – the first relates MSE to ideal risk, while the latter is slightly more convenient for the application to adaptive estimation of $\int f^2$ in Section 4. The proof is given in the appendix.

Corollary 2.1 *Let* $t_d = \sigma_d\sqrt{2\log d}$. *Then*

$$r(\xi, t_d) \leq (2\log d + 1)\{1 + \min(\xi^2, \sigma^2(\xi))\} \qquad d \geq 3, \tag{13}$$

and

$$r(\xi, t_d) \leq 2/\log d + \min\{2\xi^2, \sigma^2(\xi) + t_d^2\}, \qquad d \geq 18. \tag{14}$$

We record the arbitrary noise level version of (14) for use in Section 4.

Corollary 2.2 *Suppose that* $Y \sim N_d(\theta, \epsilon^2 I)$ *and that one seeks to estimate* $\rho = |\theta|^2$ *using* $|Y|^2 \sim \epsilon^2\chi^2(\rho/\epsilon^2)$. *Suppose that the estimator* $\hat{\rho}_{t_d}$ *and variance function* $\sigma^2(\rho)$ *are defined by* (1) *and* (2) *respectively. Then*

$$E(\hat{\rho}_{t_d} - \rho)^2 \leq \frac{2\epsilon^4}{\log d} + \min\{2\rho^2, \sigma^2(\rho) + \epsilon^4 t_d^2\}. \tag{15}$$

3 Lower Bounds

This section argues that the bounds (13) and (14) are sharp for d large, in the sense that no other estimator can asymptotically satisfy a uniformly better bound.

We use a standard Bayesian two point prior method, but with non-standard loss function. The resulting bound in the Gaussian case, Proposition 3.1, is then carried over to the chi-square setting via asymptotic normality, to give Proposition 3.2.

Let $\{P_\theta\}$ be a family of probability measures on $\mathbb{R}$, indexed by $\theta \in \Theta \subset \mathbb{R}$. Denote point mass at θ by ν_θ and consider two point prior distributions $\pi = \pi_0\nu_{\theta_0} + \pi_1\nu_{\theta_1}$. To use *weighted* squared error measure $\tilde{L}(a, \theta) = l(\theta)(a - \theta)^2$, we will need loss-weighted versions of these priors. With $l_i = l(\theta_i)$, these are

$$\tilde{\pi} = \tfrac{l_0\pi_0}{L}\,\nu_{\theta_0} + \tfrac{l_1\pi_1}{L}\,\nu_{\theta_1}, \qquad L = l_0\pi_0 + l_1\pi_1.$$

Denote the corresponding posterior probabilities *for* $\tilde{\pi}$ by

$$\eta(x) = P_{\tilde{\pi}}(\{\theta_0\}|x), \qquad \bar{\eta}(x) = P_{\tilde{\pi}}(\{\theta_1\}|x) = 1 - \eta(x).$$

Lemma 3.1 *With the previous definitions and for* $\theta_0 = 0$,

$$\tilde{R} := \inf_{\hat{\theta}} \sup_{\theta}\; l(\theta)E_\theta[\hat{\theta}(X) - \theta]^2 \geq l_1\pi_1\theta_1^2\; E_{\theta_1}\eta^2(X). \tag{16}$$

Proof The minimax risk $\tilde{R}$ is bounded below by the Bayes risk $\tilde{B}(\pi)$, using prior distribution π and loss function $\tilde{L}(a,\theta)$:

$$\tilde{R} \geq \tilde{B}(\pi) = \inf_{\hat{\theta}} \int \pi(d\theta)\, E_\theta l(\theta)[\hat{\theta}(X) - \theta]^2.$$

At least to aid intuition, it helps to convert this into a Bayes risk for squared error loss with modified prior $\tilde{\pi}$ given above, so that

$$\tilde{B}(\pi) = L \inf_{\hat{\theta}} \int \tilde{\pi}(d\theta)\, E_\theta[\hat{\theta}(X) - \theta]^2 =: B(\tilde{\pi}).$$

For squared error loss, now, the Bayes estimator $\hat{\theta}_{\tilde{\pi}}$ that attains the minimum $B(\tilde{\pi})$ is given by the posterior mean, which in the two point case with $\theta_0 = 0$ takes the simple form

$$\hat{\theta}_{\tilde{\pi}}(x) = E_{\tilde{\pi}}[\theta|x] = \theta_1 P(\{\theta_1\}|x) = \theta_1 \bar{\eta}(x),$$

which implies the desired formula (16):

$$B(\tilde{\pi}) = L \int \tilde{\pi}(d\theta) E_\theta[\hat{\theta}_{\tilde{\pi}}(X) - \theta]^2 \geq L\tilde{\pi}(\{\theta_1\}) E_{\theta_1}[\theta_1 \bar{\eta}(X) - \theta_1]^2. \qquad \blacksquare$$

Proposition 3.1 *Suppose* $X \sim N(\theta, 1)$. *Then as* $d \to \infty$,

$$\inf_{\hat{\theta}} \sup_{\theta} \frac{E(\hat{\theta}(X) - \theta)^2}{d^{-1} + (\theta^2 \wedge 1)} \geq (2 \log d)(1 + o(1)). \tag{17}$$

Proof In Lemma 3.1, let P_θ correspond to $X \sim N(\theta, 1)$ and take $l(\theta) = [d^{-1} + (\theta^2 \wedge 1)]^{-1}$. Choose $\theta_0 = 0$ and $\theta_1 = \theta_d \gg 1$ (to be specified below), so that $l_0 = d$ and $l_1 = (1 + d^{-1})^{-1}$.

Set $\pi_0 = 1/\log d$ and $\pi_1 = 1 - \pi_0$ so that $L = \pi_0 l_0 + \pi_1 l_1 \sim d/\log d$ and the loss weighted prior $\tilde{\pi} = (1-\epsilon)\nu_0 + \epsilon \nu_{\theta_d}$ with $\epsilon = \pi_1 l_1 / L \sim \log d/d$ small.

The idea is that with ϵ small, we choose θ_d so that even for x near θ_d, $\eta(x) = P_{\tilde{\pi}}(\{0\}|x) \approx 1$. Thus, with probability essentially ϵ we estimate $\hat{\theta}_{\tilde{\pi}} \approx 0$ even though $\theta = \theta_d$ and so incur an error of about θ_d^2.

Now the details. Write $g(x;\theta)$ for the $N(\theta, 1)$ density and, since we will recenter at θ_d, put $x = \theta_d + z$. Then the likelihood ratio

$$l_\infty(z;\theta_d) = \frac{g(x;\theta_d)}{g(x;0)} = \frac{\phi(z)}{\phi(\theta_d + z)} = \exp\{\theta_d z + \theta_d^2/2\}. \tag{18}$$

Of course, the posterior probability $\eta(x)$ can be written in terms of the likelihood ratio as

$$\eta(\theta_d + z) = [1 + \tfrac{\epsilon}{1-\epsilon} l_\infty(z; \theta_d)]^{-1}. \tag{19}$$

Put $a_d = \log\log d$ and specify θ_d as the solution to $\eta(\theta_d + a_d) = 1/2$, so that

$$\theta_d a_d + \theta_d^2/2 = \log\frac{1-\epsilon}{\epsilon} = \log d - \log\log d + o(1), \tag{20}$$

and hence $\theta_d \sim \sqrt{2\log d}$, and also

$$\eta(\theta_d + z) = [1 + \exp\{\theta_d(z - a_d)\}]^{-1} \to 1$$

for all fixed z. Consequently

$$E_{\theta_d}\eta^2(X) = \int \eta^2(\theta_d + z)\phi(z)dz \to 1$$

by the dominated convergence theorem. Since $l_1\pi_1 \sim 1$ and $\theta_d^2 \sim 2\log d$, the result now follows from Lemma 3.1. ∎

With the Gaussian bound as template, we turn to the corresponding result for the non-central chi-squared distributions.

Proposition 3.2 *Suppose that* $W_d \sim \chi_d^2(\xi)$. *As* $d \to \infty$,

$$\inf_{\hat{\xi}} \sup_{\xi} \frac{E(\hat{\xi}(W_d) - \xi)^2}{1 + \min\{\xi^2, \sigma^2(\xi)\}} \geq (2\log d)(1 + o(1)). \tag{21}$$

Proof The rescaled variable $X = (W - d)/\sqrt{2d}$ has mean $\theta = \xi/\sqrt{2d}$, variance $\sigma_X^2(\theta) = 1 + \theta\sqrt{8/d}$ and is asymptotically Gaussian as $d \to \infty$. Let $g_d(x;\theta)$ denote its density function. As in the proof of Proposition 3.1, we recenter at $\theta_d \sim \sqrt{2\log d}$ (to be defined precisely below) and form the likelihood ratio

$$l_d(z;\theta_d) = \frac{g_d(\theta_d + z; \theta_d)}{g_d(\theta_d + z; 0)}. \tag{22}$$

With $l_\infty(y;\theta_d)$ the corresponding Gaussian likelihood ratio defined at (18),

$$u_d(y) = \frac{l_d(y;\theta_d)}{l_\infty(y;\theta_d)} \to 1, \qquad \text{as } d \to \infty \tag{23}$$

uniformly in $|y| \leq \log d$, say (see Appendix.)

To an arbitrary estimator $\hat{\xi}(W_d)$, associate $\hat{\theta}(X) = \hat{\xi}(W_d)/\sqrt{2d}$. Hence

$$\frac{E(\hat{\xi} - \xi)^2}{1 + \min\{\xi^2, \sigma^2(\xi)\}} = \frac{E(\hat{\theta} - \theta)^2}{1/(2d) + \min\{\theta^2, \sigma_X^2(\theta)\}}.$$

Now proceed as in Proposition 3.1: set $l(\theta) = [(2d)^{-1} + \min\{\theta^2, \sigma_X^2(\theta)\}]^{-1}$ so that $l_0 = 2d, l_1 = [1 + \theta_d\sqrt{8/d} + (2d)^{-1}]^{-1}$. Set $\pi_0 = 1/\log d$, and $\pi_1 = 1 - \pi_0$ and define $\tilde{\pi}$ and ϵ as before. From Lemma 3.1,

$$\tilde{R} \geq l_1\pi_1\theta_d^2 \int \eta^2(\theta_d + z)g_d(\theta_d + z)dz \tag{24}$$

and $l_1\pi_1 \sim 1$. The posterior probability $\eta(\theta_d + z)$ is again defined by (19), with l_∞ now replaced by l_d. Again put $a_d = \log\log d$ and define θ_d as the solution to $\eta(\theta_d + a_d) = 1/2$, which is equivalent to

$$\tfrac{\epsilon}{1-\epsilon} u_d(a_d) e^{a_d\theta_d + \theta_d^2/2} = 1.$$

In view of (23), (20) shows that here too $\theta_d \sim \sqrt{2\log d}$. From the definition of θ_d and using (23),

$$\eta(\theta_d + z) = \left[1 + \frac{l_d(z;\theta_d)}{l_d(a_d;\theta_d)}\right]^{-1} = \left[1 + \frac{u_d(z)}{u_d(a_d)} e^{\theta_d(z - a_d)}\right]^{-1} \to 1.$$

By Pratt's version of the dominated convergence theorem Pratt (1960) or Hall and Heyde (1980, p 281.), the integral in (24) converges to one and this completes the proof of (21). ∎

Remark. An analagous result holds if the denominator in (21) is replaced by (RHS of (14))/$2\log d$.

4 Illustration: Estimation of $\int f^2$

The estimation of quadratic functionals such as $\int f^2$ or more generally $\int (D^l f)^2$ for non-negative integer l has received sustained attention in the last three decades. See, for example Hall and Marron (1987); Bickel and Ritov (1988); Hall and Johnstone (1992) Ibragimov et al. (1986); Donoho and Nussbaum (1990); Fan (1991); Birgé and Massart (1995); Laurent (1996); Gayraud and Tribouley (1999); Laurent and Massart (1998) and the references therein.

Bickel and Ritov (1988) found a curious 'elbow' phenomenon: for $\int f^2$, if f has Hölder smoothness $\alpha > 1/4$, efficient estimation at mean squared error rate n^{-1} is possible, while for $\alpha \leq 1/4$, the best MSE rate is $n^{-r} = n^{-8\alpha/(1+4\alpha)}$. The problem of "adaptive estimation" concerns whether one can, *without knowledge of* α, build estimators that achieve these optimal

rates for every α in a suitable range. Alas, Efromovich and Low (1996a,b) showed that this is *not* possible as soon as $0 < \alpha \leq 1/4$. They go on to adapt version of Lepskii's general purpose step-down adaptivity construction (Lepskii, 1991) to build an estimator that is efficient for $\alpha > 1/4$ and attains the best rate (logarithmically worse than n^{-r}) that is possible simultaneously for $0 < \alpha \leq 1/4$.

The treatment here is simply an illustration of chi-square thresholding to obtain the Efromovich-Low result. Two recent works (received after the first draft was completed) go much further with the $\int f^2$ problem. Gayraud and Tribouley (1999) use hard thresholding to derive Efromovich-Low, and go on to provide limiting Gaussian distribution results and even confidence intervals. Laurent and Massart (1998) derive *non-asymptotic* risk bounds via penalized least squares model selection and consider a wide family of functional classes including ℓ_p and Besov bodies.

Consider the white noise model, where one observes $Y_t = \int_0^t f(s)ds + \epsilon W_t$, $0 \leq t \leq 1$, where W_t is standard Brownian motion and $f \in L_2([0,1])$ is unknown. It is desired to estimate $Qf = \int f^2$. We use the Haar orthonormal basis, defined by $h(t) = I_{[0,1/2]} - I_{[1/2,1]}$ and $\psi_I(t) = 2^{j/2}h(2^j t - k)$ for indices $I = (j,k)$ with $j \in \mathbb{N}$ and $k \in \mathcal{I}_j = \{1, \ldots, 2^j\}$. We add the scaling function $\psi_{(-1,0)} = I_{[0,1]}$. In terms of the orthonormal coefficients, the observations take the dyadic sequence form

$$y_I = \theta_I + \epsilon z_I \tag{25}$$

where $\theta_I = \int \psi_I f$ and the noise variables z_I are i.i.d. standard Gaussian. By Parseval's identity, $Qf = \sum \theta_I^2$, and we group the coefficients by level j:

$$Qf = \sum_j \rho_j, \qquad \rho_j = \sum_{\mathcal{I}_j} \theta_I^2,$$

where $|\mathcal{I}_j| = d_j = 2^j$ (and equals 1 for $j = -1$). The corresponding sums of data coefficients have non-central χ^2 distributions:

$$|y_j|^2 = \sum_{\mathcal{I}_j} y_I^2 \sim \epsilon^2 \chi^2_{d_j}(\xi_j), \qquad \xi_j = \epsilon^{-2}\rho_j.$$

We estimate Qf by estimating ρ_j at each level separately and then adding.

To quantify smoothness, we use, for simplicity, the Hölder classes, which can be expressed for $\alpha < 1$ in terms of the Haar wavelet coefficients as

$$\Theta^\alpha(C) = \{\theta : |\theta_I| \leq C 2^{-(\alpha+1/2)j} \text{ for all } I\}. \tag{26}$$

See Meyer (1990, Sec. 6.4), or Härdle et al. (1998, Theorem 9.6) for a specific result. Thus, in terms of the levelwise squared ℓ_2 norms:

$$\theta \in \Theta^\alpha(C) \;\Rightarrow\; \rho_j \leq \bar{\rho}_j = C^2 2^{-2\alpha j} \qquad \text{for all } j \geq 0. \tag{27}$$

In smoother cases, the low frequencies are most important, whereas in rough settings, higher frequencies are critical. For the lower frequencies, define $\mathcal{I}_e = \{I : j \le j_0\}$. The estimate combines unbiased estimation at these lower frequencies (where efficiency is the goal)

$$\hat{Q}_e = \sum_{I \in \mathcal{I}_e} (y_I^2 - \epsilon^2), \qquad 2^{-j_0} = \epsilon^2 \sqrt{\log_2 \epsilon^{-2}}, \tag{28}$$

with thresholding at higher frequencies

$$\hat{Q}_t = \sum_{j=j_0+1}^{j_1} \hat{\rho}_j, \qquad 2^{-j_1} = \epsilon^4 \tag{29}$$

where, in notation matching Section 2, we put $\sqrt{2d_j} \cdot t_j = \sqrt{2 \log d_j}$ and

$$\hat{\rho}_j = (|y_j|^2 - d_j\epsilon^2 - t_j\epsilon^2)_+.$$

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

Theorem 4.1 *Let observations be taken from the Gaussian dyadic sequence model* (25) *and let the estimator* $\hat{Q} = \hat{Q}_e + \hat{Q}_t$ *of* $Qf = \int f^2$ *be defined via* (28) *and* (29). *Let* $r = 8\alpha/(1 + 4\alpha)$,

(i) For $0 < \alpha \le 1/4$,

$$\sup_{f \in \Theta^\alpha(C)} E(\hat{Q} - Qf)^2 \le cC^{2(2-r)}(\epsilon^2 \sqrt{\log(C\epsilon^{-1})})^r(1 + o(1)), \tag{30}$$

(ii) For $\alpha > 1/4$,

$$\sup_{f \in \Theta^\alpha(C)} |E(\hat{Q} - Qf)^2 - 4\epsilon^2 Qf| = o(\epsilon^2). \tag{31}$$

Proof. Decompose $Qf = \sum \rho_j = Q_e f + Q_t f + Q_r f$ where the ranges of summation match those of $\hat{Q}_e$ and $\hat{Q}_t$ in (28) and (29). From the triangle inequality for $\|\delta\| = \sqrt{E\delta^2}$,

$$\sqrt{E(\hat{Q} - Qf)^2} \le \sqrt{E(\hat{Q}_e - Q_e f)^2} + \sqrt{E(\hat{Q}_t - Q_t f)^2} + Q_r f. \tag{32}$$

The tail bound is negligible in all cases: from (27) and (29)

$$Q_r f \le C^2 \sum_{j_1}^{\infty} 2^{-2\alpha j} \le cC^2 2^{-2\alpha j_1} = cC^2 \epsilon^{8\alpha}.$$

Efficient Term. Since $\hat{Q}_t$ is unbiased, we have, using (28) and (2),

$$E(\hat{Q}_e - Q_e f)^2 = \operatorname{Var} \hat{Q}_e = 4\epsilon^2 \sum_{I \in \mathcal{I}_e} \theta_I^2 + 2^{j_0+2}\epsilon^4. \tag{33}$$

The second term is always negligible: from (28), $2^{j_0}\epsilon^4 = \epsilon^2(\log_2 \epsilon^{-2})^{-1/2} = o(\epsilon^2)$. Further, using (27), for $f \in \Theta^\alpha(C)$,

$$Qf - \sum_{\mathcal{I}_e} \theta_I^2 = \sum_{j_0+1}^{\infty} \rho_j \le cC^2 2^{-2\alpha j_0} = o(1).$$

Combining the two previous displays

$$\sup_{f \in \Theta^\sigma(C)} |E(\hat{Q}_e - Q_e f)^2 - 4\epsilon^2 Qf| = o(\epsilon^2). \tag{34}$$

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

$$\sqrt{E(\hat{Q}_t - Q_t f)^2} \le \sum_{j_0}^{j_1} \sqrt{E(\hat{\rho}_j - \rho_j)^2} =: T_\epsilon(f). \tag{35}$$

The oracle inequality (15) yields

$$E(\hat{\rho}_j - \rho_j)^2 \le 2\epsilon^4 + \min\{2\rho_j^2, \sigma^2(\rho_j) + t_j^2\epsilon^4\}, \tag{36}$$

where $\sigma^2(\rho_j) = 2^{j+1}\epsilon^4 + 4\epsilon^2\rho_j$. First, we evaluate

$$\bar{T}_\epsilon = \sum_{j_0}^{j_1} \min\{\bar{\rho}_j, t_j\epsilon^2\}. \tag{37}$$

Since $\bar{\rho}_j = C^2 2^{-2\alpha j}$ is geometrically decreasing in j and $t_j\epsilon^2 = cj^{1/2}2^{j/2}\epsilon^2$ is geometrically increasing in j, we must have $\bar{T}_\epsilon \le c(\alpha)\bar{\rho}_{j_2}$, where $j_2 = j_2(\epsilon; C, \alpha)$ is the crossing point, namely the (usually non-integer) solution to $j2^{(1+4\alpha)j} = cC^4\epsilon^{-4}$. As spelled out in (56) in the Appendix, as $\epsilon \to 0$,

$$\bar{T}_\epsilon \le c(\alpha)\bar{\rho}_{j_2} = cC^2 2^{-2\alpha j_2} \sim cC^{2-r}(\epsilon^2 \log(C\epsilon^{-1}))^{r/2}. \tag{38}$$

We conclude by checking that on $\Theta^\alpha(C)$, $T_\epsilon(f) \le c\bar{T}_\epsilon$ for small ϵ. Looking at the terms in (36), we observe first that $j_1\epsilon^2 = o(\bar{T}_\epsilon)$. Now let j_3 be the solution to $\bar{\rho}_j = t_j^2\epsilon^2$, or equivalently $j2^{(1+2\alpha)j} = cC^2\epsilon^{-2}$. Again using (56),

$$2^{-j_3} \sim c\left(\tfrac{\epsilon^2}{C^2} \log_2 \tfrac{C}{\epsilon}\right)^{\frac{1}{1+2\alpha}},$$

and so (28) shows that for small ϵ, $j_3(\epsilon) \le j_0(\epsilon)$. From this it follows that for $\theta \in \Theta^\alpha(C)$ and $j \ge j_0$ we have $\sigma^2(\rho_j) \le ct_j^2\epsilon^4$, so that (37) is indeed the dominant term in (35).

In the efficient case, (38) is negligible, so that (31) follows from (32) and (35). In the nonparametric zone $0 < \alpha \leq 1/4$, (38) shows that (34) is negligible relative to (35), from which we obtain (30). ∎

Remark. Haar wavelets have been ingeniously used by Kerkyacharian and Picard (1996) in the context of estimating $\int f^2$ and especially $\int f^3$. However, thresholding is not used there, nor is adaptivity considered.

5 Remarks on weighted chi-square

Suppose, as before, that $y_k \sim N(\theta_k, \epsilon^2), k = 1, \dots d$ are independent, but that now we desire to estimate $\rho_\alpha = \sum_1^d \alpha_k \theta_k^2$ with $\alpha_k > 0$. Such a scenario emerges in estimation of $\int (D^l f)^2$, for example. Then the natural unbiased estimator $\tilde{\rho}_{U,\alpha} = \sum_1^d \alpha_k (y_k^2 - \epsilon^2)$ is no longer a shift of a chi-square variate.

If the weights are comparable, say $1 \leq \alpha_k \leq \bar{\alpha}$ for all k, then an extension of the risk bounds of Theorem 2.1 is possible. We cite here only the extension of Corollary 2.2, referring to Johnstone (2000) for further results and details.

Proposition 5.1 *With the above notations, set $t_d = \sigma_d \sqrt{2 \log d}$. There exists an absolute constant γ such that*

$$E[\hat{\rho}(U_\alpha; \bar{\alpha} t_d) - \rho_\alpha]^2 \leq \gamma \bar{\alpha}^2 \Big[\frac{2\epsilon^4}{\log d} + \min\{2\rho_\alpha^2, \sigma^2(\rho_\alpha) + \epsilon^4 t_d^2\} \Big].$$

6 Appendix

6.1 Central χ^2 distributions

Write $f_d(w) = e^{-w/2} w^{d/2-1} / 2^{d/2} \Gamma(d/2)$ for the density function of χ_d^2 and $\tilde{F}_d(w) = \int_w^\infty f_d(u) du$ for the survivor form of the distribution function. We note the relations

$$w f_d(w) = d\, f_{d+2}(w), \tag{39}$$

$$w^2 f_d(w) = d(d+2) f_{d+4}(w), \tag{40}$$

$$D_w f_{d+2}(w) = \tfrac{1}{2}[f_d(w) - f_{d+2}(w)], \tag{41}$$

where D_w denotes partial derivative w.r.t. w. Recall that the Poisson p.d.f. is denoted by $p_\lambda(x) = e^{-\lambda} \lambda^x / \Gamma(x+1)$. From (41) or via probabilistic arguments,

$$\tilde{F}_{d+2}(w) - \tilde{F}_d(w) = p_{w/2}(d/2). \tag{42}$$

6.2 A moderate deviations bound for central χ^2

Lemma 6.1 *Let $W_d \sim \chi^2_d$ and $\sigma_d^2 = VarW_d = 2d$. If $0 \le s \le \sqrt{d/8}$, then*

$$P(W_d - d \ge s\sigma_d) \le (\frac{1}{s} + \frac{2}{\sigma_d})\frac{1}{\sqrt{2\pi}} \exp\{\frac{-s^2/2}{1 + 4s/3\sigma_d}\} \tag{43}$$

In consequence, if $d \ge 16$ and $0 \le s \le d^{1/6}$, then

$$P(W_d - d \ge s\sigma_d) \le s^{-1}e^{-s^2/2}. \tag{44}$$

This bound is an analogue of the Gaussian tail bound $P(Z \ge s) \le \phi(s)/s$, to which it reduces as $d \to \infty$ whenever $s = o(d^{1/2})$. It may be compared with two existing bounds, each derived for more general purposes. First, the Tsirelson-Ibragimov-Sudakov inequality for Lipschitz functions of a standard Gaussian vector yields, for $s \ge 0$,

$$P(W_d - d \ge \sqrt{2}s\sigma_d + s^2) \le e^{-s^2/2},$$

while the more refined inequality of Laurent and Massart (1998, Lemma 1) has as corollary, for positive s:

$$P(W_d - d \ge s\sigma_d + s^2) \le e^{-s^2/2}.$$

Substituting $s = \sqrt{2\log d}$ in this latter inequality shows that it does not suffice for conclusion (12).

Proof For $w \ge d$, $f_d(w) \le f_{d+2}(w)$, so it suffices to bound $\tilde{F}_{d+2}(s_1)$, where we have set $s_1 = d + s\sigma_d$. Equalities (39) and (41) combine to give

$$f_{d+2}(w) = (1 - d/w)^{-1}[-2D_w f_{d+2}(w)].$$

Now use the idea behind the bound $\tilde{\Phi}(s) \le \phi(s)/s$: for $w \ge s_1$, $1 - d/w \ge 1 - d/s_1$ and so

$$\tilde{F}_{d+2}(s_1) \le 2(1 - d/s_1)^{-1} f_{d+2}(s_1). \tag{45}$$

Stirling's formula, $\Gamma(d/2 + 1) \ge \sqrt{2\pi}e^{-d/2}(d/2)^{(d+1)/2}$, implies

$$\begin{aligned} 2f_{d+2}(s_1) &\le (\pi d)^{-1/2}(s_1/d)^{d/2}e^{-(s_1-d)/2} \\ &= (\pi d)^{-1/2}\exp\{(d/2)[\log(1+v) - v]\}, \end{aligned} \tag{46}$$

where $s_1 = d + dv$. The inequality

$$\log(1+v) - v \le \frac{-v^2/2}{1 + 2v/3} \qquad 0 \le v \le \tfrac{1}{2}, \tag{47}$$

holds because the left side equals $(-v^2/2)\sum_0^\infty \frac{2}{k+2}(-v)^k$, while the right side equals $(-v^2/2)\sum_0^\infty (\frac{2}{3})^k(-v)^k$ and successive sums of pairs of terms in the first series dominate the corresponding pair sums in the second for $0 \le v \le 1/2$.

If $s_1 = d + s\sqrt{2d}$, then $v = s\sqrt{2/d}$ and inserting (47) in (46),

$$\sqrt{2d} f_{d+2}(d + s\sqrt{2d}) \le \frac{1}{\sqrt{2\pi}} \exp\{\frac{-s^2/2}{1 + 4s/3\sigma_d}\}.$$

Substituting this into (45) yields (43). The ratio of the right side of (43) to (44) is bounded above for $d \ge 16$ by

$$\left(1 + \tfrac{2s}{\sigma_d}\right) \exp\left\{\tfrac{s^2}{2}\left(1 - \tfrac{1}{1+4s/3\sigma_d}\right)\right\} \le \left(1 + \tfrac{\sqrt{2}}{d^{1/3}}\right) e^{\sqrt{2}/3} \le 1. \quad \blacksquare$$

6.3 Non-central $\chi_d^2(\xi)$

A non-central $\chi_d^2(\xi)$ may be defined as the distribution of $W_d = |X|^2$ for $X \sim N_d(\theta, I)$, with the non-centrality parameter $\xi = |\theta|^2$. We have

$$EW_d = d + \xi, \qquad \text{Var } W_d = 2d + 4\xi.$$

It may be realized as a Poisson$(\xi/2)$ mixture of central χ_{d+2j}^2 distributions:

$$f_{\xi,d}(w) = \sum_{j=0}^{\infty} p_{\xi/2}(j) f_{d+2j}(w) = e^{\xi\Delta/2} f_d(w), \tag{48}$$

where the latter equality is a formal representation in terms of the difference operator $\Delta f_d = f_{d+2} - f_d$. (attributed by Johnson et al. (1995, p. 439) to Bol'shev and Kuznetzov (1963)). Formally differentiating (48) with respect to w and using the difference relation (41), one obtains a useful identity (parts of which appear in Johnson et al. (1995, pp. 442-3)):

$$D_\xi f_{\xi,d}(w) = \tfrac{1}{2}[f_{\xi,d+2}(w) - f_{\xi,d}(w)] = -D_w f_{\xi,d+2}(w). \tag{49}$$

Proof of Theorem 2.1 To obtain (10), differentiate $r(\xi, t) = \int_0^\infty [(w - t_1)_+ - \xi]^2 f_{\xi,d}(w)dw$ to get

(50)

$$D_\xi r(\xi, t) = -2\int_0^\infty [(w - t_1)_+ - \xi] f_{\xi,d} + \int_0^\infty D_w[(w - t_1)_+ - \xi]^2 f_{\xi,d+2}$$

$$= 2\xi \int_0^{t_1} f_{\xi,d} - 2\int_{t_1}^\infty (w - t_1 - \xi)[f_{\xi,d} - f_{\xi,d+2}] \tag{51}$$

$$= 2\xi \int_0^{t_1} f_{\xi,d+2} + 4\xi f_{\xi,d+2}(t_1) - 4\int_{t_1}^\infty (w - t_1 - \xi) D_w f_{\xi,d+2}. \tag{52}$$

Here (50) uses (49) and partial integration, (51) collects terms, (52) uses (49) twice, and a final partial integration leads from (52) to (10).

To derive (11), differentiate (10) and use (49) to get

$$D_\xi^2 r(\xi,t) = 2\int_0^{t_1} f_{\xi,d+2} - 2\xi\int_0^{t_1} D_w f_{\xi,d+4} - 4\int_{t_1}^{\infty} D_w f_{\xi,d+4},$$

and evaluate the final two integrals.

To complete the proof of (7), combine (39) with (48):

$$(53)\quad f_{\xi,d+4}(t_1) = \sum_0^\infty p_{\xi/2}(j)\frac{t_1}{d+2+2j} f_{d+2+2j}(t_1) \le (1+t/d) f_{\xi,d+2}(t_1).$$

Again using (49), $2f_{\xi,d+2}(t_1) = \int_{t_1}^\infty f_{\xi,d+2} - f_{\xi,d} \le \int_{t_1}^\infty f_{\xi,d+2}$. Thus, in combination with (11) and (53), we arrive at (7):

$$D_\xi^2 r(\xi,t_1) \le 2\int_0^{t_1} f_{\xi,d+2} + 2(1+t/d)\int_{t_1}^\infty f_{\xi,d+2} \le 2(1+t/d).$$

Finally, for $r(0,t)$, we derive first an expression useful for numerical evaluation. Using (8), then (40) and (39), followed by (42) and the Poisson p.d.f. identity $(x+1)p_\lambda(x+1) = \lambda p_\lambda(x)$, we obtain

$$\begin{aligned} r(0,t) &= \int_{t_1}^\infty (w-t_1)^2 f_d(w)dw \\ &= d(d+2)\tilde F_{d+4}(t_1) - 2t_1 d\tilde F_{d+2}(t_1) + t_1^2 \tilde F_d(t_1), \\ &= [(d-t_1)^2 + 2d]\tilde F_d(t_1) + d(d+2)p_{t_1/2}(1+d/2) \\ &\qquad + d(d+2-2t_1)p_{t_1/2}(d/2)] \\ &= (t^2+2d)\tilde F_d(t_1) - d(t-2)p_{t_1/2}(d/2). \end{aligned}$$

Turning now the bound (6), use twice both integration by parts and (45) to obtain

$$r(0,t) = 2\int_{t_1}^\infty dw \int_w^\infty dv\, \tilde F_d(v) \le \frac{8t_1^2}{(t+2)^2}\tilde F_d(t_1). \qquad \blacksquare$$

Proof of Corollary 2.1 Since $\sigma^2(\xi) = 2d+4\xi$ and $t_d^2 = 2d\cdot 2\log d$, bound (4) yields immediately

$$r(\xi,t_d) \le (2\log d + 1)\sigma^2(\xi).$$

For (13) it remains to check that $r(\xi, t_d) \le (2\log d + 1) + (2\log d + 1)\xi^2$. In view of (5), this will follow from

$$\eta_2 \le 2\log d, \qquad r(0, t_d) + \eta_1 \le 2\log d + 1. \tag{54}$$

Bearing in mind (3) and bound (6), it can be verified numerically that the inequalities (54) are valid for $d \ge 3$ and $d \ge 2$ respectively.

For (14) we need only verify that $r(\xi, t_d) \le 2(\log d)^{-1} + 2\xi^2$ and this will follow from more stringent versions of (54):

$$\eta_2 \le 1, \qquad r(0, t_d) + \eta_1 \le 2/\log d. \tag{55}$$

These can be verified numerically for $d \ge 14$ and $d \ge 18$ respectively. For completeness, we show that (55) (and hence (54)) follow, for *large* d, from (44). Indeed, some algebra shows

$$2\tilde{F}_d(d + t_d) \le 2\tilde{F}_{d+2}(d + t_d) = \eta_1 \le \frac{c}{d\log d}.$$

Using bound (6), we then have

$$r(0, t_d) \le \frac{c}{d\log d}\Big(1 + \frac{d}{t_d}\Big)^2 = \frac{c}{\log d}\Big(\frac{1}{\sqrt{d}} + \frac{1}{2\sqrt{\log d}}\Big)^2.$$

Taking d large in these last two displays implies (55). ∎

Proof of (23) Let $N \sim \text{Poisson}(\xi/2)$. From (48)

$$\frac{f_{\xi,d}(w)}{f_d(w)} = E\,\frac{f_{d+2N}(w)}{f_d(w)} = E\Big[\Big(\frac{w}{2}\Big)^N \frac{\Gamma(d/2)}{\Gamma(d/2 + N)}\Big].$$

The likelihood ratios (22) involve the transformed variable $X = (W - d)/\sqrt{2d}$. Thus with $\theta = \xi/\sqrt{2d}$, we have $g_d(x;\theta) = \sqrt{2d} f_{\xi,d}(d + x\sqrt{2d})$. Using Stirling's formula $\Gamma(s) = \sqrt{2\pi} e^{-s+\gamma/s} s^{s-1/2}$, with $|\gamma| \le 1/12$, and setting $m = d/2$ and $x = \theta_d + z$, we obtain the representation

$$\frac{l_d(z;\theta_d)}{l_\infty(z;\theta_d)} = E\exp\{L_m + R_m\}, \qquad R_m = \tfrac{\gamma}{m} + \tfrac{\gamma}{m+N},$$

and

$$L_m = N - (m + N - \tfrac{1}{2})\log(1 + \tfrac{N}{m}) + N\log(1 + \tfrac{\theta_{2m}+z}{\sqrt{m}}) - \theta_{2m} z - \theta_{2m}^2/2.$$

Write the Poisson variable in the form $N = \theta_d\sqrt{m} + Z_m\sqrt{\theta_{2m}}m^{1/4}$, so that $Z_m \xrightarrow{\mathcal{D}} Z$ as $m \to \infty$. Expanding the logarithms in Taylor series and collecting

terms shows that $L_m = o_p(1)$ uniformly in $|z| \le \log d$, and along with $R_m = o_p(1)$, this completes the proof of (23). ∎

Details for Theorem 4.1. Fix $\gamma > 0$. Let $x(z)$ be the solution of $x2^{\gamma x} = z$. The approximate solution $x_0 = x_0(z)$ given by

$$2^{\gamma x_0(z)} = \frac{\gamma z}{\log_2 z} \sim 2^{\gamma x(z)} \tag{56}$$

as $z \to \infty$. More precisely, for large z,

$$2^{\gamma x_0(z)} \in [c(z), 1]2^{\gamma x(z)}, \qquad c(z) \ge 1 - \frac{\log_2 \log_2 z}{\log_2 z}. \tag{57}$$

To verify (57), set $w = \log_2 z$, so that $\log_2 x(z) = w - \gamma x(z)$. Then

$$2^{\gamma x_0 - \gamma x} = \frac{\gamma x(z)}{w} = 1 - \frac{\log_2 x(z)}{w} \ge 1 - \frac{\log_2 w}{w}.$$

Acknowledgements. Many thanks to David Donoho for his ideas and collaboration in the first iteration of this project some years back, in a version based solely on Gaussian approximation. Thanks also to the Australian National University for hospitality and support when the first draft of this manuscript was written and to the referees for their careful reading and comments.

REFERENCES

Bickel, P. and Ritov, Y. (1988), 'Estimating integrated squared density derivatives: sharp best order of convergence estimates.', *Sankhya, Series A* **50**, 381–393.

Birgé, L. and Massart, P. (1995), 'Estimation of integral functionals of a density', *Annals of Statistics* **23**, 11–29.

Bol'shev, L. and Kuznetzov, P. (1963), 'On computing the integral $p(x, y) =$ ...', *Shurnal Vychislitelnoj Matematiki i Matematicheskoi Fiziki* **3**, 419–430. In Russian.

Chow, M. S. (1987), 'A complete class theorem for estimating a noncentrality parameter', *Annals of Statistics* **15**, 800–804.

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

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

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

Efromovich, S. and Low, M. (1996*b*), 'On optimal adaptive estimation of a quadratic functional', *Annals of Statistics* **24**, 1106–1125.

Fan, J. (1991), 'On estimation of quadratic functionals', *Annals of Statistics* **19**, 1273–1294.

Gayraud, G. and Tribouley, K. (1999), 'Wavelet methods to estimate an integrated quadratic functional: Adaptivity and asymptotic law', *Statistics and Probability Letters* **44**, 109–122.

Hall, P. G. and Marron, J. S. (1987), 'Estimation of integrated squared density derivatives', *Statistics and Probability Letters* **6**, 109–115.

Hall, P. and Heyde, C. (1980), *Martingale Limit Theory and Its Application*, Academic Press.

Hall, P. and Johnstone, I. (1992), 'Empirical functionals and efficient smoothing parameter selection', *Journal of the Royal Statistical Society, Series B* **54**, 475–530. with discussion.

Härdle, W., Kerkyacharian, G., Picard, D. and Tsybakov, A. (1998), *Wavelets, Approximation and Statistical Applications*, Springer.

Ibragimov, I. A., Nemirovskii, A. S. and Khas'minskii, R. Z. (1986), 'Some problems on nonparametric estimation in gaussian white noise', *Theory of Probability and its Applications* **31**, 391– 406.

Johnson, N. L., Kotz, S. and Balakrishnan, N. (1995), *Continuous Univariate Distributions, Volume 2, Second edition*, John Wiley and Sons.

Johnstone, I. M. (2000), Thresholding and oracle inequalities for weighted χ^2, To appear, *Statistica Sinica*.

Kerkyacharian, G. and Picard, D. (1996), 'Estimating nonquadratic functionals of a density using haar wavelets', *Annals of Statistics* **24**, 485–507.

Laurent, B. (1996), 'Efficient estimation of integral functionals of a density', *Annals of Statistics* **24**, 659–681.

Laurent, B. and Massart, P. (1998), Adaptive estimation of a quadratic functional by model selection, Technical report, Université de Paris-Sud, Mathématiques.

Lepskii, O. (1991), 'On a problem of adaptive estimation in gaussian white noise', *Theory of Probability and its Applications* **35**, 454–466.

Meyer, Y. (1990), *Ondelettes et Opérateurs, I: Ondelettes, II: Opérateurs de Calderón-Zygmund, III:* (with R. Coifman), *Opérateurs multilinéaires*, Hermann, Paris. English translation of first volume is published by Cambridge University Press.

Pratt, J. W. (1960), 'On interchanging limits and integrals', *Annals of Mathematical Statistics* **31**, 74–77.

Saxena, K. M. L. and Alam, K. (1982), 'Estimation of the non-centrality parameter of a chi-squared distribution', *Annals of Statistics* **10**, 1012–1016.

Department of Statistics
Sequoia Hall
Stanford University
Stanford CA 94305
U.S.A.
imj@stat.stanford.edu