# A penalised data-driven block shrinkage approach to empirical Bayes wavelet estimation

Xue Wang[*,a], Stephen G. Walker[a]

[a] *University of Kent at Canterbury, Canterbury, UK*

**Abstract**

In this paper we propose a simple Bayesian block wavelet shrinkage method for estimating an unknown function in the presence of Gaussian noise. A data–driven procedure which can adaptively choose the block size and the shrinkage level at each resolution level is provided. The asymptotic property of the proposed method, BBN (Bayesian BlockNorm shrinkage), is investigated in the Besov sequence space. The numerical performance and comparisons with some of existing wavelet block denoising methods show that the new method can achieve good performance but with the least computational time.

*Key words:* Asymptotics, Besov space, BlockNorm shrinkage, Adaptivity

## 1. Introduction

Consider the nonparametric regression model:

$$y_i = f(i/n) + \sigma z_i, \qquad i = 1, \ldots, n, \tag{1}$$

where $\sigma$ is the noise level and $\{z_i\}$ are independent standard normal random variables. The problem of interest is to estimate the unknown regression function $f(\cdot)$, which belongs to a certain class of function $F[0, 1]$, using the observed sample $\{y_i\}$.

Wavelet based procedures have shown their suitability for such settings, and non-parametric estimators of $f(\cdot)$ can be readily obtained by applying various shrinkage rules on the wavelet transformed data.

[*] Corresponding author

*Email addresses:* `X.Wang@kent.ac.uk` (Xue Wang), `S.G.Walker@kent.ac.uk` (Stephen G. Walker)

A variety of shrinkage methods based on classical and Bayesian statistical models in the wavelet domain have been proposed and studied after the works by Donoho and Johnstone (1994), and Donoho et al.(1995a, 1995b). In this broad context of function estimation, Bayesian procedures have proved effective for their capability to incorporate prior information about the unknown signal.

It is now well known that classical wavelet estimators' performance can be improved by grouping wavelet coefficients into blocks. Block wavelet estimation has generated considerable interest in recent years through the articles by Hall et al. (1997, 1999), where local blockwise thresholding procedures with fixed block sizes were introduced and their excellent mean square error (MSE) performance has been reported. Cai (1999) considered block James-Stein rules and investigated the adaptive effect of the block size and the threshold level using an oracle inequality approach. Cai and Zhou (2009) also investigated a data-driven block thresholding procedure which adaptively chose the block size and the threshold level at each resolution level by minimizing Stein's unbiased risk estimate.

Recently, the block methodology has been explored from the Bayesian point of view. In Abramovich et al. (2002), an empirical Bayes approach to incorporate information of neighbouring wavelet coefficients was considered. In their work, wavelet coefficients at each resolution level were grouped in blocks of a given size and the modeling was accomplished by using a mixture of a point mass at zero and a multi–normal distribution. Various Bayesian models have also been proposed since then. Their performance has been reported through simulation results; see, for example, De Canditiis and Vidakovic (2004), where a mixture of two normal-inverse gamma distributions (replicating a point mass at zero and a normal but with priors on the variances, one large and one small) as a prior was used. Also, Wang and Wood (2006) investigated a block shrinkage method based on the sum of squares of wavelet coefficients in the block, where a mixture of a point mass at zero and a non-central chi-squared distribution was used.

The block size and the threshold level play an important role in the performance of a block thresholding estimator. Although the Bayesian block methods mentioned above reported better performance compared with term-by-term thresholding methods and classical block thresholding methods, the choice of the block size was not adaptive. The systematic study of consistency and optimality still remains a challenge. Furthermore, in terms of computational time, most of those existing methods are still quite time consuming.

We propose a data-driven approach to empirically select both the block size and the threshold level at individual resolution levels. At each resolution level, the block size and the threshold level are chosen by minimizing a penalised maximum likelihood estimate (PMLE). The theoretical properties of the proposed estimator are considered here in the Besov sequence space formulation that is by now classical for the analysis of wavelet regression methods. It is shown that the minimizer of PMLE yields an optimal estimate of the wavelet coefficient vector. Simulation results with four different functions at various sample sizes are provided, and their comparisons with other existing methods show that the new method can achieve good performance but with the least computational time.

The paper is organized as follows. A basic review of Besov spaces and standard theoretical properties of the analysis of wavelet regression methods will be given in Section 2. We also introduce our approach: the Bayesian BlockNorm (BBN) procedure in this section. Asymptotic properties of the BBN estimator are presented in Section 3 and its numerical implementation and finite sample performance will be discussed in Section 4. Finally, we conclude with a brief discussion in Section 5.

## 2. The model

### *2.1. Wavelet series and Besov spaces*

The Besov space, $B^{\alpha}_{p,q}$, contains functions having $\alpha$ bounded derivatives in the $L^p$ norm, while the parameter $q$ provides a finer gradation of smoothness. The Besov spaces are a rich class of function spaces and contain many traditional smooth spaces such as the Hölder ($B^{\alpha}_{\infty,\infty}$) and Sobolev ($B^{\alpha}_{2,2}$) spaces. For full details of Besov spaces, please see Triebel (1983) and DeVore and Popov (1988).

For a given $r$-regular mother wavelet $\psi$ with $r > \alpha$, a corresponding father wavelet $\phi$ and a fixed primary resolution level $j_0$, a function $f$ can be expanded as

$$f(t) = \sum_{k=0}^{2^{j_0}} w_{j_0,k}\phi_k(t) + \sum_{j=j_0}^{\infty}\sum_{k=0}^{2^j-1} w_{j,k}\psi_{j,k}(t)$$

and the Besov sequence norm $||\cdot||_{b^{\alpha}_{p,q}}$ of the wavelet coefficients of $f$ is then defined as

$$||f||_{b^{\alpha}_{p,q}} = ||\mathbf{w}_{j_0}||_p + \left(\sum_{j=j_0}^{\infty}(2^{js}||\mathbf{w}_j||_p)^q\right)^{1/q}$$

where $\mathbf{w}_{j_0}$ is the vector of the father wavelet coefficients at the primary resolution level $j_0$, $\mathbf{w}_j$ is the vector of the wavelet coefficients at level $j$, and $s = \alpha + 1/2 - 1/p > 0$. Note that the Besov function norm of index $(\alpha, p, q)$ of a function $f$ is equivalent to the sequence norm of the wavelet coefficients of the function. For this see Meyer (1992).

*2.2. Bayesian BlockNorm estimator*

Now let us suppose that $f \in B^{\alpha}_{p,q}$. Performing the wavelet transform on (1), we have

$$x_{jk} = w_{jk} + n^{-1/2}\sigma z_{jk}, \qquad j \geq j_0, k = 0, \ldots, 2^j - 1, \tag{2}$$

where $j_0$ is fixed, $\{z_{jk}\}$ are independent standard normal random variables and the noise level $\sigma$ is assumed known (we will define this precisely later on). Let $\mathbf{x}$ represent the observation vector and $\mathbf{w}$ be the true signal vector, then the nonparametric problem of estimating $f$ turns into a problem of estimating a high dimensional normal mean vector $\mathbf{w}$ based on the observations $\mathbf{x}$ under certain rules.

For each fixed resolution level $j \geq j_0$, let $L \geq 1$ be the possible length of each block, and $M = 2^j/L$ be the number of blocks. (For simplicity we shall assume that $L$ is a power of 2.) Let $\mathbf{x}_b = (x_{(b-1)L+1}, \ldots, x_{bL})$ represent observations in the $b$-th block on level $j$, and similarly $\mathbf{w}_b = (w_{(b-1)L+1}, \ldots, w_{bL})$ and $\mathbf{z}_b = (z_{(b-1)L+1}, \ldots, z_{bL})$, also $\sigma_n = n^{-1/2}\sigma$, we have

$$\mathbf{x}_b = \mathbf{w}_b + \sigma_n \mathbf{z}_b. \tag{3}$$

We impose the prior on the wavelet coefficients in $b$-th block to be

$$\mathbf{w}_b \sim N(\mathbf{0}, \tau_b^2 I_{L\times L}), \tag{4}$$

where $\tau_b^2$, the noise level for the $b$-th block, will be estimated from the data, as an empirical Bayes exercise.

Subject to the model (3) and prior (4), the posterior distribution of $[\mathbf{w}_b|\mathbf{x}_b]$ follows a normal distribution:

$$[\mathbf{w}_b|\mathbf{x}_b] \sim N\left(\frac{\tau_b^2}{\sigma_n^2 + \tau_b^2}\mathbf{x}_b, \frac{\sigma_n^2\tau_b^2}{(\sigma_n^2 + \tau_b^2)} I_{L\times L}\right).$$

The Bayes rule we will consider here corresponds to the $L^2$-loss and yields the posterior mean

$$\widehat{\mathbf{w}}_b = \mathrm{E}(\mathbf{w}_b|\mathbf{x}_b) = \frac{\tau_b^2}{\sigma_n^2 + \tau_b^2}\mathbf{x}_b. \tag{5}$$

This will be our proposed estimator for $\mathbf{w}_b$ and in the next subsection we show how to derive a natural estimator for $\tau_b$.

*2.3. The penalised adaptive procedure*

According to the model (3) and prior (4), we know, by integration, that the observation $\mathbf{x}_b = (x_{(b-1)L+1}, \ldots, x_{bL})$ is a realization of a multinormal random variable following $N(\mathbf{0}, (\sigma_n^2 + \tau_b^2)I_{L\times L})$ with likelihood function $L(\tau_b, L; \mathbf{x}_b, \sigma_n)$ as follows:

$$L(\tau_b, L; \mathbf{x}_b, \sigma_n) = \left[2\pi(\tau_b^2 + \sigma_n^2)\right]^{-L/2} \exp\left\{-\frac{||\mathbf{x}_b||_2^2}{2(\sigma_n^2 + \tau_b^2)}\right\}. \tag{6}$$

Also, we know through simple calculations, that the maximum likelihood estimator of $\tau_b^2$ is

$$\hat{\tau}_b^2 = \max\left\{0, \frac{||\mathbf{x}_b||_2^2}{L} - \sigma_n^2\right\}. \tag{7}$$

In this way, we can see that $\hat{\tau}_b^2$ provides the "natural" threshold rule. When the average of sum of squares of observations in the same block is less than the noise level, we set all the coefficients in that block to zero.

It is of course necessary to specify the value of the parameter $L$ at each level. The maximum likelihood estimator is usually asymptotically optimal in the parametric context, but it has too many degrees of freedom here. We regularise the maximum likelihood problem by adding constraints. For simplicity, we assume that the best choice of the block size for each resolution level $j$ lies in the set $\{1, 2, \ldots, 2^j\}$. If the probability for each choice is $p_k = \theta^k$, where $0 < \theta < 1$ and $k = 0, \ldots, j$ corresponding to the block set $\{1, 2, \ldots, 2^j\}$, the weighted maximum likelihood function for the whole level $j$ is

$$\theta^k \prod_{b=1}^{M} [2\pi(\tau_b^2 + \sigma_n^2)]^{-L/2} \exp\left\{-\frac{\parallel \mathbf{x}_b \parallel_2^2}{2(\tau_b^2 + \sigma_n^2)}\right\}. \tag{8}$$

We can obtain the optimal block size through minimizing the following penalized log–likelihood function,

$$l^* = \min_L \left\{-\sum_{b=1}^{M} l[\tau_b(L), L; \mathbf{x}_b, \sigma_n] - k\log_2\theta\right\}, \tag{9}$$

where $l[\tau_b(L), L; \mathbf{x}_b, \sigma_n]$ is the log version of the function (6) for the level $j$, and $k = \log_2 L$.

Theoretical results given in Section 3 show that the pair $(L^*, \tau_b^2(L^*))$, $L^*$ is the minimiser of (9), obtained by the above procedure is asymptotically optimal in the sense that the resulting block thresholding estimator adaptively attains the ideal block thresholding risk.

Note that, in contrast to some other block thresholding methods, both the block size and the threshold level of the proposed procedure are chosen empirically and vary from resolution level to level. Both the theoretical and numerical results given in the next two sections will show that the proposed estimator outperforms classical block thresholding estimators and are comparable with Bayes Block estimators with a fixed block size.

## 3. Theoretical properties

In this section we investigate the asymptotic minimax properties of the proposed approach above. Assume $f \in B^{\alpha}_{p,q}$, $\alpha > \max(0, 1/p - 1/2)$ and $p, q \geq 1$. Among all possible estimators $\widehat{f}$, the minimax mean squared error is defined as

$$R(n, B^{\alpha}_{p,q}) = \inf_{\widehat{f}} \sup_{f \in B^{\alpha}_{p,q}} \mathrm{E}||\widehat{f} - f||^2_{L^2[0,1]}.$$

Now let $\psi$ be the mother wavelet of regularity $r > \alpha$. The set of the corresponding wavelet coefficients $w$ of $f$ belongs to a Besov ball of radius $R$, that is $b^{\alpha}_{p,q}(R) = \{w : ||w||_{b^{\alpha}_{p,q}} \leq R\}$. Due to the orthonormality of a wavelet basis, it is that

$$R(n, B^{\alpha}_{p,q}) = \inf_{\widehat{w}} \sup_{w \in b^{\alpha}_{p,q}(R)} \mathrm{E}||\widehat{w} - w||^2_{l^2},$$

where $\widehat{w}$ is the wavelet coefficient of $\widehat{f}$.

**Proposition 1.** *Let a mother wavelet $\psi$ have regularity $r$, with $\max(0, 1/p - 1/2) < \alpha < r$ and $p, q \geq 1$. Then*

$$sup_{w \in b^{\alpha}_{p,q}(R)} E||w - \widehat{w}||^2_2 = \begin{cases} \mathrm{O}(n^{-2\alpha/(2\alpha+1)}) & p \geq 2 \\ \mathrm{O}(n^{-(2\alpha+1-2/p)/(2\alpha+2-2/p)}) & 1 \leq p < 2. \end{cases}$$

The proof is given in the Appendix and follows a similar path to the one given in Abramovich et al. (2004). Our estimator achieves the optimal minimax rate within a small constant factor over any Besov space $B^{\alpha}_{p,q}$, in both cases of $p \geq 2$ and $1 \leq p < 2$.

## 4. Numerical results

The purpose here is to illustrate the practical performance of our proposed approach. In practice the noise level $\sigma$ in (2), which we assumed known for simplicity, needs to be estimated from the data and we will use the following robust estimator of $\sigma$ given in Donoho and Johnstone (1994a). This estimator $\sigma$ is based on the empirical wavelet coefficients at the highest resolution level,

$$\hat{\sigma} = \frac{1}{0.6745}\text{median}\left(|x_{J-1,k}| : 1 \leq k \leq 2^{J-1}\right).$$

Four functions, 'HeaviSine','Blocks', 'Bumps' and 'Doppler', representing different levels of spatial variability, are used as test functions for the purpose of our simulation study. Each test function is rescaled to achieve different signal-to-noise ratios (SNR), and the standard normal noise is added to the functions. The average MSE for the estimator $\hat{f}$ of $f$ is defined as

$$MSE_f = \frac{1}{n}\sum_{i=0}^{n-1}\left[\left\{\hat{f}(i/n) - f(i/n)\right\}^2\right] \tag{10}$$

In order to examine the effect of the proposed method on numerical performance, we performed a preliminary simulation study to examine the parameter $\theta$ introduced in (8). We found that a value of $\theta \in [0.5, 0.8]$ was the reasonable choice for the four test functions with various sample sizes and SNRs. Therefore we take $\theta = 0.6$ for all test functions in this simulation study.

In Table 1, the proposed method BBN is compared with some of the recently proposed methods in the literature: SureBlock (Cai and Zhou 2009), BlockJS (Cai, 1999), BPM (Abramovich et al., 2002) and NCP (Wang and Wood, 2006). BlockJS is a classical block thresholding procedure with a fixed block size $L = \log_2(n)$ and a fixed threshold level. SureBlock is a classical block thresholding procedure which empirically chooses the block size and the threshold level at each resolution level by minimizing Stein's unbiased risk estimation. BPM is a Bayes block threshold method which imposes a mixture of multinormal distribution and a point mass as the prior with block length $L_j = 2^{[\log_2(j/2)]}$ at resolution level $j$. The posterior mean is used here as the shrinkage rule. NCP is a Bayesian block shrinkage approach based on the block sum of squares with a fixed block size. It imposes a mixture of a non-central chi-squared distribution and a point mass. The "power" prior as

**Table 1: Simulation results of five methods (BlockJS, SureBlock, BBN, BPM and NCP) with 100 simulation runs, where AMSE was obtained with SNR=7 and sample sizes N=(256,512,1024 or 2048).**

| Methods | HeaviSine | | | | BlockJS | | | |
|---|---|---|---|---|---|---|---|---|
| | 256 | 512 | 1024 | 2048 | 256 | 512 | 1024 | 2048 |
| BlockJS | 0.371 | 0.197 | 0.125 | 0.074 | 0.990 | 0.707 | 0.473 | 0.289 |
| SureBlock | 0.228∗ | 0.160∗ | 0.087 | 0.056 | 0.583 | 0.481 | 0.334 | 0.251 |
| BBN | 0.250 | 0.166 | 0.096 | 0.063 | 0.407 | 0.310 | 0.285 | 0.213 |
| BPM | 0.269 | 0.189 | 0.120 | 0.079 | 0.268∗ | 0.233 | 0.169 | 0.110 |
| NCP | 0.269 | 0.171 | 0.085∗ | 0.050∗ | 0.312 | 0.232∗ | 0.153∗ | 0.095∗ |
| Mathods | Bumps | | | | Doppler | | | |
| | 256 | 512 | 1024 | 2048 | 256 | 512 | 1024 | 2048 |
| BlockJS | 0.865 | 0.571 | 0.389 | 0.193 | 0.405 | 0.208 | 0.146 | 0.074 |
| SureBlock | 0.636 | 0.487 | 0.378 | 0.237 | 0.359 | 0.304 | 0.174 | 0.110 |
| BBN | 0.575∗ | 0.466 | 0.348 | 0.215 | 0.337∗ | 0.216 | 0.135 | 0.081 |
| BPM | 0.701 | 0.474 | 0.302 | 0.197 | 0.344 | 0.189∗ | 0.139 | 0.095 |
| NCP | 0.677 | 0.426∗ | 0.275∗ | 0.178∗ | 0.351 | 0.195 | 0.125∗ | 0.071∗ |

the distribution of the hyperparameter and posterior mean as the shrinkage rule are used here and the block size is fixed as $L = 2$. Each of these wavelet estimators has been shown to perform well numerically.

The average MSE (AMSE) results with 100 simulation runs for four test functions with SNR=7 at different sample sizes (256, 512, 1024 and 2048) are provided. In Table 1, for each simulation condition, the best AMSE is highlighted with a star. The simulation results show that the proposed method is very competitive with the best of the existing block methods. In three of the four functions, the new method does better than classical methods: BlockJS and SureBlock. Generally, BBN is comparable with Bayesian methods: BPM and NCP.

In Table 2 the average CPU time in seconds as reported by the MATLAB program is shown, where each average value is taken from 100 simulation runs. The "Bumps" signal is tested by four methods (BlockJS, BBN BPM and NCP) at sample sizes 256, 512, 1024 and 2048. As anticipated, the BBN method along with the BlockJS are superior in terms of CPU time to the BPM and NCP methods.

**Table 2: Average CPU time involved in computing 100 simulation runs for the Bumps signal using four methods with SNR=7 and sample sizes N=(256,512,1024, or 2048). Standard errors are given in parentheses.**

| Bumps | 256 | 512 | 1024 | 2048 |
|---|---|---|---|---|
| BlockJS | 0.0189 (.0349) | 0.0266 (.0318) | 0.0372 (.0354) | 0.0483 (.0322) |
| BBN | 0.0080 (.0079) | 0.0136 (.0053) | 0.0242 (.0078) | 0.0453 (.0047) |
| BPM | 1.5241 (.1455) | 2.3287 (.1734) | 8.1894 (.8263) | 29.307 (1.131) |
| NCP | 0.7925 (.0312) | 1.6948 (.0803) | 3.7175 (.1276) | 7.8644 (.2522) |

## 5. Conclusion

We have studied the theoretical properties of a proposed Bayesian block wavelet shrinkage method, where the prior distribution imposed on wavelet coefficients is a multinormal distribution and parameters are estimated via empirical Bayes ideas. This provides us with a natural threshold level without a usual mixture model and an explicit estimator for the wavelet coefficients, which makes it easy to implement and fast to compute with alternative Bayesian block methods.

Comparisons with other existing methods using simulation studies of four different test functions at various sample sizes show that the new method is comparable (similar or better performance) with the existing methods. An important advantage of this new method is that its computational cost in terms of CPU time is the least among all the compared methods.

## A. Appendix

Without loss of generality we assume that $\sigma_n^2 = 1/n$ in the model (2). For each fixed resolution level $j$, we spilt the wavelet coefficients into $M_j$ non-overlapping blocks each with $L_j$ coefficients, so $M_j \times L_j = 2^j$. The parameter $\tau_{jl}^2$ for the $l-$th block at level $j$ will be estimated by the coefficients within the block, which is $\widehat{\tau}_{jl}^2 = \max\{0, ||\mathbf{x}_{jl}||_2^2/L_j - \sigma_n^2\}$. Then we have $\widehat{\mathbf{w}}_{jl} = E(\mathbf{w}_{jl}|\mathbf{x}_{jl}) = b_{jl}\mathbf{x}_{jl}$, where $b_{jl} = \widehat{\tau}_{jl}^2/(\widehat{\tau}_{jl}^2 + \sigma_n^2)$ according to (5) for $j \geq 0$ and $l = 1, \ldots, M_j$.

For any sequence of wavelet coefficients $\{w_{jk}\} \in b_{p,q}^{\alpha}(R)$, we consider

$$\begin{aligned} \text{MSE} &= \sum_{j=0}^{\infty}\sum_{l=1}^{M_j} \mathrm{E}||b_{jl}\mathbf{x}_{jl} - \mathbf{w}_{jl}||_2^2 = \sum_{j=0}^{\infty}\sum_{l=1}^{M_j} \mathrm{E}||b_{jl}(\mathbf{x}_{jl} - \mathbf{w}_{jl}) - (1-b_{jl})\mathbf{w}_{jl}||_2^2 \\ &\leq \sum_{j=0}^{\infty}\sum_{l=1}^{M_j} \{\mathrm{E}||b_{jl}(\mathbf{x}_{jl} - \mathbf{w}_{jl})||_2^2 + \mathrm{E}||(1-b_{jl})\mathbf{w}_{jl}||_2^2\} := A + B. \end{aligned} \tag{11}$$

For $b_{jl}$, we have

$$b_{jl} = \frac{\max\left\{0, \frac{||\mathbf{x}_{jl}||_2^2}{L_j} - \sigma_n^2\right\}}{\max\left\{\sigma_n^2, \frac{||\mathbf{x}_{jl}||_2^2}{L_j}\right\}} = \max\left\{0, 1 - \frac{L_j\sigma_n^2}{||\mathbf{x}_{jl}||_2^2}\right\}.$$

Let $J_s = s^{-1}\log_2 n$ be a fixed resolution level. Then using the fact that $1 - L_j\sigma_n^2/||\mathbf{x}_{jl}||_2^2$ is a concave function of $||\mathbf{x}_{jl}||_2^2$, we have the following inequalities:

$$E(b_{jl}) \leq \begin{cases} ||\mathbf{w}_{jl}||_2^2/L_j\sigma_n^2 & j > J_s, \\ ||\mathbf{w}_{jl}||_2^2/||\mathbf{w}_{jl}||_2^2 \quad = 1 & j \leq J_s. \end{cases}$$

Also, for $j > J_s$,

$$\sum_{l=1}^{M_j} \mathrm{E}(b_{jl}) \leq \frac{\sum_{l=1}^{M_j} ||\mathbf{w}_{jl}||_2^2}{L_j\sigma_n^2} = \frac{\sum_{j=0}^{2^j-1} w_{jk}^2}{L_j\sigma_n^2}.$$

Let $\alpha' = \alpha$ for $p > 2$ and $\alpha' = \alpha + 1/2 - 1/p$ otherwise. For sequences from Besov balls, we have that $\sum_{j=0}^{2^j-1} w_{jk}^2 \leq C2^{-2j\alpha'}$. For further information, see Johnstone (2002).

Hence, the first term $A$ of (11) is

$$A = \sum_{j=0}^{J_s}\sum_{l=1}^{M_j}\{\mathrm{E}||b_{jl}(\mathbf{x}_{jl} - \mathbf{w}_{jl})||_2^2 + \sum_{j=J_s+1}^{\infty}\sum_{l=1}^{M_j}\{\mathrm{E}||b_{jl}(\mathbf{x}_{jl} - \mathbf{w}_{jl})||_2^2,$$

with

$$\begin{aligned} &\sum_{l=1}^{M_j} \mathrm{E}||b_{jl}(\mathbf{x}_{jl} - \mathbf{w}_{jl})||_2^2 \leq \sum_{k=0}^{2^j-1} \sqrt{\mathrm{E}b_{jl(k)}^4}\sqrt{\mathrm{E}(x_{jk} - w_{jk})^4} \\ &= C\sigma_n^2 L_j \sum_{l=1}^{M_j} [\mathrm{E}b_{jl}^4]^{1/2} \leq C\sigma_n^2 L_j \left(M_j \sum_{l=1}^{M_j} \mathrm{E}b_{jl}^4\right)^{1/2} \leq C \cdot \sigma_n^2 L_j \left(M_j \sum_{l=1}^{M_j} \mathrm{E}b_{jl}\right)^{1/2}. \end{aligned}$$

We have

$$A \leq \mathcal{O}(n^{-(s-1)/s}) + \mathcal{O}(n^{-(s+\alpha-1/2)/s})$$

for $p > 2$ and $A \leq \mathcal{O}(n^{-(s-1)/s}) + \mathcal{O}(n^{-(s+\alpha-1/p)/s})$ for $1 \leq p < 2$.

Now consider the second term $B$ of (11). Let

$$1 - b_{jl} = \frac{\sigma_n^2}{\sigma_n^2 + \widehat{\tau}_{jl}^2} = \frac{1}{1 + \max\left\{0, ||\mathbf{x}_{jl}||_2^2/(L_j\sigma_n^2) - 1\right\}}.$$

Using the fixed resolution level $J_s = s^{-1}\log_2 n$, we have

$$1 - b_{jl} = \frac{1}{\max\{1, ||\mathbf{x}_{jl}||_2^2/L_j\sigma_n^2\}} \leq \begin{cases} 1 & j > J_s \\ L_j\sigma_n^2/||\mathbf{x}_{jl}||_2^2 & j \leq J_s. \end{cases}$$

Hence

$$\begin{aligned} B &= \sum_{j=0}^{\infty}\sum_{l=1}^{M_j} \mathrm{E}||(1-b_{jl})\mathbf{w}_{jl}||_2^2 \leq \sum_{j=0}^{\infty}\sum_{l=1}^{M_j} ||\mathbf{w}_{jl}||_2^2 \cdot \mathrm{E}(1-b_{jl})^2 \\ &\leq \sum_{j=0}^{J_s}\sum_{l=1}^{M_j} ||\mathbf{w}_{jl}||_2^2 \cdot \mathrm{E}(1-b_{jl}) + \sum_{J_{s+1}}^{\infty}\sum_{l=1}^{M_j} ||\mathbf{w}_{jl}||_2^2 := B_1 + B_2. \end{aligned}$$

Similarly, let $\alpha' = \alpha$ if $p > 2$ and $\alpha' = \alpha + 1/2 - 1/p$ otherwise. Then

$$B_2 = \sum_{J_{s+1}}^{\infty}\sum_{l=1}^{M_j} ||\mathbf{w}_{jl}||_2^2 = \sum_{J_{s+1}}^{\infty}\sum_{k=0}^{2^j-1} w_{jk}^2 \leq \mathcal{O}(n^{-2\alpha/s})$$

for $p > 2$ and $B_2 \leq \mathcal{O}(n^{-(2\alpha+1-2/p)/s})$ otherwise.

Now consider $B_1$. From the result (e.g. Moser, 2007, Theorem 4), the expectation of the reciprocal value of a non-central chi-square random variable $Y$ with an even number $2m$ ($m > 1$) of degrees of freedom is $\mathrm{E}(Y^{-1}) = g'_{m-1}(\xi)$, where $\xi$ is the non-centrality and

$$g'_m(\xi) = e^{-\xi}\sum_{k=0}^{\infty}\frac{1}{k!}\cdot\frac{1}{k+m}\xi^k,$$

which can be bounded by

$$\frac{1}{\xi+m} \leq g'_m(\xi) \leq \min\left\{\frac{m+1}{m(\xi+m+1)}, \frac{1}{\xi+m-1}\right\}.$$

Hence, for $L_j = 2m$ and $m > 1$, we have

$$\begin{aligned}
\mathrm{E}(1-b_{jl}) &\leq L_j \min\left\{\frac{L_j/2+1}{L_j/2(||\mathbf{w}_{jl}||_2^2/\sigma_n^2 + L_j/2+1)}, \frac{1}{||\mathbf{w}_{jl}||_2^2/\sigma_n^2 + L_j/2-1}\right\} \\
&\leq \frac{L_j+2}{||\mathbf{w}_{jl}||_2^2/\sigma_n^2 + L_j/2+1} \leq \frac{\sigma_n^2(L_j+2)}{||\mathbf{w}_{jl}||_2^2}.
\end{aligned}$$

We know that the optimal block size increases with the increase of the resolution level (also shown by the simulations) and, for simplicity, we assume that the best choice of block size for each resolution level $j$ lies in the set $\{1, 2, \ldots, 2^j\}$. Without losing the generality, we can find a $J_1 < J_s$ for all large $n$, so that the chosen block size is less than 4 only when $j \leq J_1$, and a $C^*$ so that $C^* \leq ||\mathbf{w}_{jl}||_2^2$ for the resolution levels $j < J_1$ and $l = 1, \ldots, M_j$ with $||\mathbf{x}_{jl}||_2^2 \neq 0$.

Hence

$$\begin{aligned}
B_2 &\leq \sum_{j=0}^{J_1}\sum_{l=1}^{M_j} ||\mathbf{w}_{jl}||_2^2 \cdot \frac{L_j\sigma_n^2}{C^*} + \sum_{j=J_1+1}^{J_s}\sum_{l=1}^{M_j} ||\mathbf{w}_{jl}||_2^2 \cdot \frac{\sigma_n^2(L_j+2)}{||\mathbf{w}_{jl}||_2^2} \\
&\leq \mathcal{O}(n^{-1}) + \mathcal{O}(n^{-(s-1)/s}).
\end{aligned}$$

If we choose $s = 2\alpha + 1$ for $p > 2$ and $s = 2\alpha + 2 - 2/p$ otherwise, the proof is completed. $\diamond$

## References


[1] Abramovich, F., Besbeas, P. and Sapatinas, T. (2002). Empirical Bayes approach to block wavelet function estimation. *Computational Statistics and Data Analysis,* **39**, 435–451.

[2] Abramovich, F., Amato, U. and Angelini, C. (2004). Optimality of Bayesian wavelet estimators. *Scandinavian Journal of Statistics,* **31**, 217-234.

[3] Cai, T. T. (1999). Adaptive wavelet estimation: a block thresholding and oracle inequality approach. *Ann. Statist.,* **27**, 898–924.

[4] Cai, T.T. and Zhou, H.H. (2009). A data-driven block thresholding approach to wavelet estimation. *The Annals of Statistics*, **37**, 569–595.

[5] De Canditiis, D. and Vidakovic, B. (2004). Wavelet Bayesian block shrinkage via mixtures of normal-inverse gamma priors. *J. Comp. Graph. Statist.*, **13**, 382–398.

[6] DeVore, R. and Popov, V. (1988). Interpolation of Besov spaces. *Trans. Amer. Math. Soc.* **305**, 397-414.

[7] Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage.*Biometrika*, **81**, 425-455.

[8] Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D (1995a). Wavelet shrinkage: asymptopia (with discussion)? *J. R. Statist. Soc. B*, **57**, 301-365.

[9] Donoho, D. L. and Johnstone, I. M. (1995b). Adapting to unknown smoothness via wavelet shrinkage. *J. Am. Statist. Ass.*, **90**, 1200-1224.

[10] Donoho, D. L. and Johnstone, I. M. (1998). Minimax estimation via wavelet shrinkage. *Ann. Statist.*, **26**, 879-921.

[11] Hall, P., Penev, S., Kerkyacharian, G. and Picard, D. (1997). Numerical performance of block thresholded wavelet estimators. *Statistical Computing*, **7**, 115-124.

[12] Hall, P., Kerkyacharian, G. and Picard, D. (1999). On the minimax optimality of block thresholded wavelet estimator. *Statistica Sinica*, **9**, 33-50.

[13] Johnstone I. (2002). *Function Estimation and Gaussian Sequence Models.* unpublished monograph, http://www-stat.stanford.edu/ imj.

[14] Meyer, Y. (1992). *Wavelets and Operators.* Cambridge University Press, Cambridge.

[15] Moder, S. M. (2007). Some expectations of a non-central chi-square distribution with an even number of degrees of freedom. *TENCON 2007-2007 IEEE Region 10 Conference.*

[16] Triebel, H. (1983). *Theory of Function Spaces.* Birkhäuser Verlag, Basel.

[17] Wang, X. and Wood, A. T. A. (2006). Empirical Bayes block shrinkage of wavelet coefficients via the noncentral $\chi^2$ distribution. **93**, 705-722.