Adaptive Structured Block Sparsity Via Dyadic Partitioning

Gabriel Peyré, Jalal M. Fadili, Christophe Chesneau

ADAPTIVE STRUCTURED BLOCK SPARSITY VIA DYADIC PARTITIONING

Gabriel Peyré[1], *Jalal Fadili*[2] *and Christophe Chesneau*[3]

[1]Ceremade
CNRS-Univ. Paris-Dauphine, France
Gabriel.Peyre@ceremade.dauphine.fr

[2]GREYC
CNRS-ENSICAEN-Univ. Caen, France
Jalal.Fadili@greyc.ensicaen.fr

[3]LMNO
CNRS-Univ. Caen, France
Chesneau.Christophe@math.unicaen.fr

ABSTRACT

This paper proposes a novel method to adapt the block-sparsity structure to the observed noisy data. Towards this goal, the Stein risk estimator framework is exploited, and the block-sparsity is dyadically organized in a tree. The adaptation of the sparsity structure is obtained by finding the best recursive dyadic partition, whose terminal nodes (leaves) are the blocks, that minimizes a data-driven estimator of the risk. Our main contributions are (i) analytical expression of the risk; (ii) a novel estimator of the risk; (iii) a fast algorithm that yields the best partition. Numerical results on wavelet-domain denoising of synthetic and natural images illustrate the improvement brought by our adaptive approach.

1. INTRODUCTION

In the last decade or so, sparsity has emerged as one of the attractive theoretical and practical signal properties in a wide range of signal processing applications. The interest in sparsity has arisen owing to the new sampling theory, *compressed sensing*, which revitalizes the vision of to the well-known Shannon sampling theory.

In multiscale decompositions of natural images, it has been widely observed that large detail coefficients are not only sparse but also tend to cluster into *blocks* (or groups) along the singularities. It can then be safely anticipated that exploiting this typical structure of the sparsity as a prior would be beneficial in many subsequent processing steps, for instance in restoration and inverse problems.

Block- sparsity and estimation A now classical approach to denoising [14, 13] operates by individual thresholding/shrinkage of the coefficients in a suitable basis (e.g. wavelets), where the signal/image is sparse. However, the individual thresholding achieves a degree of trade-off between variance and bias contribution to the quadratic risk which is not optimal. One way to increase estimation precision is by exploiting information about neighboring coefficients. In other words, the coefficients could be thresholded in blocks rather than individually. It has been shown that such a procedure, with an appropriate fixed block size, achieves the optimal minimax convergence rates over several functional spaces (e.g. Besov) [4, 5, 8]. From a practical point view, the results of [8, 7, 21] clearly show the notable improvement brought by block thresholding. In [7] the authors propose a multi-channel block denoising algorithm in the wavelet domain. [21] advocate the use of block denoising for audio signals with anisotropic blocks. In these works, and following [6], the denoiser hyperparameters (threshold, block size) are optimized by minimizing the Stein Unbiased Risk Estimator (SURE). Block thresholding with fixed block size has been extended to deconvolution in [9]. Moreover, there has been a recent wave of interest in convex block-sparsity regularization in the statistical and machine learning literature (e.g. [22, 1]), as well as in the compressed sensing community; e.g. [2]. In all these works, the block-sparsity structure is supposedly known in advance.

Dyadic partitioning The CART methodology of tree-structured adaptive non-parametric regression has been widely used in statistical data analysis since its inception nearly two decades ago [3]. It is built around ideas of recursive partitioning. There are several variants of CART, depending on the procedure used to construct the partition. In this work, we are interested in optimal (non-greedy) dyadic recursive partitioning. That is, starting from the trivial partition containing the whole domain, partitions are built by splitting it into two pieces vertically and horizontally. The same splitting is applied recursively to each ancestor piece, generating the children partitions. A recursive dyadic partition is any partition reachable by successive application of these rules. While the original CART can split rectangles in all proportions, the dyadic CART we develop here, can split rectangles only in half (i.e. squares). This allows a more limited range of possible partitions, which makes it possible to find an optimal partition in linear time. There are close connections between the dyadic CART and best-basis algorithms in a tree-structured dictionary [12, 10]. There is also a mathematical connection between the areas, where the main results driving them assert that, given an additive cost function, which assigns numerical values to a 2^d-tree and its subtrees, the optimal subtree minimizing the cost function is obtained by dynamic programming/bottom-up pruning of the complete tree. This link has been capitalized on in [16] for denoising, for non-linear approximation of geometrical images in both the spatial [11, 19], and wavelet domains [15, 20, 17]. Adaptivity offered by the segmentation adaptation is crucial to improve over classical non-adaptive approximation (e.g. wavelets).

Learning the block structure. In a recent work, [18] propose to learn the sparsity structure and the sparsifying dictionary from a set of exemplars. However, the solution to the marginal optimization problem with respect to the spar-

sity structure is obviously NP-hard, the authors approach it, without any guarantee, using an agglomerative clustering algorithm.

2. BLOCK NON-LINEAR ESTIMATORS AND RISKS

We consider a simple denoising scenario, where one observes $y = x + w \in \mathbb{R}^N$, with x the unknown clean signal to recover, and w is an additive white Gaussian noise of variance σ^2. The prior on x is that it exhibits a block-sparsity pattern.

2.1 Block Estimators

A block segmentation B corresponds to a disjoint union of the indices set

$$\{0,\dots,N-1\} = \bigcup_{b\in B} b, \quad b\cap b' = \emptyset,\ \forall b \neq b' .$$

A block-based thresholding estimator of x is defined as $\widehat{x} = S_{\lambda,B}(y)$, where for each block $b \in B$

$$\forall i \in b, \quad \widehat{x}[i] = \rho_\lambda(\|y_b\|)y[i], \quad \text{where} \quad \|y_b\|^2 = \frac{1}{|b|}\sum_{i\in b}|y[i]|^2 , \tag{1}$$

$|b|$ is the cardinality of the block, and $y_b \in \mathbb{R}^{|b|}$ is the vector of observations within block b. In words, this estimator operates by jointly thresholding all coefficients within a block if the block signal-to-noise ratio is below a threshold parameterized by $\lambda \geqslant 0$. Otherwise, a common attenuation rule is applied to the coefficients.

In this paper, we consider two popular thresholding estimators; namely soft and James-Stein (JS) thresholdings defined respectively as[1]

$$\rho_\lambda^{\text{Soft}}(a) = \max\left(0, 1-\frac{\lambda}{\|a\|}\right), \quad \rho_\lambda^{\text{JS}}(a) = \max\left(0, 1-\frac{\lambda^2}{\|a\|^2}\right). \tag{2}$$

One can show that these estimators are the unique MAP estimators associated to closed-form penalties that indeed promote block-sparsity. We omit this result for obvious space limitations. Note that unlike soft thresholding, which has a constant bias, the JS rule entails less bias that decreases as the block energy increases.

2.2 Estimators risks

SURE The quadratic risk associated to a denoiser S is $\mathbb{E}_w(\|S(y)-x\|^2)$ where the expectation is taken with respect to the noise distribution. The best denoiser (in ℓ_2 sense) should be the one that minimizes this risk. However, without access to an oracle that provides some information on x, minimizing this risk seems out of reach.

Under weak differentiability of $y \in \mathbb{R}^n \mapsto S(y)$, Stein lemma allows to get an unbiased estimator of the risk, coined SURE, which solely depends on the observation y. Its general expression reads

$$J(y,S) = n\sigma^2 + \|y-S(y)\|^2 + 2\sigma^2 \mathrm{div}(S-\mathrm{Id})(y) , \tag{3}$$

where n is the dimension of x and the divergence of a multi-valued mapping $f = (f_i)_i : \mathbb{R}^n \to \mathbb{R}^n$ is $\mathrm{div}(f)(x) = \sum_i \frac{\partial f_i}{\partial x_i}(x)$. The SURE is an unbiased estimator of the risk since

$$\mathbb{E}_w(\|S(y)-x\|^2) = \mathbb{E}_w(J(y,S)).$$

It is then possible to use (3), computed from a single realization of y, as a risk estimator, and to minimize it in order to adapt the hyperparameters (e.g. threshold λ, block size $|b|$) of the denoiser S. In this paper, we propose this framework to infer both the optimal block-sparsity structure and the threshold λ.

SURE on blocks Applying the SURE formula (3), the SURE on each block $b \in B$ corresponding to the estimators $y_b \mapsto \rho_\lambda(\|y_b\|)y_b$ defined in (1)-(2) reads

$$\begin{aligned} J^{\text{Soft}}(y_b,\lambda,\sigma) &= |b|\sigma^2 + \Big(|b|\|y_b\|^2 - 2|b|\sigma^2\Big) I(\|y_b\| < \lambda) \\ &\quad + \left(|b|\lambda^2 - 2\sigma^2(|b|-1)\frac{\lambda}{\|y_b\|}\right) I(\|y_b\| \geqslant \lambda) \qquad (4) \\ J^{\text{JS}}(y_b,\lambda,\sigma) &= |b|\sigma^2 + \Big(|b|\|y_b\|^2 - 2|b|\sigma^2\Big) I(\|y_b\| < \lambda) \\ &\quad + \frac{|b|\lambda^2 - 2\sigma^2(|b|-2)}{\|y_b\|^2/\lambda^2} I(\|y_b\| \geqslant \lambda) , \qquad (5) \end{aligned}$$

where $I(\omega)$ is the indicator of the event ω. Let's point out that $J^{\cdot}(y_b,\lambda,\sigma) = \sigma^2 J^{\cdot}(y_b/\sigma,\lambda/\sigma,1)$, and $|b|\|y_b\|^2/\sigma^2 \sim \chi^2_{|b|}(|b|\|x_b\|^2/\sigma^2)$ a non-central chi-square distribution with $|b|$ degrees of freedom and non-centrality parameter $|b|\|x_b\|^2/\sigma^2$. This simple observation and some algebra yield the following result.

Proposition 1. *Assume $\sigma = 1$. Let $L = |b|$, $\mu = L\|x_b\|^2$ and $\lambda_b = \lambda\sqrt{L}$. For all $L > 2$,*

$$\begin{aligned} \mathbb{E}_w(J^{Soft}(y_b,\lambda,1)) &= \mu + \sum_{k\in\mathbb{N}} \tfrac{e^{-\mu/2}(\mu/2)^k}{k!}\Big(-\sqrt{2}\lambda_b(L-1) \\ &\quad \tfrac{\Gamma\left(\frac{L-1}{2}+k\right)}{\Gamma\left(\frac{L}{2}+k\right)} P_{L+2k-2} - (L+2k)P_{L+2k+2} + L(\lambda_b^2+2)P_{L+2k}\Big) \qquad (6) \end{aligned}$$

$$\begin{aligned} \mathbb{E}_w(J^{JS}(y_b,\lambda,1)) &= \mu + \sum_{k\in\mathbb{N}} \tfrac{e^{-\mu/2}(\mu/2)^k}{k!}\Big(\tfrac{\lambda_b^2(4-L(\lambda_b^2-2))}{L+2k-2} \\ &\quad P_{L+2k-2} - (L+2k)P_{L+2k+2} + 2LP_{L+2k}\Big) \qquad (7) \end{aligned}$$

where $P_\nu = \Pr(\chi^2_\nu > \lambda_b^2) = \frac{\Gamma(\nu/2,\lambda_b^2)}{\Gamma(\nu/2)}$.

These expressions of the risk only depend on μ, and not x_b.

A new risk estimator Although the SURE (with a single realization of y) is an unbiased estimator of the risk, it may exhibit a non-negligible variance. It can be shown that this is the case for the SURE associated with the block Soft and JS estimators. Consequently, inferring the optimal partition based on the SURE turns to be quite inaccurate. In order to

[1] Other thresholding rules could be considered as well, with the proviso that the mapping $a \mapsto \rho_\lambda(a)$ be weakly differentiable for the Stein lemma to apply, see hereafter.

circumvent this difficulty, we propose to use Monte-Carlo integration from expressions (4)-(5)[2]. Indeed, a potentially biased, but with a lower variance, of $\mathbb{E}_w(J^{\mathrm{JS}}(y_b,\lambda,\sigma))$ is given by

$$\widehat{J}^{\mathrm{JS}}(y_b,\lambda,\sigma)=\frac{1}{K}\sum_{i=1}^{K}\Big(|b|\sigma^2+(z_i-2|b|\sigma^2)I(z_i<\lambda^2|b|) + \frac{\lambda^2|b|(\lambda^2|b|-2\sigma^2(|b|-2))}{z_i}I(z_i\geqslant\lambda^2|b|)\Big)\,, \quad (8)$$

where z_i are realizations of a random variable (RV) $\sigma^2\chi^2_{|b|}(\hat{\mu})$, $\hat{\mu}=|b|\max(0,\|y_b\|^2/\sigma^2-1)$. A similar formula to (8) can be written for block soft thresholding (4). The bias in this risk estimator originates from the estimator $\hat{\mu}$ of the non-centrality parameter. In our experiments, $K=100$ was sufficient to get good results. In order to alleviate any ambiguity in the sequel, the terminology SURE will be avoided, and our risk estimator will be dubbed Stein risk.

Now, given some block structure B, the overall associated risk is hence

$$\widehat{J}_B(y,\lambda,\sigma)=\sum_{b\in B}\widehat{J}(y_b,\lambda,\sigma)\,, \quad (9)$$

where $\widehat{J}$ is either $\widehat{J}^{\mathrm{Soft}}$ or $\widehat{J}^{\mathrm{JS}}$.

3. BLOCK SPARSITY BY DYADIC PARTITIONING

Let's recall that we seek $(\lambda^\star,B^\star)$ the global minimizer of $\min_{\lambda,B}\mathbb{E}_w(\|S_{B,\lambda}(y)-x\|^2)$. Without knowledge of x, this solution is approached by the global minimizer of the Stein risk (8). This leads to an intractable combinatorial problem if one does not assume additional structure on the set of allowable partitions. To obtain a fast algorithm, inspired by the CART methodology, we assume that the blocks are organized in a recursive hierarchical structure, obtained by an iterative dyadic subdivision of the blocks.

3.1 Quadtree Partitioning

For simplicity, we detail in this section the 2D case of quadtree subdivisions, which is useful for image processing applications. Note however that our approach is general and extends to any dimension and can handle arbitrary subdivision schemes (possibly non regular and non-stationary). In the following, the set $\{0,\ldots,N-1\}$ indexes 2D pixels, and the blocks b corresponds to squares grouping a subset of the pixels.

A block partition B is associated to a quadtree T, which is a subtree of the whole quadtree containing all possible partitions. The blocks of the partition are the leaves $\mathscr{L}(T)$ of T. The latter lists a set of blocks $b_{j,i}$ indexed by their depth $0\leqslant j\leqslant J=\log_2(N)/2$, and their position $0\leqslant i<4^j$. Fig. 1shows an example of a quadtree T (right) and the associated dyadic partition (left), where each block $b_{j,i}$ corresponds to a square regrouping $N/4^j$ pixels. The block are defined iteratively by splitting $b_{j,i}$ into four regular sub-squares $(b_{j+1,4i},b_{j+1,4i+1},b_{j+1,4i+2},b_{j+1,4i+3})$ of equal size.

[2]Or alternatively (6)-(7) by generating realizations k of a Poisson RV of intensity $\hat{\mu}$.

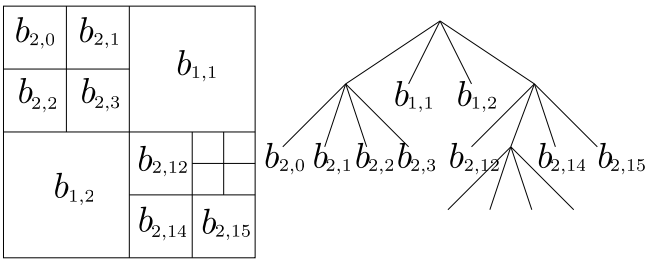


Figure 1: *Example of a quadtree B (right) and the associated dyadic partition (left).*

3.2 Fast Block Partitioning

The goal is to minimize the Stein risk (8) of a block estimator with respect to the dyadic partition and λ

$$\min_{T,\lambda}\sum_{b\in\mathscr{L}(T)}\widehat{J}(y_b,\lambda,\sigma).$$

As λ is a scalar, its value can be optimized by any dichotomic-like (e.g. Golden Section) search algorithm over the cost function marginally minimized with respect to T. Given that this is an additive cost function over the leaves of T, minimizing it with respect to T (or equivalently B) for fixed λ is achieved using the CART dynamic programming algorithm detailed next.

Step 1 **Compute the risk on each block.** For each possible dyadic block $b_{j,i}$, compute $J_{j,i}=\widehat{J}(y_{b_{j,i}},\lambda,\sigma)$.

Step 2 **Best blocks selection.** A bottom-up recursive pruning of the complete tree. For each $j=J-1\ldots,0$ and $0\leqslant i<4^j$,

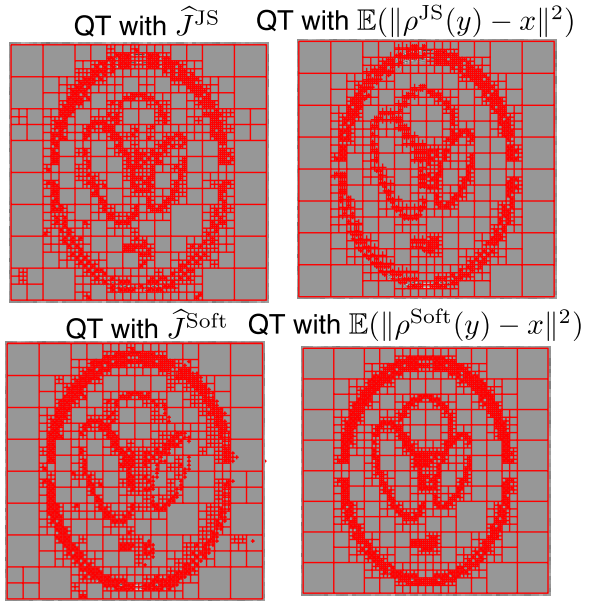
$$\tilde{J}_{j,i}=\min(J_{j,i},J^c_{j,i})\quad\text{where}\quad J^c_{j,i}=\sum_{\eta=0}^{3}\tilde{J}_{j+1,4i+\eta}.$$

If $\tilde{J}_{j,i}=J_{j,i}$, then the node (j,i) is declared as a leaf of the tree T. Otherwise (j,i) is an interior node.

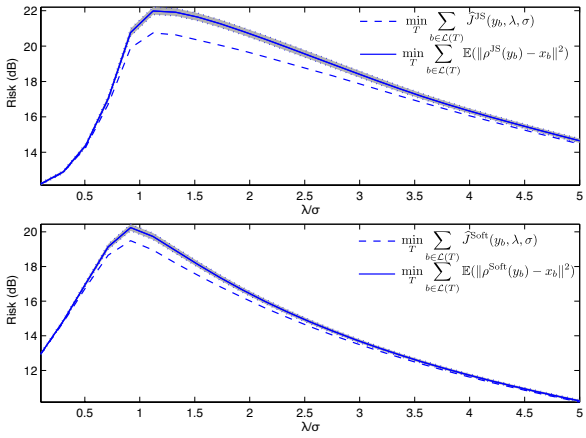
The complexity of this algorithm is dominated by that of computing (8) on all blocks which costs $O(KN)$ operations, and the number of comparisons in the dyadic CART which is $O(N)$.

4. NUMERICAL EXAMPLES

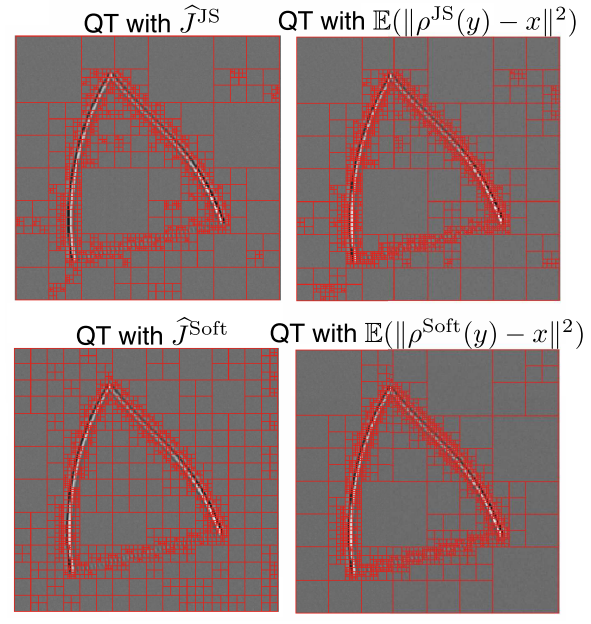
Synthetic images Fig. 2(a) shows an example of a dyadic partition obtained on a subband of the wavelet coefficients of two noisy geometrical images, here the Shepp-Logan phantom and a triangle, with PSNR=20dB. The displayed partitions are those minimizing the exact risk (knowing x) and our Stein risk estimator, with each of the block thresholding rules Soft and JS. One can clearly see that the optimal block structure provided by our estimator is very close to the oracle one. Both partitions are adaptively refined to match the geometry of the singularities as expected. Fig. 2(b) depicts the exact and estimated Stein risks (dB) at the optimal partition for each thresholding rule as a function of λ/σ. The curves are indeed unimodal and have a similar behavior. In spite of the anticipated but slight bias in the estimated Stein risk, the values of λ optimizing both risks coincide.



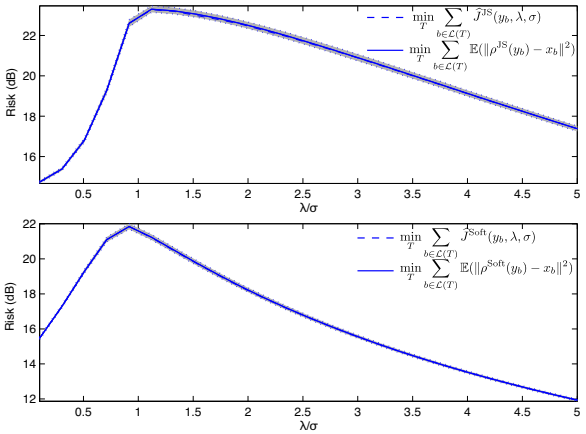
(a)



(b)



(c)



(d)

Figure 2: *(a)-(c): Dyadic partitions (QT) of a wavelet subband of two images by minimizing the exact risk and our Stein risk estimator. (b)-(d): Exact and Stein risks (dB) vs λ/σ. The shaded area corresponds to the confidence interval as the exact risk is estimated with the empirical mean over 100 realizations.*

TI wavelet-domain denoising Fig. 3 depicts some denoising results in the translation-invariant (TI) wavelet domain. For block-thresholding with a fixed block size, the value of λ was optimized by minimizing the SURE [6]. As the JS rule entails less denoising bias, it is systematically better than the Soft rule. Compared to fixed-block denoising, adapting the block sparsity structure enhances notably the denoising performance both visually and quantitatively.

5. CONCLUSION

We have proposed a framework to adapt the block-sparsity structure from noisy data. Its potential application has been illustrated on a denoising. The extension of this approach to inverse problems as well as some open theoretical questions concerning the properties of our risk estimator are currently under investigation.

REFERENCES

[1] F. R. Bach. Consistency of the group lasso and multiple kernel learning. *J. Mach. Learn. Res.*, 9:1179–1225, June 2008.

[2] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde. Model-based compressive sensing. *IEEE Trans. Inf. Theo.*, 56:1982–2001, 2010.

[3] L. Breiman, J. H. Friedman, R. A. Olshen, , and C. J. Stone. *Classification and Regression Trees*. Wadsworth, Belmont, CA, 1984.

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

[5] T. Cai and B.W. Silverman. Incorporate Information on Neighboring Coefficients into Wavelet Estimation. *Sankhya*, 63:127–148, 2001.

[6] T. Cai and H. Zhou. A data-driven block thresholding

approach to wavelet estimation. *Annals of Statistics*, 37(2):569—595, 2009.

[7] C. Chaux, L. Duval, A. Benazza-Benyahia, and J.-C. Pesquet. A nonlinear stein based estimator for multichannel image denoising. *IEEE Trans. on Sig.*, 56(8):3855–3870, 2008.

[8] C. Chesneau, M.J. Fadili, and J.-L. Starck. Stein block thresholding for image denoising. *App. Comp. Harm. Anal.*, 28:67–88, 2009.

[9] C. Chesneau, M.J. Fadili, and J.-L. Starck. Stein block thresholding for wavelet-based image deconvolution. *Elec. J. Statistics*, 4:415–435, 2010.

[10] R. Coifman and V. Wickerhauser. Entropy-based algorithms for best basis selection. *IEEE Trans. Inf. Theo.*, IT–38(2):713–718, Mar. 1992.

[11] D. Donoho. Wedgelets: Nearly minimax estimation of edges. *Annals of Stat.*, 27(3):859–897, 1999.

[12] D. L. Donoho. Cart and best-ortho-basis: A connection. *Annals of Stat.*, 25(5):1870–1911, 1997.

[13] D. L. Donoho and I. Johnstone. Minimax estimation by wavelet shrinkage. *Annals of Stat.*, 26:879–921, 1998.

[14] D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, 81:425–455, 1994.

[15] P.L. Dragotti and M. Vetterli. Wavelet footprints: Theory, algorithms and applications. *IEEE Trans. Signal Proc.*, 51(5):1306–1323, Mat 2003.

[16] H. Krim, D. Tucker, S. Mallat, and D. Donoho. On denoising and best signal representation. *IEEE Trans. Info. Theory*, 45(7):319–335, Nov 1999.

[17] S. Mallat and G. Peyré. Orthogonal bandlet bases for geometric images approximation. *Commun. on Pure and Appl. Math.*, 61(9):1173–1212, 2008.

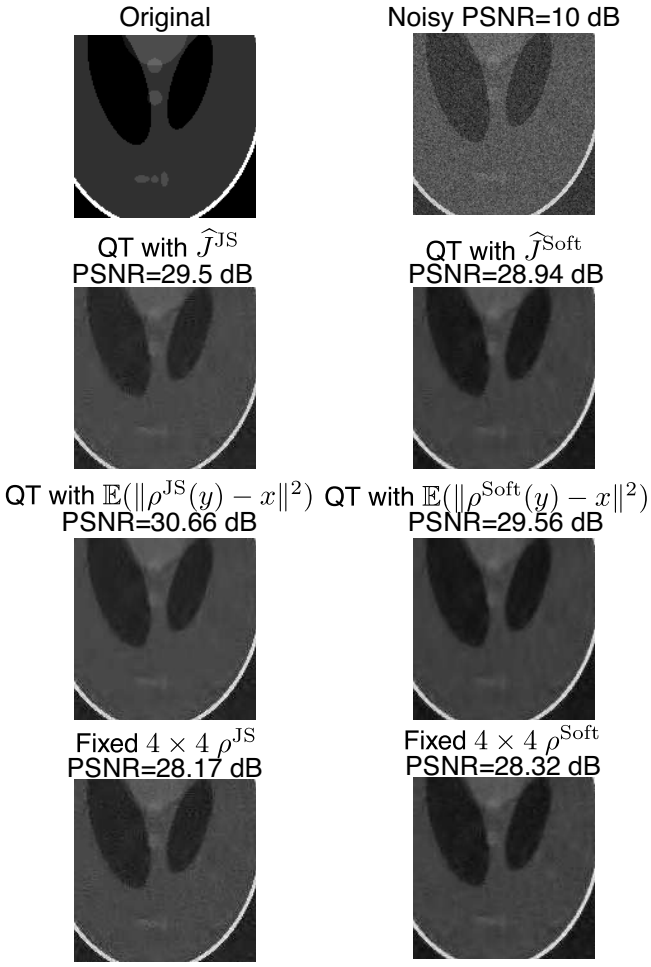
[18] K. Rosenblum, L. Zelnik-Manor, and Y. Eldar. Dictionary optimization for block-sparse representations. In *AAAI Fall 2010 Symposium on Manifold Learning*, 2010.

[19] R. Shukla, P.L. Dragotti, M. Do, and M. Vetterli. Rate distortion optimized tree structured compression algorithms for piecewise smooth images. *IEEE Trans. Image Proc.*, 14(3), 2005.

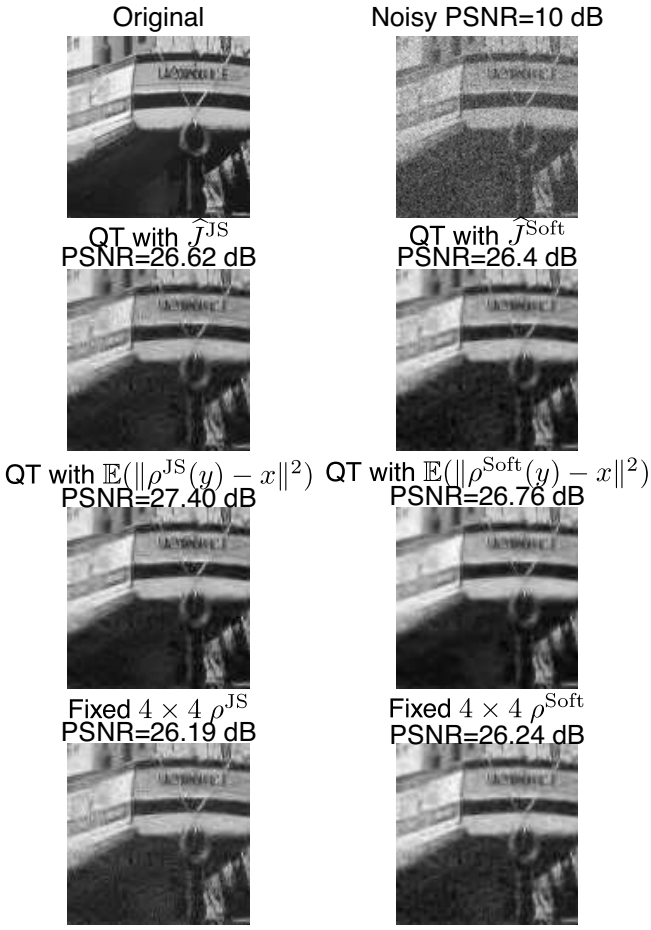
[20] M. Wakin, J. Romberg, H. Choi, and R. Baraniuk. Wavelet-domain approximation and compression of piecewise smooth images. *IEEE Trans. Image Proc.*, 15(5):1071–1087, May 2006.

[21] G. Yu, S. Mallat, and E. Bacry. Audio denoising by time-frequency block thresholding. *IEEE Trans. on Sig.*, 56(5):1830–1839, 2008.

[22] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *J. of The Roy. Stat. Soc. B*, 68(1):49–67, 2006.



(a)



(b)

Figure 3: *Zoom on denoising results with input PSNR=10dB. (a): Synthetic image. (b): Natural image.*