MARKOV-TREE BAYESIAN GROUP-SPARSE MODELING WITH WAVELETS

Ganchi Zhang and Nick Kingsbury

Signal Processing Group, Dept. of Engineering, University of Cambridge, UK

ABSTRACT

In this paper, we propose a new Markov-tree Bayesian modeling of wavelet coefficients. Based on a group-sparse Gaussian Scale Mixtures model with 2-layer cascaded Gamma distributions for the variances, the proposed method effectively exploits both intrascale and interscale relationships across wavelet subbands. To determine the posterior distribution, we apply Variational Bayesian inference with a subband adaptive majorization-minimization method to make the method tractable for large problems.

***Index Terms*—** Group-sparse modeling, Markov-tree, majorization minimization, variational Bayesian, dual-tree complex wavelets.

1. INTRODUCTION

Linear inverse problems appear often in many applications of image processing where a noisy indirect observation $\mathbf{y}$, of an original image $\mathbf{x}$, is modeled as $\mathbf{y} = \mathbf{B}\mathbf{x} + \mathbf{n}$, where $\mathbf{B}$ of size $M \times N$ is the matrix representation of a direct linear operator and $\mathbf{n}$ is usually additive Gaussian noise with variance ν^2. Wavelet-based methods are good for solving ill-posed image restoration problems because natural images can often be sparsified using a wavelet basis [1]. Note that, the statistical properties of wavelet coefficients can often be modeled by heavy-tailed Gaussian scale mixture (GSM) priors that capture the intrascale relationships among wavelet coefficients [2, 3]. However, many authors have argued that there is a strong persistence of large/small wavelet coefficients across scales, and such interscale relationships are beneficial for modeling wavelet coefficients [4, 5, 6]. In general, this interscale dependency mechanism can be well represented using a wavelet tree structure where child coefficient energy relates strongly to parent energy [6]. Various methods such as bivariate shrinkage [7], Hidden Markov Tree (HMT) [5], and overlapping-group penalties [6] have been used to exploit the parent-child relationship.

2. MODEL FORMULATION

In this paper, our contribution is a new Markov-tree based model that explores both intrascale and interscale dependencies among wavelet coefficients. Assume we can represent the image $\mathbf{x}$ by wavelet expansion as $\mathbf{x} = \mathbf{M}\mathbf{w}$ where $\mathbf{M}$ is the inverse wavelet transform, and $\mathbf{w}$ is an $N \times 1$ vector which contains all wavelet coefficients. This results in a wavelet-based formulation as $\mathbf{y} = \mathbf{B}\mathbf{M}\mathbf{w} + \mathbf{n}$. It is noted that for an orthogonal basis, $\mathbf{M}$ is a square orthogonal matrix, whereas for an over-complete dictionary (e.g. a tight frame), $\mathbf{M}$ has N columns and M rows, with $N > M$ [1]. The resulting likelihood of the data can be shown to be

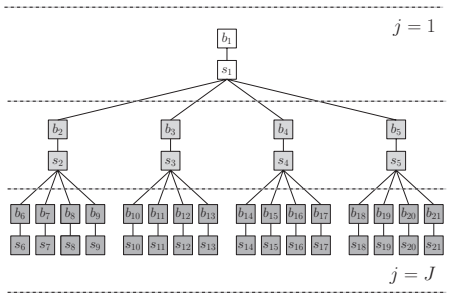


Fig. 1: Joint probability of $\mathbf{s}$ and $\mathbf{b}$ based on a Markov-tree model.

$$p(\mathbf{y}|\mathbf{w}, \nu^2) = \left(2\pi\nu^2\right)^{-\frac{M}{2}} \exp\{-\frac{1}{2\nu^2}\|\mathbf{y} - \mathbf{B}\mathbf{M}\mathbf{w}\|^2\} \quad (1)$$

Following the assumptions in [8], we use a non-overlapped group-sparse GSM model to model $\mathbf{w}$, and the prior of $\mathbf{w}$, conditioned on $\mathbf{S}$, can then be expressed as

$$p\left(\mathbf{w}|\mathbf{S}\right) = \prod_{i=1}^{G} \mathcal{N}\left(\mathbf{w}_i|0, \sigma_i^2\right) = \mathcal{N}\left(\mathbf{w}|0, \mathbf{S}^{-1}\right) \quad (2)$$

where the i^{th} group $\mathbf{w}_i$ is a vector of size g_i whose elements are drawn from a zero-mean Gaussian distribution with a signal variance σ_i^2 (as yet unknown), and where G is the number of groups and $\mathbf{S}$ is a diagonal matrix formed from the vector $\mathbf{s}$ whose i^{th} entry is $s_i = 1/\sigma_i^2$. Note that $N = \sum_{i=1}^{G} g_i$, and that, because $\mathbf{S}$ needs to be of size $N \times N$, when $N > G$ the diagonal of $\mathbf{S}$ is an expanded form of $\mathbf{s}$ in which each s_i is repeated g_i times [8]. From (1) and (2), the posterior distribution for $\mathbf{w}$ is

$$p\left(\mathbf{w}|\mathbf{y}, \mathbf{S}, \nu^2\right) = \frac{p\left(\mathbf{y}|\mathbf{w}, \nu^2\right) \times p\left(\mathbf{w}|\mathbf{S}\right)}{p\left(\mathbf{y}|\mathbf{S}, \nu^2\right)} \quad (3)$$

978-1-4673-7802-4/16/$31.00 ©2016 IEEE

which can be rearranged into a Gaussian form as

$$p\left(\mathbf{w}|\mathbf{y},\mathbf{S},\nu^2\right) = \mathcal{N}\left(\mathbf{w}|\mu,\Sigma\right) \quad (4)$$

with

$$\mu = \nu^{-2}\Sigma\mathbf{M}^T\mathbf{B}^T\mathbf{y} \quad (5)$$

$$\Sigma = \left(\nu^{-2}\mathbf{M}^T\mathbf{B}^T\mathbf{B}\mathbf{M}+\mathbf{S}\right)^{-1} \quad (6)$$

The computation of the posterior variance Σ requires inversion of the $N \times N$ square matrix $(\nu^{-2}\mathbf{M}^T\mathbf{B}^T\mathbf{B}\mathbf{M}+\mathbf{S})$. This operation is not computationally feasible for large images and 3D datasets. To overcome this, in [8], we introduced the Bayesian Majorization Minimization (MM) framework to simplify the posterior:

$$\overline{p}\left(\mathbf{w},\mathbf{z}|\mathbf{y},\mathbf{S},\nu^2\right) = p\left(\mathbf{z}|\mathbf{w}\right) \times p\left(\mathbf{w}|\mathbf{y},\mathbf{S},\nu^2\right) \quad (7)$$

where

$$p\left(\mathbf{z}|\mathbf{w}\right) = \mathcal{N}\left(\mathbf{w}|\mathbf{z},\Sigma_{\mathbf{z}}\right)$$

$$\propto \exp\{-(\mathbf{w}-\mathbf{z})^T\frac{\Lambda_\alpha - \mathbf{M}^T\mathbf{B}^T\mathbf{B}\mathbf{M}}{2\nu^2}(\mathbf{w}-\mathbf{z})\} \quad (8)$$

Λ_α is an $N \times N$ diagonal matrix formed from a vector α whose elements α_j may be optimized independently for each subspace/subband j of $\mathbf{M}$, such that $(\Lambda_\alpha - \mathbf{M}^T\mathbf{B}^T\mathbf{B}\mathbf{M})$ is positive definite [1, 9, 10]. When $\mathbf{z}$ is given (typically as a previous estimate for $\mathbf{w}$), the approximation model $\overline{p}\left(\mathbf{w},\mathbf{z}|\mathbf{y},\mathbf{S},\nu^2\right)$ can be shown as

$$\overline{p}\left(\mathbf{w}|\mathbf{y},\mathbf{z},\mathbf{S},\nu^2\right) = \mathcal{N}\left(\mathbf{w}|\overline{\mu},\overline{\Sigma}\right) \quad (9)$$

with

$$\overline{\mu} = \nu^{-2}\overline{\Sigma}[(\Lambda_\alpha - \mathbf{M}^T\mathbf{B}^T\mathbf{B}\mathbf{M})\mathbf{z} + \mathbf{M}^T\mathbf{B}^T\mathbf{y}] \quad (10)$$

$$\overline{\Sigma} = (\nu^{-2}\Lambda_\alpha + \mathbf{S})^{-1} \quad (11)$$

3. MARKOV-TREE BASED VB STRATEGY

To proceed with Bayesian Inference and model dependency and persistence across scale, we propose a joint probability density between $\mathbf{s}$ and a hidden variable $\mathbf{b}$ based on a Markov-tree model, as shown in Fig. 1. This is the main contribution of this paper. In this tree structure, we denote the parent node of node i by $d(i)$. We use $l(i)$ to indicate the level of node i, and denote J as the number of levels of wavelet decomposition. A key feature of this new Markov-tree model is that there is a hidden node b_i linking node s_i to its parent node $s_{d(i)}$, which differs distinctly from the conventional HMT model where s_i and $s_{d(i)}$ are linked using a predefined transition matrix. We thus have

$$p(\mathbf{s},\mathbf{b}) = p(\mathbf{s}_1|\mathbf{b}_1)p_0(\mathbf{b}_1)\prod_{j=2}^{J} p(\mathbf{s}_j,\mathbf{b}_j|\mathbf{s}_{j-1}) \quad (12)$$

where, for level 1 (the root level),

$$p(\mathbf{s}_1|\mathbf{b}_1) = \prod_{i\in\{l(i)=1\}} p(s_i|a_1,b_i)$$

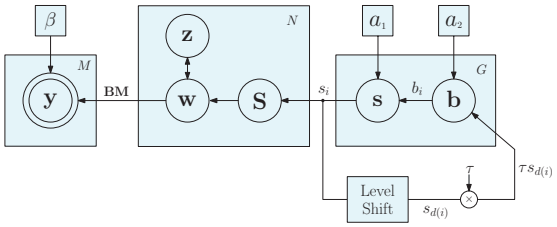


Fig. 2: The graphic model of linear regression with hierarchical priors. y and z are Gaussian distributions, w is a GSM, s and b are Gamma distributions.

and, for levels $2 \le j \le J$,

$$p(\mathbf{s}_j,\mathbf{b}_j|\mathbf{s}_{j-1}) = \prod_{i\in\{l(i)=j\}} p(s_i|a_1,b_i)p(b_i|a_2,\tau s_{d(i)})$$

To strongly encourage sparsity, we assume $\mathbf{S}$ and $\mathbf{b}$ are associated with Gamma priors such that $p(s_i|a_1,b_i) = \mathcal{G}(s_i;a_1,b_i)$ and $p(b_i|a_2,\tau s_{d(i)}) = \mathcal{G}(b_i;a_2,\tau s_{d(i)})$, where a_1 and a_2 are shape factors and τ is an energy gain factor. Since we do not have prior knowledge about root level nodes, we impose a noninformative Jeffrey's prior for root level $\mathbf{b}_1$ such that $p_0(\mathbf{b}_1) = \prod_{i\in\{l(i)=1\}}\frac{1}{b_i}$. The complete graphical model is shown in Fig. 2. As a result, the posterior of hidden variables now becomes

$$p\left(\mathbf{w},\mathbf{z},\mathbf{s},\mathbf{b}|\mathbf{y}\right) = \frac{p(\mathbf{w},\mathbf{z},\mathbf{s},\mathbf{b},\mathbf{y})}{p(\mathbf{y})}$$

$$= \frac{p(\mathbf{y}|\mathbf{w},\beta)\,p(\mathbf{z}|\mathbf{w})\,p(\mathbf{w}|\mathbf{S})\,p(\mathbf{s},\mathbf{b})}{p(\mathbf{y})} \quad (13)$$

where in different applications, $\beta = \nu^{-2}$ is either a user parameter for adjusting the regularization strength or can be estimated by imposing a Gamma distribution.

However the exact Bayesian inference of (13) cannot be performed as $p(\mathbf{y})$ is intractable. To approximate the posterior $p(\xi|\mathbf{y})$ where $\xi = \{\mathbf{w},\mathbf{z},\mathbf{s},\mathbf{b}\}$, we adopt the variational Bayesian (VB) approximation, which provides a distribution $q(\xi)$ to approximate $p(\xi|\mathbf{y})$ [11]. Specifically, $q(\xi)$ is determined by minimizing the Kullback-Leibler (KL) divergence between $q(\xi)$ and $p(\xi|\mathbf{y})$:

$$\mathrm{KL}(q(\xi)\|p(\xi|\mathbf{y})) = -\int q(\xi)\ln\left(\frac{p(\xi|\mathbf{y})}{q(\xi)}\right)d\xi \quad (14)$$

To find $q(\xi)$, we use the mean field approximation as

$$q(\xi) = q(\mathbf{w},\mathbf{z},\mathbf{s},\mathbf{b}) \approx q(\mathbf{w})q(\mathbf{z})q(\mathbf{s})q(\mathbf{b}) \quad (15)$$

Based on this factorization, the distribution of each variable $q(\lambda)$, $\lambda \in \xi$ can be optimized as [11]

$$\ln q(\lambda) = \langle \ln p(\xi|\mathbf{y})\rangle_{q(\xi\setminus\lambda)} = \langle \ln p(\xi,\mathbf{y})\rangle_{q(\xi\setminus\lambda)} + \text{const} \quad (16)$$

where $\langle\cdot\rangle_{q(\xi\setminus\lambda)}$ denotes expectation over all the factors of $q(\xi)$ except $q(\lambda)$. The key steps of Markov-tree based VBMM image restoration algorithm are shown in Algorithm 1.

Algorithm 1 Markov-tree based VBMM algorithm

Inputs: parameters for the sensing matrix $\mathbf{B}$, observation $\mathbf{y}$, Λ_α, a_1, a_2, β, initial estimations of $\mathbf{z}^{(0)}$, $\mathbf{s}^{(0)}$ and $\mathbf{b}^{(0)}$.
while iteration $t = 0 : t_{\max}$ or $\mathbf{z}$ has converged, **do**
  $\overline{\Sigma}^{(t)} = \left(\beta\Lambda_\alpha + \mathbf{S}^{(t)}\right)^{-1}$
  $\mu^t = \beta\overline{\Sigma}^{(t)}[\Lambda_\alpha \mathbf{z}^{(t)} - \mathbf{M}^T\mathbf{B}^T(\mathbf{B}\mathbf{M}\mathbf{z}^{(t)} - \mathbf{y})]$
  $\mathbf{w}^{(t+1)} = \mu^t$
  $\mathbf{z}^{(t+1)} = \mathbf{w}^{(t+1)}$
  for $i = 1...G$ **do**
    if $1 \le l(i) \le J-1$
    $s_i^{(t+1)} = \dfrac{g_i + 2(a_1 + 4a_2)}{\|\overline{\mu}_i^{(t)}\|^2 + tr[\overline{\Sigma}_i^{(t)}] + 2(b_i^{(t)} + \tau\sum_{k\in\mathbf{c}(i)} b_k^{(t)})}$
    elseif $l(i) = J$
    $s_i^{(t+1)} = \dfrac{g_i + 2a_1}{\|\overline{\mu}_i^{(t)}\|^2 + tr[\overline{\Sigma}_i^{(t)}] + 2b_i^{(t)}}$
  end for
  for $i = 1...G$ **do**
    if $l(i) = 1$
    ${b_i}^{(t+1)} = \dfrac{a_1}{{s_i}^{(t+1)}}$
    elseif $l(i) = 2 \le l(i) \le J$
    ${b_i}^{(t+1)} = \dfrac{a_1 + a_2}{{s_i}^{(t+1)} + \tau s_{d(i)}^{(t+1)}}$
  end for
  $\beta^{(t+1)} = \dfrac{M + 2c}{\|\mathbf{y} - \mathbf{B}\mathbf{M}\overline{\mu}^{(t)}\|^2 + \mathrm{tr}[\Lambda_\alpha\overline{\Sigma}^{(t)}] + 2d}$ (Optional)
end while
Output restored image $\mathbf{x} = \mathbf{M}\mathbf{z}^{t+1}$

4. SELECTION OF HYPERPARAMETERS

In this work, we have chosen the dual-tree complex wavelet transform (DT-CWT) as our redundant sparsifying transform as it has good sparsity inducing properties and is efficient to compute [12]. We optimize the parameters a_1, a_2 and τ based on the statistics of complex coefficients from natural images as shown in Figure 3, where we minimize the KL divergence between histograms of parent reweighted complex wavelet coefficient magnitudes $V_i = \frac{|\mathbf{w}_i|}{|\mathbf{w}_{d(i)}|}$ and pdfs of synthesized Gamma-Rayleigh distributed models for V_i, given by random samples drawn as follows

$$V_i \sim s_i V e^{\frac{-V^2 s_i}{2}}, \text{ with } \begin{cases} s_i \sim \mathcal{G}(s; a_1, b_i) \\ b_i \sim \mathcal{G}(b; a_2, \tau) \end{cases} \quad (17)$$

This takes into account the fact that the marginal distribution of DT-CWT coefficient magnitudes can be approximated by the Rayleigh law [13, 14]. Using several standard test images including Lenna, Cameraman, House, Peppers and Boat, we empirically find that the set of values are $a_1 = 11$, $a_2 = 1.5$ and $\tau = 1.2$ as shown in Fig. 4.

5. RESULTS

We present a set of experiments to evaluate our proposed Markov-tree VBMM (VM) algorithm. We show that the performance is significantly better than the VBMM algorithm (VC) described in Section 2 and its overlapped group tree-structured extensions (V1 and V4 in [8]).

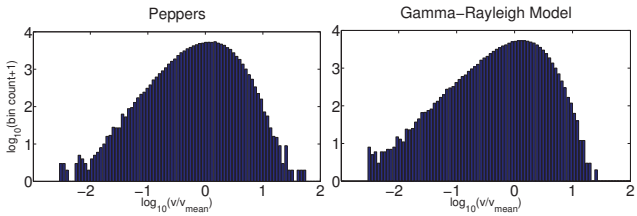


Fig. 3: Comparison of log-histograms of parent reweighted coefficient magnitudes V_i at wavelet level 1 for 256×256 Peppers image (left) with synthesized coefficients from the Gamma-Rayleigh model for V_i (right). The log-histograms for images Lenna, Cameraman, House and Boat are very similar.

5.1. 1-D heavisine signal recovery

In this section, the proposed algorithm is tested using a similar experiment to that described in [15, 16]. Since this is a 1-D CS problem, $\mathbf{B}$ becomes the sensing matrix Φ. Because it is piecewise smooth, the signal is compressible in the wavelet domain. The entries of sensing matrix Φ are generated from a Gaussian distribution $\mathcal{N}(0, 1)$. The observation is obtained from $M = 80$ noise free random Gaussian measurements. The recovery quality was assessed using root-mean-square error $\mathrm{RMSE} = \frac{\|\mathbf{x}-\mathbf{x}_r\|_2}{\|\mathbf{x}_r\|_2}$ where $\mathbf{x}$ is the reconstructed signal and $\mathbf{x}_r$ is the original (reference) signal. We set $\alpha = \rho(\Phi^T\Phi)$ in order to assure group acceleration as suggested in [15]. For wavelet bases, we have chosen the DT-CWT, and the level of decomposition is 9. Because the DT-CWT produces complex coefficients, we assume a pair of real and imaginary coefficients share the same variance and can be clustered into one group. As a result, we have $G = \frac{N}{2}$ groups for VM. We set $a_1 = 4, a_2 = 0.5$ and $\tau = 0.3$ according to the estimation of KL divergence in Section 4.

Table 1: RMSE results over 20 random samples of Φ

M	80		60		50	
	mean	std	mean	std	mean	std
VC	0.033	0.005	0.050	0.007	0.064	0.009
V1	0.032	0.007	0.049	0.006	0.064	0.006
V4	0.030	0.006	0.047	0.007	0.061	0.008
VM	0.029	0.007	0.045	0.007	0.059	0.009

To demonstrate the quality of recovery, we ran our algorithms for 100 iterations over 20 random implementations of Φ. Table 1 compares the RMSE results of VC, V1, V4 and VM. It is shown that VM achieves lowest RMSE results for all cases.

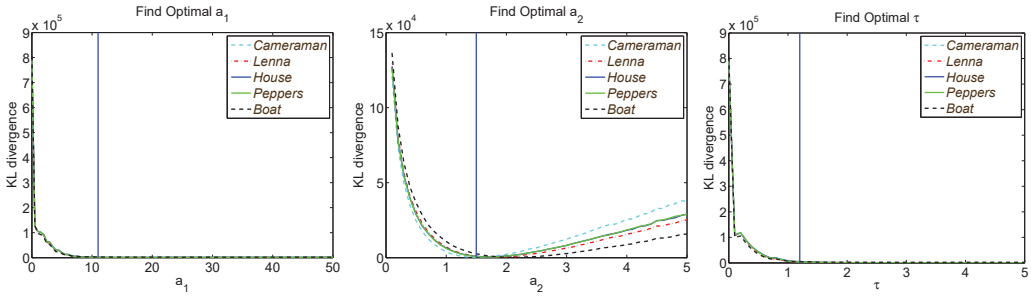


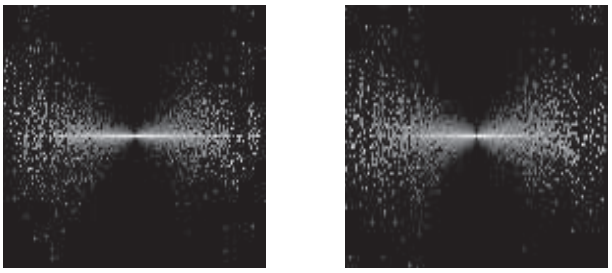
Fig. 4: Determination of optimal a_1, a_2 and τ based on the KL divergence between histograms of parent reweighted complex wavelet coefficient magnitudes $\mathbf{v}_i = \frac{|\mathbf{w}|_i}{|\mathbf{w}|_{d(i)}}$ and pdfs of synthesized Gamma-Rayleigh distributed models for $\mathbf{v}_i$.

Table 2: Average ISNR (dB) results for VC, V1, V4 and VM over 30 noise realizations.

iters		10	30	50	70	100
20dB	VC	2.66	3.15	3.35	3.45	3.52
	V1	2.95	3.55	3.67	3.71	3.74
	V4	2.96	**3.62**	3.72	3.74	3.76
	VM	**3.11**	3.60	**3.72**	**3.75**	**3.76**
40dB	VC	7.20	7.66	7.85	7.95	8.04
	V1	7.63	7.99	8.11	8.16	8.20
	V4	**7.57**	**8.01**	8.14	8.20	8.24
	VM	7.44	7.99	**8.24**	**8.36**	**8.46**
50dB	VC	10.17	10.66	10.87	10.99	11.08
	V1	10.17	10.75	10.94	11.04	11.14
	V4	10.19	10.86	11.06	11.16	11.26
	VM	**10.28**	**10.88**	**11.17**	**11.34**	**11.51**

5.2. Image deconvolution

In this section, we perform experiments for image deconvolution. Here the linear operator $\mathbf{B}$ becomes a convolution matrix $\mathbf{H}$. We have chosen the 2D directionally selective DT-CWT as our redundant sparsifying transform. Because the DT-CWT produces complex coefficients, we assume that a pair of real and imaginary coefficients share the same variance and form non-overlapping groups of size $g_i = 2$ for all i. As a result, we have $G = \frac{N}{2}$ groups for VM. In the experiment, we convolved the Cameraman image with a 9×9 uniform blur kernel. White Gaussian noise was added to the blurred image and the blurred signal-to-noise ratio (BSNR)$=10\log_{10}\frac{\|\mathbf{H}\mathbf{x}_r-\overline{\mathbf{H}\mathbf{x}_r}\|^2}{M\nu^2}$ was used to define the noise level. $\mathbf{x}_r$ is the original image and $\overline{\mathbf{H}\mathbf{x}_r}$ is the mean of $\mathbf{H}\mathbf{x}_r$. The improvement in signal-to-noise ratio (ISNR) $=10\log_{10}(\frac{\|\mathbf{y}-\mathbf{x}_r\|^2}{\|\mathbf{M}\mathbf{w}-\mathbf{x}_r\|^2})$ was used to evaluate each estimate $\mathbf{w}$. We calculated the matrix Λ_α using the method proposed in [10] where the contributions from every sub-band are accounted in determining the gain of a particular sub-band. The initial estimation of $\mathbf{x}_r$ was achieved by an under-regularized Wiener filter $\mathbf{x}_0$=$(\mathbf{H}^T\mathbf{H}+10^{-3}\nu^2\mathbf{I})^{-1}\mathbf{H}^T\mathbf{y}$. We set $a_1 = 11$, $a_2 = 1.5$ and $\tau = 1.2$ as shown in Fig. 4. In Table 2, we show average ISNR values over 30 noise realizations where three noise levels, BSNR= 20 dB, 40dB and 50 dB were considered. It is shown that VM outperforms VB, V1 and V4. To assess the dependency characteristic, Figure 5 shows the conditional histogram of parent-children wavelet coefficients across scale after 200 iterations. It is shown that compared with VC, VM imposes better persistence across scale, as desired.



(a) VC (b) VM

Fig. 5: Conditional histogram of parent-children wavelet coefficients for the Cameraman image. Note that the real part of complex wavelets are shown here while the imaginary part demonstrates similar characteristics.

6. CONCLUSION

Here we have extended the VBMM algorithm to incorporate a new Markov-tree structure, which effectively explores both intrascale and interscale dependencies among wavelet coefficients. The proposed method significantly outperforms the VBMM algorithm and its tree-structured extensions, while the computation per iteration increases by only 6%, relative to the VBMM which takes 0.09 seconds per iteration with a 256×256 image in Matlab. When deconvolving a 3D MRI dataset of size $256 \times 256 \times 256$, the algorithm takes 7.29 seconds per iteration, thus showing its order-N properties. Other related models, such as pairwise Markov trees [17] and hierarchical infinite divisibility [18], are to be studied in future research.

7. REFERENCES

[1] J. Bioucas-Dias M. Figueiredo and R. Nowak, "Majorization-minimization algorithms for wavelet-based image restoration," *IEEE Trans. on Image Process.*, vol. 16, pp. 2980–2991, 2007.

[2] J. Bioucas-Dias, "Bayesian wavelet-based image deconvolution: a GEM algorithm exploiting a class of heavy-tailed priors," *IEEE Trans. on Image Process.*, vol. 15, no. 4, pp. 937–951, 2006.

[3] J. Portilla, V. Strela, M. Wainwright, and E. Simoncelli, "Image denoising using scale mixtures of gaussians in the wavelet domain," *IEEE Trans. on Image Process.*, vol. 12, pp. 1338 – 1351, Nov. 2003.

[4] M. Crouse, R. Nowak, and R. Baraniuk, "Wavelet-based statistical signal processing using hidden markov models," *IEEE Trans. on Signal Process.*, vol. 46, no. 4, pp. 886–902, 1998.

[5] H. Choi J. Romberg and R. Baraniuk, "Bayesian tree-structured image modeling using wavelet-domain hidden markov models," *IEEE Trans. Image Process.*, vol. 10, pp. 1056–1068, 2001.

[6] N. Rao, R. Nowak, S. Wright, and N. Kingsbury, "Convex approaches to model wavelet sparsity patterns," in *IEEE International Conference on Image Processing 2011*, Sept. 11-14 2011, pp. 1917–1920.

[7] L. Sendur and I. Selesnick, "Bivariate shrinkage functions for wavelet-based denoising exploiting interscale dependency," *IEEE Trans. on Signal Process.*, vol. 50, no. 11, pp. 2744–2756, 2002.

[8] G. Zhang and N. Kingsbury, "Variational bayesian image restoration with group-sparse modeling of wavelet coefficients," *Digital Signal Process.*, vol. 47, pp. 157–168, 2015.

[9] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Comm. Pure Appl. Math.*, vol. 57, pp. 1413–1541, Nov. 2004.

[10] I. Bayram and I. Selesnick, "A subband adaptive iterative shrinkage/thresholding algorithm," *IEEE Trans. on Signal Process.*, vol. 58, pp. 1131 – 1143, March 2010.

[11] D. Tzikas, A. Likas, and N. Galatsanos, "The variational approximation for bayesian inference," *IEEE Signal Process. Mag.*, vol. 25, no. 6, pp. 131–146, 2008.

[12] I. Selesnick, R. Baraniuk, and N. Kingsbury, "The dual-tree complex wavelet transform," *IEEE Signal Process. Mag.*, vol. 22, pp. 123 – 151, Nov. 2005.

[13] S. Voloshynovskiy, O. Koval, and T. Pun, "Wavelet-based image denoising using nonstationary stochastic geometrical image priors," in *Electronic Imaging 2003*. International Society for Optics and Photonics, 2003, pp. 675–687.

[14] P. Hill, A. Achim, D. Bull, and M. Al-Mualla, "Dual-tree complex wavelet coefficient magnitude modelling using the bivariate cauchy–rayleigh distribution for image denoising," *Signal Processing*, vol. 105, pp. 464–472, 2014.

[15] Y. Zhang and N. Kingsbury, "Fast L_0-based sparse signal recovery," in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP) 2010*, Kittila, Finland, Aug. 29-Sept. 1 2010, pp. 403 – 408.

[16] R. Baraniuk, V. Cevher, M. Duarte, and C. Hegde, "Model-based compressive sensing," *IEEE Trans. Inform. Theory*, vol. 56, pp. 1982 – 2001, April 2010.

[17] F. Desbouvries, J. Lecomte, and W. Pieczynski, "Kalman filtering in pairwise markov trees," *Signal process.*, vol. 86, no. 5, pp. 1049–1054, 2006.

[18] X. Yuan, S. Rao, V.and Han, and L. Carin, "Hierarchical infinite divisibility for multiscale shrinkage," *Signal Processing, IEEE Trans. on*, vol. 62, no. 17, pp. 4363–4374, 2014.