MARKOV-TREE BAYESIAN GROUP-SPARSE MODELING: EFFICIENT SOLUTIONS TO LARGE INVERSE PROBLEMS

GANCHI ZHANG† AND NICK KINGSBURY†

Abstract. In this paper, we propose a new Markov-tree Bayesian modeling of wavelet coefficients. Based on a group-sparse GSM 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.

Key words. Image deconvolution, 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 $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}$, where $\mathbf{H}$ 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 [5] and overlapping-group penalties [6] have been used to exploit the parent-child relationship.

2. Model Formulation. We propose a new Markov-tree based model for exploring 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 $\mathbf{y} = \mathbf{H}\mathbf{M}\mathbf{w} + \mathbf{n}$. 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

$$p(\mathbf{y}|\mathbf{w}, \nu^2) = \left(2\pi\nu^2\right)^{-\frac{M}{2}} \exp\{-\frac{1}{2\nu^2}\|\mathbf{y} - \mathbf{H}\mathbf{M}\mathbf{w}\|^2\} \tag{2.1}$$

Similarly to [8], we use a non-overlapped group-sparse GSM model to model $\mathbf{w}$, and the conditional prior of $\mathbf{w}$ can then be expressed as

$$p\left(\mathbf{w}|\mathbf{S}\right) = \prod_{i=1}^{G} \mathcal{N}\left(\mathbf{w}_i|0, \sigma_i^2\right) = \mathcal{N}\left(\mathbf{w}|0, \mathbf{S}^{-1}\right) \tag{2.2}$$

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

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

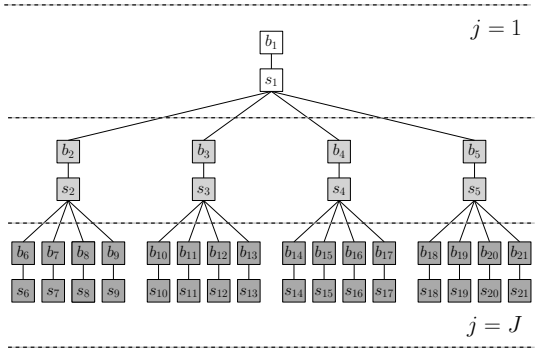

FIG. 2.1. *Joint probability of* **s** *and* **b** *based on a Markov-tree model*

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 (2.1) and (2.2), the posterior distribution for $\mathbf{w}$ is

$$p\left(\mathbf{w}|\mathbf{y},\mathbf{S},\nu^2\right) = \frac{p\left(\mathbf{y}|\mathbf{w},\nu^2\right) \times p\left(\mathbf{w}|\mathbf{S}\right)}{p\left(\mathbf{y}|\mathbf{S},\nu^2\right)} \tag{2.3}$$

which can be rearranged into a Gaussian form as

$$p\left(\mathbf{w}|\mathbf{y},\mathbf{S},\nu^2\right) = \mathcal{N}\left(\mathbf{w}|\mu,\Sigma\right) \tag{2.4}$$

with

$$\mu = \nu^{-2}\Sigma\mathbf{M}^T\mathbf{H}^T\mathbf{y}, \quad \Sigma = \left(\nu^{-2}\mathbf{M}^T\mathbf{H}^T\mathbf{H}\mathbf{M} + \mathbf{S}\right)^{-1} \tag{2.5}$$

The computation of the posterior variance Σ requires inversion of the $N \times N$ square matrix $(\nu^{-2}\mathbf{M}^T\mathbf{H}^T\mathbf{H}\mathbf{M} + \mathbf{S})$. This operation is not computationally feasible for large images and 3D datasets. To overcome this, we introduce a hidden variable $\mathbf{z}$ and the following approximation model for its posterior distribution:

$$\overline{p}\left(\mathbf{w},\mathbf{z}|\mathbf{y},\mathbf{S},\nu^2\right) = p\left(\mathbf{z}|\mathbf{w}\right) \times p\left(\mathbf{w}|\mathbf{y},\mathbf{S},\nu^2\right) \tag{2.6}$$

where

$$p\left(\mathbf{z}|\mathbf{w}\right) = \mathcal{N}\left(\mathbf{w}|\mathbf{z},\Sigma_{\mathbf{z}}\right) \propto \exp\{-(\mathbf{w}-\mathbf{z})^T \frac{\Lambda_\alpha - \mathbf{M}^T\mathbf{H}^T\mathbf{H}\mathbf{M}}{2\nu^2}(\mathbf{w}-\mathbf{z})\} \tag{2.7}$$

Λ_α is a $N \times N$ 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 $\nu^2\Sigma_{\mathbf{z}} = (\Lambda_\alpha - \mathbf{M}^T\mathbf{H}^T\mathbf{H}\mathbf{M})$ 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

$$\overline{p}\left(\mathbf{w}|\mathbf{y},\mathbf{z},\mathbf{S},\nu^2\right) = \mathcal{N}\left(\mathbf{w}|\overline{\mu},\overline{\Sigma}\right) \tag{2.8}$$

with

$$\overline{\mu} = \nu^{-2}\overline{\Sigma}[(\Lambda_\alpha - \mathbf{M}^T\mathbf{H}^T\mathbf{H}\mathbf{M})\mathbf{z} + \mathbf{M}^T\mathbf{H}^T\mathbf{y}], \quad \overline{\Sigma} = (\nu^{-2}\Lambda_\alpha + \mathbf{S})^{-1} \tag{2.9}$$

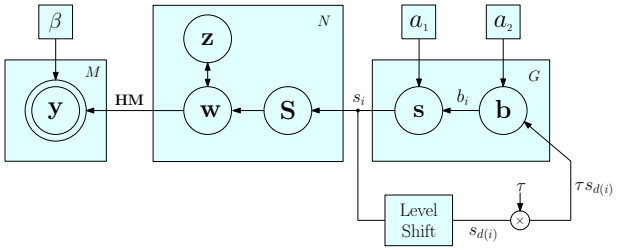

FIG. 2.2. *The graphic model of linear regression with hierarchical priors. y and z are Gaussian distributions, w is a GSM, s and b are Gamma distributions.*

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. 2.1. 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

$$p(\mathbf{s},\mathbf{b}) = p(\mathbf{s}_1|\mathbf{b}_1)p_0(\mathbf{b}_1)\prod_{j=2}^{J} p(\mathbf{s}_j,\mathbf{b}_j|\mathbf{s}_{j-1}) \tag{2.10}$$

where, for level 1 (the root level),

$$p(\mathbf{s}_1|\mathbf{b}_1) = \prod_{i\in\{l(i)=1\}} p(s_i|a_1,b_i) \tag{2.11}$$

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

$$p(\mathbf{s}_j,\mathbf{b}_j|\mathbf{s}_{j-1}) = \prod_{i\in\{l(i)=j\}} p(s_i|a_1,b_i)p(b_i|a_2,\tau s_{d(i)}) \tag{2.12}$$

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. 2.2. As a result, the posterior of hidden variables now becomes

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

where we have assumed that, in a given application, $\beta = \nu^{-2}$ is either known or is a user parameter for adjusting the regularization strength.

However the exact Bayesian inference of (2.8) 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 $p(\xi|\mathbf{y})$:

$$\mathrm{KL}(q(\xi)\|p(\xi|\mathbf{y})) = -\int q(\xi)\ln\Big(\frac{p(\xi|\mathbf{y})}{q(\xi)}\Big)\,d\xi \tag{2.13}$$

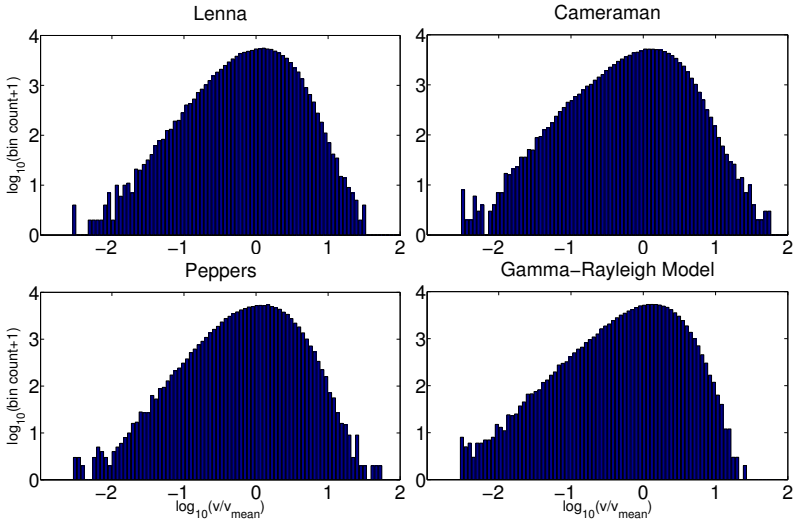

FIG. 3.1. *Comparison of log-histograms of parent reweighted coefficient magnitudes* $\mathbf{v}_i$ *at wavelet level 1 for 256 × 256 Lenna image (top left), Cameraman image (top right) and Peppers image (bottom left), with synthesized coefficients from the Gamma-Rayleigh model for* $\mathbf{v}_i$ *(bottom right).*

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

$$q(\xi) = q(\mathbf{w}, \mathbf{z}, \mathbf{s}, \mathbf{b}) \approx q(\mathbf{w})q(\mathbf{z})q(\mathbf{s})q(\mathbf{b}) \tag{2.14}$$

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

$$\ln q(\lambda) = \langle \ln p(\xi|\mathbf{y}) \rangle_{q(\xi \setminus \lambda)} = \langle \ln p(\xi, \mathbf{y}) \rangle_{q(\xi \setminus \lambda)} + \text{const} \tag{2.15}$$

where $\langle \cdot \rangle_{q(\xi \setminus \lambda)}$ denotes expectation over all the factors of $q(\xi)$ except $q(\lambda)$. .

3. Results. We present a set of experiments to evaluate our proposed Markov-tree VBMM (MT-VBMM) algorithm for image deconvolution where the linear operator $\mathbf{H}$ becomes a convolution matrix. We show that the performance is significantly better than the VBMM algorithm in [8].

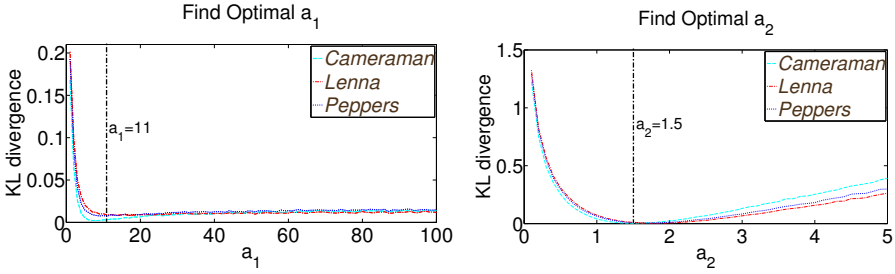

FIG. 3.2. *Determination of optimal* a_1 *and* a_2 *based on the KL divergence between histograms of parent reweighted complex wavelet coefficient magnitudes* $\mathbf{v}_i = \frac{|\mathbf{w}|_i}{|\mathbf{w}|_{d(i)}}$ *and pdfs of synthesized Gamma-Rayleigh distributed models for* $\mathbf{v}_i$.

We have chosen the DT CWT as our redundant sparsifying transform as it has good sparsity inducing properties and is efficient to compute [12]. Because the DT CWT produces complex coefficients, we assume that a pair of real and imaginary

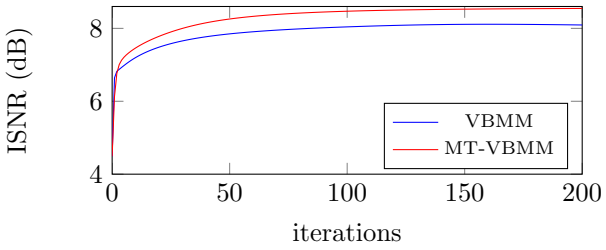

FIG. 3.3. ISNR *(dB) over 200 iterations, on Cameraman, BSNR: 40 dB.*

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 both MT-VBMM and VBMM. 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 a Wiener-type filter $\mathbf{x}_0=(\mathbf{H}^T\mathbf{H}+10^{-3}\nu^2\mathbf{I})^{-1}\mathbf{H}^T\mathbf{y}$. In the experiments, We optimize the parameters a_1 and a_2 based on the statistics of complex coefficients from natural images as shown in Fig. 3.1, where we minimize 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 $\mathbf{v}_i$, given by random saqmples drawn as follows

$$\mathbf{v}_i \sim s_i \mathbf{v} e^{\frac{-\mathbf{v}^2 s_i}{2}} \tag{3.1}$$

with

$$s_i \sim \mathcal{G}(s; a_1, b_i), \qquad b_i \sim \mathcal{G}(b; a_2, \tau) \tag{3.2}$$

It is found that the optimal values are $a_1 = 11$ and $a_2 = 1.5$ as shown in Fig. 3.2. Fig. 3.3 compares the ISNR results of MT-VBMM to VBMM over 200 iterations when the BSNR of the observation is 40dB and deconvolution results are shown in Fig. 3.4 for visual comparison.

(a) Original (b) Blurred (c) VBMM, ISNR: 8.09 dB (d) MT-VBMM, ISNR: 8.55 dB

FIG. 3.4. *Deconvolution and ISNR (dB) results, on Cameraman, BSNR: 40 dB.*

TABLE 4.1
AVERAGE ISNR RESULTS OVER 30 NOISE REALIZATIONS, 'V' stands for VBMM, 'MT-V' stands for Markov-tree VBMM

BSNR	20 dB		40 dB		50 dB	
Method	V	MT-V	V	MT-V	V	MT-V
10 iters	2.731	**3.113**	7.107	**7.439**	10.148	**10.281**
30 iters	3.282	**3.601**	7.531	**7.992**	10.656	**10.879**
50 iters	3.491	**3.719**	7.730	**8.239**	10.879	**11.165**
70 iters	3.582	**3.750**	7.842	**8.363**	10.996	**11.342**
100 iters	3.646	**3.759**	7.939	**8.458**	11.085	**11.506**

It is shown that proposed Markov-tree VBMM significantly outperforms the VBMM algorithm in terms of both ISNR results and visual quality. In Table 4.1, we show average ISNR values obtained from repeating our experiments over 30 noise realizations where another two noise levels, BSNR=20 dB, 50 dB were also considered. It is shown that the MT-VBMM converges faster than the VBMM in all cases.

4. 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, 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.

REFERENCES

[1] M. Figueiredo, J. Bioucas-Dias and R. Nowak, *Majorization–minimization algorithms for wavelet-based image restoration*, IEEE Trans. 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. Image Process. , vol. 15, 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. Image Process., vol. 12, pp. 1338-1351, 2003.
[4] M. Crouse, R. Nowak and R. Baraniuk, *Wavelet-based statistical signal processing using hidden* Markov *models*, IEEE Trans. Signal Process., vol. 46, pp. 886-902, 1998.
[5] J. Romberg, H. Choi 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*, Proc. IEEE ICIP 2011, pp. 1917-1920, 2011.
[7] L. Sendur and I. Selesnick, *Bivariate shrinkage functions for wavelet-based denoising exploiting interscale dependency*, IEEE Trans. Signal Process., vol. 50, pp. 2744-2756, 2002.
[8] G. Zhang and N. Kingsbury, *Fast* L0*-based Image Deconvolution with Variational* Bayesian *Inference and Majorization-Minimization*, in Proc. IEEE GlobalSIP 2013, pp. 1081-1084.
[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. 1413-1457, 2004.
[10] I. Bayram and I. Selesnick, *A subband adaptive iterative shrinkage/thresholding algorithm*, IEEE Trans. Signal Process., vol. 58, pp. 1131-1143, 2010.
[11] D. Tzikas, A. Likas and N. Galatsanos, *The variational approximation for Bayesian inference: Life after the EM algotrithm*, IEEE Signal Process. Mag., vol. 25, 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, 2005.