Variational Bayesian Image Restoration with Group-sparse Modeling of Wavelet Coefficients

Ganchi Zhang*, Nick Kingsbury

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

Abstract

In this work, we present a recent wavelet-based image restoration framework based on a group-sparse Gaussian scale mixture model. A hierarchical Bayesian estimation is derived using a combination of variational Bayesian inference and a subband-adaptive majorization-minimization method that simplifies computation of the posterior distribution. We show that both of these iterative methods can converge together without needing nested loops, and thus good solutions can be found rapidly in the non-convex search space. We also integrate our method, variational Bayesian with majorization minimization (VBMM), with tree-structured modeling of the wavelet coefficients. This extension achieves significant gains in performance over the coefficient-sparse version of the algorithm. The experimental results demonstrate that the proposed method and its tree-structured extensions are effective for various imaging applications such as image deconvolution, image superresolution and compressive sensing magnetic resonance imaging (MRI) reconstruction, and that they outperform more conventional sparsity-inducing methods based on the l_1-norm.

Keywords:
image restoration, wavelet group-sparse modeling, variational Bayesian inference, majorization minimization, dual-tree complex wavelets

*Corresponding author.

Email addresses: gz243@cam.ac.uk (Ganchi Zhang), ngk@eng.cam.ac.uk (Nick Kingsbury)

1. Introduction

Linear inverse problems appear often in many applications of image processing such as restoration, motion estimation, reconstruction and segmentation, where a noisy indirect observation $\mathbf{y}$, of an original image $\mathbf{x}$, is modeled as [1, 2]

$$\mathbf{y} = \mathbf{B}\mathbf{x} + \mathbf{n} \tag{1}$$

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.

In many scenarios, this inverse problem is highly ill-posed, i.e. the direct operator does not have an inverse or it is nearly singular so that its inverse is very sensitive to noise [3]. Thus it can only be solved satisfactorily by incorporating some regularization techniques, often using Bayesian inference with prior information [4]. In previous works, it is found that wavelet-based tools, such as the Discrete Wavelet Transform (DWT), are powerful for modeling this prior knowledge [4, 5, 6].

In the past two decades, the DWT has been exploited for a wide range of signal processing applications such as denoising, deconvolution, superresolution, compression and classification (see, e.g., [7, 8, 9, 10, 11]). The DWT provides an efficient implementation based on a filter bank structure utilizing decimation and two discrete filters, a low-pass and a high-pass filter [12]. Wavelet-based regularization methods are good for image restoration problems because wavelet coefficients tend to be sparse for most image types.

Although the DWT is compact, it suffers from shift dependency, lack of directionality, oscillation and aliasing [13]. These will significantly constrain the performance of a DWT-based signal processing system. To solve these shortcomings, the dual-tree complex wavelet transform (DT $\mathbb{C}$WT) first proposed by Kingsbury, is a recent simple and efficient redundant transform that has been widely used in solving diverse signal processing problems. The DT $\mathbb{C}$WT is better than the DWT for image restoration problems due to the fact that directional filters encourage greater sparsity and complex coefficients show more consistent persistence across scale. Other recent extensions of the DWT, such as curvelets [14] and contourlets [15], would also work in this context but few, if any, combine the efficiency and good performance of the dual-tree approach.

It is known that the wavelet coefficients of natural images display non-Gaussian statistics and their marginal distributions typically show a large peak at zero with long heavy tails [16, 17]. To account for this non-Gaussian

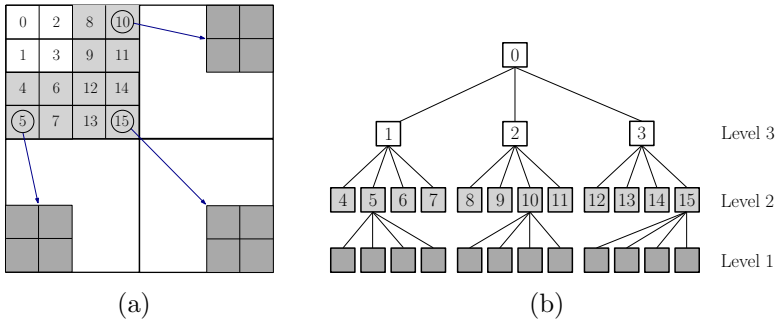


Figure 1: (a) 8×8 image with 3-level 2D DWT decomposition. (b) quadtree structure of wavelet coefficients.

behavior, many univariate parametric models such as generalized Laplacian distributions [17] and Bessel K form density models [18] have been previously used to model the wavelet coefficients. However, these models do not consider the persistence across scales of wavelet coefficients [19]. In fact, the energies of wavelet coefficients of natural images exhibit a strong characteristic signal-dependent structure. Fig. 1 depicts an example of quadtree structure that corresponds to an 8×8 image with 3-level 2D DWT decomposition. To well capture the statistical dependencies, bivariate shrinkage [20], Hidden Markov Tree models (HMM) [21, 22] and Gaussian Scale Mixture Models (GSM) [16, 23] have been widely applied to model wavelet coefficients whose energies are not randomly distributed. Among those methods, it is acknowledged that the GSM model can be used in the framework of sparse Bayesian learning (SBL) where the sparsity is obtained by reweighting the Gaussian prior [24, 25]. Based on this connection, several researchers have shown that Bayesian methods are applicable for wavelet-based regularization problems [4, 5, 16].

Recently, Bayesian group-sparse (or block sparse) modeling has emerged where the sparsity is imposed on groups instead of individual components [27, 28]. In [28], variational Bayesian (VB) inference is used for group-sparse modeling and has been shown to find sparse solutions effectively. These approaches can potentially be used in the wavelet domain since a pair of coefficients at a certain location and adjacent scales are typically both large or both small in amplitude [29]. Tree-structure existing in the wavelet domain allows group-sparse models to be easily constructed and used. One of the major contributions of our work is to investigate the use of Bayesian group-sparse modeling for wavelet-based regularization problems.

In [26], we proposed a hierarchical Bayesian modeling of wavelet coef-

ficients derived from a group-sparse GSM model. Based on a combination of VB inference with a subband-adaptive majorization minimization (MM) method, the VBMM method in [26] effectively simplifies computation of the posterior distribution and finds good solutions in the non-convex search space. In addition, the VBMM method has also shown good potential with group-sparse modeling. In [30], we incorporate the VBMM method with a wavelet tree structure based on overlapped groups, which leads to an improved solution compared with unstructured coefficient-sparse modeling.

In this paper, we extend the ideas from [26] to generalize the VBMM method and discuss the theoretical foundations in some detail. Different from [26] and [30], we also include the results of image superresolution and MRI image reconstruction. The proposed method can handle very large data sets with a good performance and low computation cost. The paper is organized as follows. Section 2 describes our proposed VBMM image restoration framework. Section 3 discusses the tree-structured extensions of VBMM. Experimental results are shown in Section 4. Conclusions are provided in Section 5.

2. VBMM Image restoration

In this section, we describe our proposed VBMM Image restoration framework and its tree-structured extensions.

2.1. Model Formulations

To obtain a wavelet-based formulation, we note that the image $\mathbf{x}$ can be represented by wavelet expansion as $\mathbf{x} = \mathbf{M}\mathbf{w}$ where $\mathbf{w}$ is a $N \times 1$ vector representing all wavelet coefficients, and $\mathbf{M}$ is the inverse wavelet transform whose columns are the wavelet basis functions. In the case of 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 $N > M$ [6]. The linear model in (1) then becomes

$$\mathbf{y} = \mathbf{B}\mathbf{M}\mathbf{w} + \mathbf{n} \tag{2}$$

and the resulting likelihood of the data assuming Gaussian noise $\mathbf{n}$ can be shown to be

$$p(\mathbf{y}|\mathbf{w}, \nu^2) = \left(2\pi\nu^2\right)^{-\frac{M}{2}} \exp\{-\frac{1}{2\nu^2}\|\mathbf{y} - \mathbf{B}\mathbf{M}\mathbf{w}\|^2\} \tag{3}$$

A GSM model is now employed to model the wavelet coefficients. Inspired from [28], we adopt a model which incorporates group sparsity such that $\mathbf{w}_i$, the i^{th} group of $\mathbf{w}$, follows a zero mean Gaussian distribution with an (as yet) unknown variance of σ_i^2 per element. Therefore the conditional prior of $\mathbf{w}$ can be expressed as

$$p\left(\mathbf{w}|\mathbf{S}\right) = \prod_{i=1}^{G} \mathcal{N}\left(\mathbf{w}_i|0,\sigma_i^2\right) = \mathcal{N}\left(\mathbf{w}|0,\mathbf{S}^{-1}\right) \tag{4}$$

where $\mathbf{w}_i$ is a vector of coefficients comprising the i^{th} group of size g_i, $\mathbf{S}$ is a diagonal matrix of size $N \times N$ formed from the vector $\mathbf{s}$ of size G whose i^{th} entry is $s_i = 1/\sigma_i^2$, and G denotes the number of groups. The case $G = N$ corresponds to independent sparse modeling of the wavelet coefficients [28]; whereas the case, $G = N/2$ and $g_i = 2$ for all i, can be used to model the real and imaginary parts of G complex coefficients, each with a 2-D circularly symmetric pdf. To be consistent with the following algebra, $\mathbf{S}$ needs to be of size $N \times N$ and, when $N > G$, its diagonal must be an expanded form of $\mathbf{s}$ where each s_i appears g_i times for the elements of group g_i, and $N = \sum_{i=1}^{G} g_i$.

To proceed with Bayesian inference, the posterior distribution can be calculated via:

$$p\left(\mathbf{w}|\mathbf{y},\mathbf{S},\nu^2\right) = \frac{p\left(\mathbf{y}|\mathbf{w},\nu^2\right) \times p\left(\mathbf{w}|\mathbf{S}\right)}{p\left(\mathbf{y}|\mathbf{S},\nu^2\right)} \tag{5}$$

Because both $p\left(\mathbf{y}|\mathbf{w},\nu^2\right)$ and $p\left(\mathbf{w}|\mathbf{S}\right)$ are Gaussian functions of $\mathbf{w}$, the posterior distribution can be rearranged into a Gaussian form as

$$p\left(\mathbf{w}|\mathbf{y},\mathbf{S},\nu^2\right) = \mathcal{N}\left(\mathbf{w}|\mu,\Sigma\right) \tag{6}$$

where, from (3) and (4):

$$\Sigma = \left(\nu^{-2}\mathbf{M}^T\mathbf{B}^T\mathbf{B}\mathbf{M} + \mathbf{S}\right)^{-1} \tag{7}$$

$$\mu = \nu^{-2}\Sigma\mathbf{M}^T\mathbf{B}^T\mathbf{y} \tag{8}$$

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, as N is often $\sim 10^7$ or more. Here we adopt the MM technique from [31], together with the recent subband-adaptive MM from [32, 33] to derive our fast algorithm. The aim is to

replace the troublesome ($\mathbf{M}^T\mathbf{B}^T\mathbf{B}\mathbf{M}$) term in (7) with the purely diagonal Λ_α, as in (16).

To derive MM from a Bayesian viewpoint, we now introduce the following approximation model for the posterior distribution of $\mathbf{w}$ and the new hidden variable $\mathbf{z}$:

$$\overline{p}\left(\mathbf{w},\mathbf{z}|\mathbf{y},\mathbf{S},\nu^2\right) = p\left(\mathbf{z}|\mathbf{w}\right) \times p\left(\mathbf{w}|\mathbf{y},\mathbf{S},\nu^2\right) \tag{9}$$

where

$$p\left(\mathbf{z}|\mathbf{w}\right) \propto \exp\{-(\mathbf{w}-\mathbf{z})^T \frac{\Lambda_\alpha - \mathbf{M}^T\mathbf{B}^T\mathbf{B}\mathbf{M}}{2\nu^2}(\mathbf{w}-\mathbf{z})\} \tag{10}$$

Note that taking the negative logarithm of both sides of (9) will give a similar surrogate function to that proposed in [32, 33]

$$\overline{J}_\alpha(\mathbf{w},\mathbf{z}) = J(\mathbf{w}) + (\mathbf{w}-\mathbf{z})^T \frac{\Lambda_\alpha - \mathbf{M}^T\mathbf{B}^T\mathbf{B}\mathbf{M}}{2\nu^2}(\mathbf{w}-\mathbf{z}) \tag{11}$$

where $\overline{J}_\alpha(\mathbf{w},\mathbf{z}) = -\ln \overline{p}(\mathbf{w},\mathbf{z}|\mathbf{y},\mathbf{S},\nu^2)$ and $J(\mathbf{w}) = -\ln p(\mathbf{w}|\mathbf{y},\mathbf{S},\nu^2)$ from (6). Λ_α is a diagonal matrix formed from a vector α whose elements α_j may be minimized independently for each subspace/subband j of $\mathbf{M}$, such that $\Lambda_\alpha - \mathbf{M}^T\mathbf{B}^T\mathbf{B}\mathbf{M}$ is just positive definite. This property ensures that $\overline{J}_\alpha(\mathbf{w},\mathbf{z}) > J(\mathbf{w})$ for any $\mathbf{w} \neq \mathbf{z}$, and $\overline{J}_\alpha(\mathbf{w},\mathbf{z}) = J(\mathbf{w})$ for $\mathbf{w} = \mathbf{z}$ [33], and hence produces monotonicity of the decay of $J(\mathbf{w})$, since for t^{th} iteration [6]

$$\begin{aligned} J\big(\mathbf{z}^{(t+1)}\big) &= \overline{J}_\alpha\big(\mathbf{z}^{(t+1)},\mathbf{z}^{(t+1)}\big) \\ &\leq \overline{J}_\alpha\big(\mathbf{z}^{(t+1)},\mathbf{z}^{(t)}\big) \\ &\leq \overline{J}_\alpha\big(\mathbf{z}^{(t)},\mathbf{z}^{(t)}\big) = J\big(\mathbf{z}^{(t)}\big) \end{aligned} \tag{12}$$

The first inequality results from $\overline{J}_\alpha(\mathbf{w},\mathbf{z}) \geq \overline{J}_\alpha(\mathbf{w},\mathbf{w}) = J(\mathbf{w})$ for any $\mathbf{w}$ and $\mathbf{z}$. The second inequality comes from the fact that $\overline{J}_\alpha\big(\mathbf{w},\mathbf{z}^{(t)}\big)$ attains its minimum for $\mathbf{w} = \mathbf{z}^{(t+1)}$ according to the definition of a MM iterative algorithm [6]:

$$\mathbf{z}^{(t+1)} = \arg\min_{\mathbf{w}} \left(\mathbf{w},\mathbf{z}^{(t)}\right) \tag{13}$$

The use of (11) is known as the subband-adaptive MM technique. Related algorithms are called subband-adaptive iterative shrinkage/thresholding (SIST)

algorithms [32, 33, 34]. In [33] and [34], fast algorithms for computing Λ_α are proposed. We refer the reader to Appendix 6.1 for more detailed discussion where we summarize ways of selecting Λ_α for both convolutional and non-convolutional kernels.

Because $p(\mathbf{z}|\mathbf{w}) \propto \mathcal{N}\left(\mathbf{w}|\mathbf{z}, \nu^2(\Lambda_\alpha - \mathbf{M}^T\mathbf{B}^T\mathbf{B}\mathbf{M})^{-1}\right)$ and $p(\mathbf{w}|\mathbf{y}, \mathbf{S}, \nu^2)$, given by (6), (7) and (8), are Gaussian functions of $\mathbf{w}$, the approximation model $\overline{p}(\mathbf{w}, \mathbf{z}|\mathbf{y}, \mathbf{S}, \nu^2)$ is also a Gaussian distribution and, when $\mathbf{z}$ is given, we can rearrange (9) into the Gaussian form:

$$\overline{p}\left(\mathbf{w}|\mathbf{y}, \mathbf{z}, \mathbf{S}, \nu^2\right) = \mathcal{N}\left(\mathbf{w}|\overline{\mu}, \overline{\Sigma}\right) \tag{14}$$

with

$$\overline{\mu} = \nu^{-2}\overline{\Sigma}[(\Lambda_\alpha - \mathbf{M}^T\mathbf{B}^T\mathbf{B}\mathbf{M})\mathbf{z} + \mathbf{M}^T\mathbf{B}^T\mathbf{y}] \tag{15}$$

$$\overline{\Sigma} = (\nu^{-2}\Lambda_\alpha + \mathbf{S})^{-1} \tag{16}$$

where $\overline{\Sigma}^{-1}$ is now purely diagonal and easy to invert. This gives the subband-adaptive MM technique in a Bayesian framework as introduced in [26], whose convergence rate is improved by keeping the spectral radius of the matrix $(\Lambda_\alpha - \mathbf{M}^T\mathbf{B}^T\mathbf{B}\mathbf{M})$ small.

In the above approximation model for $\overline{p}(\mathbf{w}|\mathbf{y}, \mathbf{z}, \mathbf{S}, \nu^2)$, it is required to estimate the inverse signal variance $\mathbf{S}$ since we do not know it yet. In [35], maximum a *posteriori* (MAP) estimation with an independent prior is used to determine $\mathbf{S}$. However, it is known that point estimates do not define much of the available signal space, and better convergence is achieved if approximate distributions of the posterior density are used. In fact, VB inference possesses this property by providing a distribution that approximates the posterior distribution of the hidden variables [36], and it has been shown in [37] that VB inference can effectively smooth out local minima and help to ensure that a near-global minimum solution is found. Compared with MAP, VB inference can be seen as a more principled approach, which should find improved solutions to inverse problems.

2.2. VB Continuation Strategy

In this section, we apply the VB approximation to derive the continuation strategies of our model. To update the variables appearing in (9), we construct a 3-layer hierarchical prior as described in [26]. To be more specific, we impose a multivariate Gamma distribution for the inverse signal variance

vector $\mathbf{s}$ (which is expanded into $\mathbf{S}$ for (4)) using shape parameter a and rate vector $\mathbf{b}$:

$$p(\mathbf{s}|a,\mathbf{b}) = \prod_{i=1}^{G} \frac{b_i^a}{\Gamma(a)} s_i^{a-1} \exp(-b_i s_i) \tag{17}$$

and a further Gamma distribution for $\mathbf{b}$:

$$p(\mathbf{b}|k,\theta) = \prod_{i=1}^{G} \frac{\theta^k}{\Gamma(k)} b_i^{k-1} \exp(-\theta b_i) \tag{18}$$

When shape k and rate θ both tend towards zero, the Gamma prior on $p(\mathbf{b})$ tends to a noninformative Jeffreys prior [38], and therefore the mean of signal variance σ^2 approximately follows a noninformative prior. This means that the posterior depends only on the data and not the prior, which is known to strongly promote sparse estimates [39]. If prior knowledge is available for the model, we can tune the hyperparameters so that the prior becomes more informative. As a result, we can move between the informative and noninformative prior flexibly using the 3-layer Gamma GSM model. Because it often leads to simple closed-form solutions, we will use it in this paper for encouraging sparsity. However, there are many alternative choices for a sparsity-favouring prior that can be used for the GSM such as Multivariate Laplace, Inverse Gaussian and Bessel function distributions [28].

Here we keep the noise variance $\nu^2 \equiv \beta^{-1}$ as a user parameter in order to be able to adjust the regularization strength. Note that although ν^2 can be estimated via Bayesian inference as discussed in Section 2.3, its estimate can be inaccurate because of the difficulty of accurately separating broadband signal components from noise. The complete graphical model is shown in Fig. 2. As a result, the posterior of hidden variables now becomes

$$\begin{aligned} p\left(\mathbf{w},\mathbf{z},\mathbf{s},\mathbf{b}|\mathbf{y},\beta\right) &= \frac{p(\mathbf{w},\mathbf{z},\mathbf{s},\mathbf{b},\mathbf{y}|\beta)}{p(\mathbf{y})} \\ &= \frac{p\left(\mathbf{y}|\mathbf{w},\beta\right) p\left(\mathbf{w}|\mathbf{S}\right) p(\mathbf{z}|\mathbf{w}) p(\mathbf{s}|a,\mathbf{b}) p(\mathbf{b}|k,\theta)}{p\left(\mathbf{y}\right)} \end{aligned} \tag{19}$$

where $\beta \equiv \nu^{-2}$. Note that the exact Bayesian posterior of (19) cannot be calculated as the marginal likelihood $p\left(\mathbf{y}\right)$ is intractable [36]. To approximate the posterior $p(\xi|\mathbf{y},\beta)$ where $\xi = \{\mathbf{w},\mathbf{z},\mathbf{s},\mathbf{b}\}$, we adopt the VB approximation, which provides a distribution $q(\xi)$ to approximate $p(\xi|\mathbf{y},\beta)$ [28, 36].

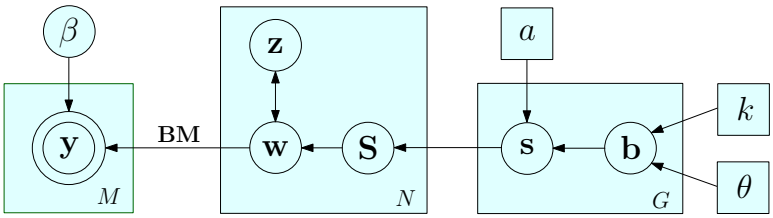


Figure 2: The graphical model of linear regression with hierarchical priors. $\mathbf{y}$ and $\mathbf{z}$ are Gaussian distributions, $\mathbf{w}$ is a GSM, $\mathbf{s}$ and $\mathbf{b}$ are Gamma distributions.

To be specific, $q(\xi)$ is determined by minimizing the Kullback-Leibler (KL) divergence between $q(\xi)$ and $p(\xi|\mathbf{y})$.

From the basic probabilistic theory, we can decompose the log marginal probability using [40]

$$\ln p(\mathbf{y}) = \mathcal{L}(q(\xi)) + \mathrm{KL}(q(\xi)\|p(\xi|\mathbf{y})) \tag{20}$$

with

$$\mathcal{L}(q(\xi)) = \int q(\xi) \ln\left(\frac{p(\xi,\mathbf{y})}{q(\xi)}\right) d\xi \tag{21}$$

$$\mathrm{KL}(q(\xi)\|p(\xi|\mathbf{y})) = -\int q(\xi) \ln\left(\frac{p(\xi|\mathbf{y})}{q(\xi)}\right) d\xi \tag{22}$$

By rearranging (20), we get

$$\mathcal{L}(q(\xi)) = \ln p(\mathbf{y}) - \mathrm{KL}(q(\xi)\|p(\xi|\mathbf{y})) \tag{23}$$

When solving for the pdfs of the parameters in ξ, we want to build the joint pdf $q(\xi)$ which most closely matches the shape of $p(\xi,\mathbf{y})$ for the given $\mathbf{y}$, i.e. which maximizes $\mathcal{L}(q(\xi))$. This is equivalent to minimizing the KL divergence [40]. Because $\ln(\frac{p(\xi|\mathbf{y})}{q(\xi)}) \leq \frac{p(\xi|\mathbf{y})}{q(\xi)} - 1$ and both $q(\xi)$ and $p(\xi|\mathbf{y})$ are valid pdfs over ξ, we get $\mathrm{KL}(q(\xi)\|p(\xi|\mathbf{y})) \geq 0$. Since $\ln p(\mathbf{y})$ is a constant, the maximum of $\mathcal{L}(q(\xi))$ occurs when $q(\xi)$ equals the posterior distribution $p(\xi|\mathbf{y})$, and then $\mathrm{KL}(q(\xi)\|p(\xi|\mathbf{y})) = 0$ and $\mathcal{L}(q(\xi)) = \ln p(\mathbf{y})$.

This can be viewed as a generalization of the Expectation-Maximization (EM) algorithm such that the VB model only contains hidden variables and no parameters [36]. This methodology is also referred to as nonparametric distribution estimation in statistics [41]. Although we can calculate $p(\xi,\mathbf{y})$

from the model, the true posterior distribution $p(\xi|\mathbf{y})$ is intractable because we do not know $p(\mathbf{y})$. This means that $q(\xi)$ cannot be simply obtained. Therefore we consider a restricted family of distributions $q(\xi)$ instead and minimize the KL divergence for each member of this family [40]. By using the mean-field approximation which assumes posterior independence between different variables [28], we can then factorize $q(\xi)$ into disjoint groups, so that

$$q(\xi) = \prod_{i=1}^{D} q_i(\xi_i) \tag{24}$$

where D is the total number of disjoint groups. For instance, in VBMM, $q(\mathbf{w})$, $q(\mathbf{z})$, $q(\mathbf{s})$ (and hence $\mathbf{S}$) and $q(\mathbf{b})$ are pdfs of disjoint groups from $q(\xi)$. The purpose is to find a set of $q_i(\xi_i)$, $i = 1 \dots D$, such that $\mathcal{L}(q(\xi))$ is maximized, so that $q(\xi) \approx p(\xi|\mathbf{y})$.

Based on this factorization, the distribution of each variable $q(\omega)$, $\omega \in \xi$, which minimizes (22) can be optimized as

$$\begin{aligned} \ln q(\omega) &= \langle \ln p(\xi|\mathbf{y}) \rangle_{q(\xi \setminus \omega)} \\ &= \langle \ln p(\xi, \mathbf{y}) \rangle_{q(\xi \setminus \omega)} + \text{const} \end{aligned} \tag{25}$$

where $\langle \cdot \rangle_q$ denotes expectation over q and $\xi \setminus \omega$ means the set of ξ with ω removed. The detailed proof is given in Appendix 6.2. By sequentially calculating $q(\omega)$ for each $\omega \in \xi$ in turn, we obtain the following updating rules.

(i) ***Optimize*** $\ln q(\mathbf{w})$ ***using (5), (9) and (14)***

$$\begin{aligned} \ln q(\mathbf{w}) &= \langle \ln p(\xi, \mathbf{y}) \rangle_{q(\mathbf{z})q(\mathbf{s})q(\mathbf{b})} + \text{const} \\ &= \langle \ln(p(\mathbf{y}|\mathbf{w}, \beta)p(\mathbf{w}|\mathbf{S})p(\mathbf{z}|\mathbf{w})) \rangle + \text{const} \\ &= \langle \ln p(\mathbf{w}|\mathbf{z}, \mathbf{y}, \beta, \mathbf{S}) \rangle + \text{const} \\ &= -\frac{1}{2}\mathbf{w}^T \overline{\Sigma}^{-1} \mathbf{w} + \mathbf{w}^T \overline{\Sigma}^{-1} \overline{\mu} + \text{const} \end{aligned} \tag{26}$$

This represents a multivariate Gaussian distribution $q(\mathbf{w})$ with mean $\overline{\mu}$ and covariance $\overline{\Sigma}$. Thus the mean of $\mathbf{w} = \overline{\mu}$, and hence

$$\mathbf{w}^{(t+1)} = \langle \mathbf{w} \rangle = \overline{\mu}^{(t)} \tag{27}$$

where $\overline{\mu}^{(t)}$ is computed using the current estimates of $\mathbf{S}^{(t)}$ and $\mathbf{z}^{(t)}$ as in (15) and (16):

$$\overline{\Sigma}^{(t)} = \left(\beta\Lambda_\alpha + \mathbf{S}^{(t)}\right)^{-1} \tag{28}$$

$$\overline{\mu}^{(t)} = \beta\overline{\Sigma}^{(t)}[\Lambda_\alpha \mathbf{z}^{(t)} - \mathbf{M}^T\mathbf{B}^T(\mathbf{B}\mathbf{M}\mathbf{z}^{(t)} - \mathbf{y})] \tag{29}$$

(ii) ***Optimize*** $\ln q(\mathbf{z})$ ***using (10)***

$$\begin{aligned}\ln q(\mathbf{z}) &= \langle \ln p(\xi, \mathbf{y})\rangle_{q(\mathbf{w})q(\mathbf{s})q(\mathbf{b})} + \text{const} \\ &= \langle \ln p(\mathbf{z}|\mathbf{w})\rangle + \text{const} \\ &= -\frac{1}{2}\mathbf{z}^T\Sigma_\mathbf{z}\mathbf{z} + \mathbf{z}^T\Sigma_\mathbf{z}\langle\mathbf{w}\rangle + \text{const}\end{aligned} \tag{30}$$

This represents a Gaussian distribution $q(\mathbf{z})$ where $\Sigma_\mathbf{z} = \nu^2(\Lambda_\alpha - \mathbf{M}^T\mathbf{B}^T\mathbf{B}\mathbf{M})^{-1}$. Thus provided $(\Lambda_\alpha - \mathbf{M}^T\mathbf{B}^T\mathbf{B}\mathbf{M})$ is positive definite, the mean of $\mathbf{z}$ occurs when

$$\mathbf{z}^{(t+1)} = \langle\mathbf{w}\rangle = \mathbf{w}^{(t+1)} \tag{31}$$

(iii) ***Optimize*** $\ln q(\mathbf{s})$ ***using (4) and (17)***

$$\begin{aligned}&\ln q(\mathbf{s}) = \langle \ln p(\xi, \mathbf{y})\rangle_{q(\mathbf{w})q(\mathbf{z})q(\mathbf{b})} + \text{const} \\ &= \langle \ln(p(\mathbf{w}|\mathbf{S})p(\mathbf{s}|a,\mathbf{b}))\rangle + \text{const} \\ &= \Big\langle \sum_{i=1}^{G} \big(\frac{g_i}{2}\ln s_i - \frac{1}{2}s_i\|\mathbf{w}_i\|^2 + (a-1)\ln s_i - b_i s_i\big)\Big\rangle + \text{const} \\ &= \sum_{i=1}^{G}\Big(\big(a + \frac{g_i}{2} - 1\big)\ln s_i - \big(\frac{\langle\|\mathbf{w}_i\|^2\rangle}{2} + \langle b_i\rangle\big)s_i\Big) + \text{const}\end{aligned} \tag{32}$$

This is the exponent of the product of G Gamma distributions [36]. Thus the mean of s_i for $i = 1 \ldots G$, occurs when

$$s_i^{(t+1)} = \langle s_i\rangle = \frac{g_i + 2a}{\left(\|\overline{\mu}_i^{(t)}\|^2 + \text{tr}[\overline{\Sigma}_i^{(t)}]\right) + 2b_i^{(t)}} \tag{33}$$

where $\overline{\mu}_i^{(t)}$ and $\overline{\Sigma}_i^{(t)}$ are the components of $\overline{\mu}^{(t)}$ and $\overline{\Sigma}^{(t)}$ corresponding to group $\mathbf{w}_i$, and $b_i^{(t)} = \langle b_i\rangle$ at iteration t. Note that the mean of a Gamma

Algorithm 1 VBMM-based image restoration algorithm

Inputs: parameters for the sensing matrix $\mathbf{B}$, observation $\mathbf{y}$, Λ_α, a, k, θ, β, initial estimations of $\mathbf{z}^{(0)}$, $\mathbf{s}^{(0)}$ and $\mathbf{b}^{(0)}$.
while iterations $t = 0 : t_{\max}$ or $\mathbf{w}$ has converged, **do**
  $\overline{\Sigma}^{(t)} = \left(\beta\Lambda_\alpha + \mathbf{S}^{(t)}\right)^{-1}$
  $\mathbf{w}^{(t+1)} = \beta\overline{\Sigma}^{(t)}[\Lambda_\alpha\mathbf{z}^{(t)} - \mathbf{M}^T\mathbf{B}^T(\mathbf{BMz}^{(t)} - \mathbf{y})]$
  $\mathbf{z}^{(t+1)} = \mathbf{w}^{(t+1)}$
  $s_i^{(t+1)} = \dfrac{g_i + 2a}{\left(\|\overline{\mu}_i^{(t)}\|^2 + \mathrm{tr}[\overline{\Sigma}_i^{(t)}]\right) + 2b_i^{(t)}}$ for $i = 1...G$
  $b_i^{(t+1)} = \dfrac{a+k}{s_i^{(t+1)} + \theta}$ for $i = 1...G$
end while
Output restored image $\mathbf{x} = \mathbf{Mw}$

distribution $p(s_i|a, b_i)$ is given by $\langle s_i\rangle = \frac{a}{b_i}$.

(iv) ***Optimize*** $\ln q(\mathbf{b})$ ***using (17) and (18)***

$$\begin{aligned}
&\ln q(\mathbf{b}) = \langle \ln p(\xi, \mathbf{y})\rangle_{q(\mathbf{w})q(\mathbf{z})q(\mathbf{s})} + \text{const} \\
&= \langle \ln(p(\mathbf{s}|a, \mathbf{b})p(\mathbf{b}|k, \theta))\rangle + \text{const} \\
&= \Big\langle \sum_{i=1}^{G} \big(a \ln b_i - b_i s_i + (k-1)\ln b_i - \theta b_i\big)\Big\rangle + \text{const} \\
&= \sum_{i=1}^{G} \Big(\big(a + k - 1\big)\ln b_i - \big(\langle s_i\rangle + \theta\big)b_i\Big) + \text{const} \qquad (34)
\end{aligned}$$

Thus each $q(b_i)$ is a Gamma distribution and the mean of b_i, for $i = 1 \ldots G$, occurs when

$$b_i^{(t+1)} = \langle b_i\rangle = \frac{a+k}{s_i^{(t+1)} + \theta} \qquad (35)$$

The updating procedure can be viewed as minimizing the KL divergence with respect to each of the factors $q_i(\xi)_i$ in turn. Therefore we aim to get a close approximation of $q(\xi)$ to $p(\xi|\mathbf{y})$, within the limitations of the factorized

form, (24). The key steps of VBMM-based image restoration algorithm are shown in Algorithm 1.

Suppose after t iterations b_i converges to a stationary point such that $b_i^{(t+1)} = b_i^{(t)}$, then (35) can be substituted into (33), producing

$$(\|\mathbf{w}_i^{(t+1)}\|^2)\big(s_i^{(t+1)}\big)^2 + (\theta\|\mathbf{w}_i^{(t+1)}\|^2 + 2k - g_i)\big(s_i^{(t+1)}\big) - (g_i + 2a)\theta = 0 \quad (36)$$

where $\|\mathbf{w}_i^{(t+1)}\|^2 = (\|\overline{\mu}_i^{(t)}\|^2 + \mathrm{tr}[\overline{\Sigma}_i^{(t)}])$. When $\theta \to 0$, (36) has a unique solution as

$$s_i^{(t+1)} = \frac{g_i - 2k}{\langle\|\mathbf{w}_i\|^2\rangle} \quad (37)$$

It can be seen from (37) that as long as $k < \frac{g_i}{2}$ the proposed method approximates an l_0 norm by reweighting an l_2 norm iteratively which can be regarded as a relaxation of Iterative Reweighted Least Squares (IRLS) studied in [42]. It is also noted that if $\theta = 0$, the solution does not depend on a. However, in practice θ should be set just above zero in order for $p(\mathbf{b}|k, \theta)$ in (18) to be a proper pdf. Where additional prior knowledge exists, a, θ and k can be chosen appropriately.

2.3. Estimation of Noise Variance

As discussed, we can keep the inverse noise variance β as a user parameter to adjust the regularization strength. However for some applications, noise variance is an unknown and potentially variable parameter, which must therefore be estimated. To estimate β, we first assume a Gamma distribution is imposed for inverse signal variance β ($\beta \equiv \nu^{-2}$), such that

$$p(\beta|c, d) = \frac{d^c}{\Gamma(c)}\beta^{c-1}\exp(-\beta d) \quad (38)$$

If we apply the same VB continuation strategy, we obtain that:

$$\beta^{(t+1)} = \frac{M + 2c}{\|\mathbf{y} - \mathbf{BM}\overline{\mu}^{(t)}\|^2 + \mathrm{tr}[\mathbf{M}^T\mathbf{B}^T\mathbf{BM}\overline{\Sigma}^{(t)}] + 2d} \quad (39)$$

Furthermore, if the value of Λ_α is properly set such that $\Lambda_\alpha - \mathbf{M}^T\mathbf{B}^T\mathbf{BM}$ is close to zero, the following approximation is obtained:

$$\mathrm{tr}[\mathbf{M}^T\mathbf{B}^T\mathbf{BM}\overline{\Sigma}^{(t)}] \approx \mathrm{tr}[\Lambda_\alpha\overline{\Sigma}^{(t)}] \quad (40)$$

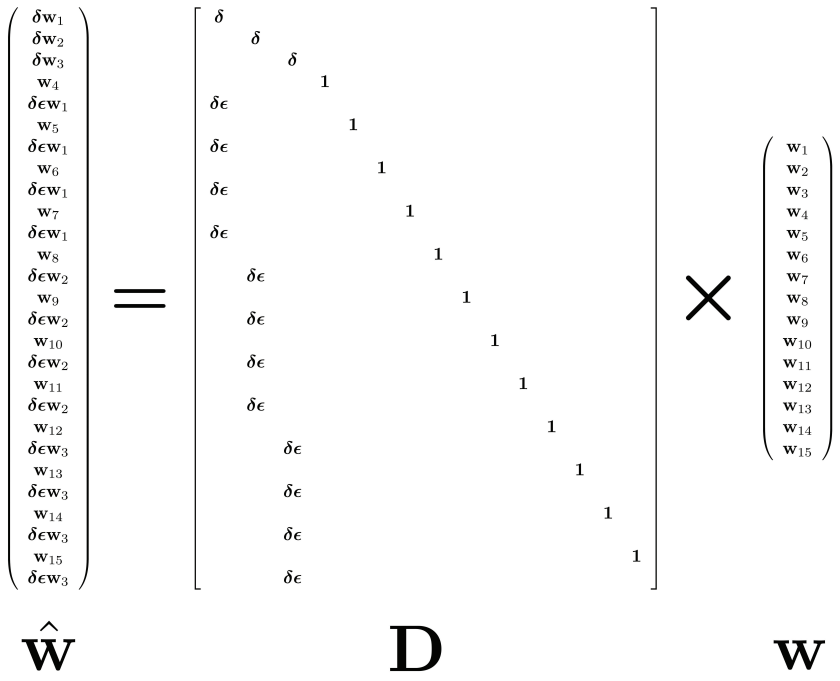


Figure 3: Simple example of the non-overlapping transformation corresponding only to levels 2 and 3 of the quadtree in Fig. 1(b), using the parent+1child grouping of Fig. 4(a), $\delta = \frac{1}{\sqrt{1+4\epsilon^2}}, \epsilon \in (0, 1]$.

This assumption is likely to be valid if most of the energy of $\mathbf{M}^T\mathbf{B}^T\mathbf{B}\mathbf{M}$ is compressed into the leading diagonal terms [35]. Accordingly, β can then be estimated as:

$$\beta^{(t+1)} = \frac{M + 2c}{\|\mathbf{y} - \mathbf{B}\mathbf{M}\overline{\mu}^{(t)}\|^2 + \mathrm{tr}[\Lambda_\alpha \overline{\Sigma}^{(t)}] + 2d} \tag{41}$$

3. Extension for Tree-structured Modeling

To achieve the goal of a fully *overlapped group sparse* solution, it is possible to incorporate the VBMM model with a wavelet tree structure as shown in [30]. Recent works have demonstrated that modeling wavelet parent-children relationships can be viewed as overlapping group regularization [29, 43]. Inspired from [43], we adopt a transformation to a non-overlapping redundant space $\hat{\mathbf{w}} = \mathbf{D}\mathbf{w}$. To encourage the persistence of large/small coefficients

across scales, where $\hat{\mathbf{w}}$ is a $P \times 1$ vector that forms groups of wavelet coefficients in the non-overlapping space, and the sparse transformation matrix $\mathbf{D}$ indicates the presence or absence of correspondence between the overlapping and non-overlapping spaces. Unlike [30], in this paper we construct the $\mathbf{D}$ matrix to be a Parseval tight frame such that $\mathbf{D}^T\mathbf{D} = \mathbf{I}$. A simple example of this non-overlapping redundant transformation is shown in Fig. 3.

In this case, we set entries of the $\mathbf{D}$ matrix that correspond to the finest-level coefficients as 1 and set entries that represent all parent coefficients as δ. Because the magnitudes of the wavelet coefficients tend to decay across scale [22], we further introduce a parameter $\epsilon \in (0, 1]$ to adjust the ratio between the magnitudes of the parent coefficient and its replicated copies, 4 copies for "parent+1child" and 1 copy for "parent+4children". The entries that correspond to the replicated parent coefficients are $\delta\epsilon$. Note that here we keep $\delta = \frac{1}{\sqrt{1+4\epsilon^2}}$ for "parent+1child" and $\delta = \frac{1}{\sqrt{1+\epsilon^2}}$ for "parent+4children" to ensure $\mathbf{D}$ is a Parseval tight frame so that $\mathbf{D}^T\mathbf{D} = \mathbf{I}$.

As a result, the likelihood of the data in (3) can be extended to be

$$p(\mathbf{y}|\hat{\mathbf{w}}, \nu^2) = \left(2\pi\nu^2\right)^{-\frac{M}{2}} \exp\{-\frac{1}{2\nu^2}\|\mathbf{y} - \mathbf{B}\mathbf{M}\mathbf{D}^T\hat{\mathbf{w}}\|^2\} \tag{42}$$

where $\hat{\mathbf{w}} = \mathbf{D}\mathbf{w}$. Then we can model $\hat{\mathbf{w}}$ using a group sparse GSM model similar to (4):

$$p\left(\hat{\mathbf{w}}|\mathbf{S}\right) = \prod_{i=1}^{G} \mathcal{N}\left(\hat{\mathbf{w}}_i|0, \sigma_i^2\right) = \mathcal{N}\left(\hat{\mathbf{w}}|0, \hat{\mathbf{S}}^{-1}\right) \tag{43}$$

Now $\hat{\mathbf{S}}$ needs to be of size $P \times P$, and when $P > G$, its diagonal is an expanded form of $\mathbf{s}$ where each $\hat{s}_i$ is repeated g_i times.

However, how best to group coefficients is not clear and becomes an important question. Here, we consider two grouping schemes: "parent+1child" and "parent+4children" as illustrated in Fig. 4. In the case of "parent+1child" scheme, the parent coefficient is grouped separately with each child coefficient, whereas for "parent+4children" scheme the parent coefficient is grouped with all 4 of its children. Note that in both cases, we group the root-level coefficients individually since they do not have a parent.

Based on Bayes' rule, the posterior distribution can be then calculated

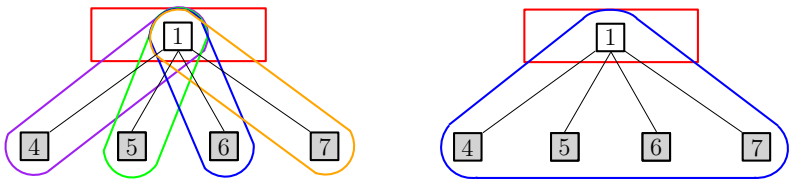


Figure 4: Illustration of different grouping strategies: (a) parent+1child, (b) parent+4children. The root-level coefficients are grouped individually shown as the red rectangles which represent singleton groups.

via:

$$p\left(\hat{\mathbf{w}}|\mathbf{y},\hat{\mathbf{S}},\nu^2\right)=\frac{p\left(\mathbf{y}|\hat{\mathbf{w}},\nu^2\right)\times p\left(\hat{\mathbf{w}}|\hat{\mathbf{S}}\right)}{p\left(\mathbf{y}|\hat{\mathbf{S}},\nu^2\right)} \tag{44}$$

Because both $p\left(\mathbf{y}|\hat{\mathbf{w}},\nu^2\right)$ and $p\left(\hat{\mathbf{w}}|\hat{\mathbf{S}}\right)$ are Gaussian functions of $\hat{\mathbf{w}}$, the posterior can be rearranged as

$$p\left(\hat{\mathbf{w}}|\mathbf{y},\hat{\mathbf{S}},\nu^2\right)=\mathcal{N}\left(\hat{\mathbf{w}}|\hat{\mu},\hat{\Sigma}\right) \tag{45}$$

with

$$\hat{\mu}=\nu^{-2}\hat{\Sigma}\mathbf{D}\mathbf{M}^T\mathbf{B}^T\mathbf{y} \tag{46}$$

$$\hat{\Sigma}=\left(\nu^{-2}\mathbf{D}\mathbf{M}^T\mathbf{B}^T\mathbf{B}\mathbf{M}\mathbf{D}^T+\hat{\mathbf{S}}\right)^{-1} \tag{47}$$

However, now the $P\times P$ square matrix $\left(\nu^{-2}\mathbf{D}\mathbf{M}^T\mathbf{B}^T\mathbf{B}\mathbf{M}\mathbf{D}^T+\hat{\mathbf{S}}\right)$needs to be inverted, which is not computationally feasible for big data sets and images. So, we again adopt the Bayesian MM framework as in (9). Here we introduce a hidden vector $\mathbf{z}$ of size N as before and the following approximation model for its posterior distribution:

$$\overline{p}\left(\hat{\mathbf{w}},\mathbf{z}|\mathbf{y},\hat{\mathbf{S}},\nu^2\right)=p\left(\mathbf{z}|\hat{\mathbf{w}}\right)\times p\left(\hat{\mathbf{w}}|\mathbf{y},\hat{\mathbf{S}},\nu^2\right) \tag{48}$$

and

$$p\left(\mathbf{z}|\hat{\mathbf{w}}\right)=\exp\{-(\hat{\mathbf{w}}-\mathbf{D}\mathbf{z})^T\mathbf{Q}(\hat{\mathbf{w}}-\mathbf{D}\mathbf{z})\} \tag{49}$$

where

$$\mathbf{Q}=\frac{\hat{\Lambda}_\alpha-\mathbf{D}\mathbf{M}^T\mathbf{B}^T\mathbf{B}\mathbf{M}\mathbf{D}^T}{2\nu^2} \tag{50}$$

Algorithm 2 Tree-structured VBMM Image Restoration Algorithm

1: **Inputs**: parameters for the sensing matrix $\mathbf{B}$, observation $\mathbf{y}$, $\hat{\Lambda}_\alpha$, a, β, k, θ, initial estimations of $\mathbf{z}^{(0)}$, $\hat{\mathbf{s}}^{(0)}$ and $\hat{\mathbf{b}}^{(0)}$.
2: **while** iterations $t = 0 : t_{\max}$ or $\mathbf{z}$ has converged, **do**
3: $\hat{\overline{\Sigma}}^{(t)} = \left(\beta\hat{\Lambda}_\alpha + \hat{\mathbf{S}}^{(t)}\right)^{-1}$
4: $\hat{\overline{\mu}}^{(t)} = \beta\hat{\overline{\Sigma}}^{(t)}[\hat{\Lambda}_\alpha\mathbf{D}\mathbf{z}^{(t)} - \mathbf{D}\mathbf{M}^T\mathbf{B}^T(\mathbf{B}\mathbf{M}\mathbf{z}^{(t)} - \mathbf{y})]$
5: $\hat{\mathbf{w}}^{(t+1)} = \hat{\overline{\mu}}^{(t)}$
6: $\mathbf{z}^{(t+1)} = \mathbf{D}^T\hat{\mathbf{w}}^{(t+1)}$
7: $\hat{s}_i^{(t+1)} = \dfrac{g_i + 2a}{\left(\|\hat{\overline{\mu}}_i^{(t)}\|^2 + \mathrm{tr}[\hat{\overline{\Sigma}}_i^{(t)}]\right) + 2\hat{b}_i^{(t)}}$ for $i = 1...G$
8: $\hat{b}_i^{(t+1)} = \dfrac{a + k}{\hat{s}_i^{(t+1)} + \theta}$ for $i = 1...G$
9: **end while**
10: **Output** restored image $\mathbf{x} = \mathbf{M}\mathbf{z}^{(t+1)}$

We find that $\mathbf{Q}$ should be positive definite to ensure convergence and hence:

$$\hat{\Lambda}_\alpha \succ \mathbf{D}\mathbf{M}^T\mathbf{B}^T\mathbf{B}\mathbf{M}\mathbf{D}^T \tag{51}$$

which is equivalent to requiring that:

$$\begin{aligned}\mathbf{D}^T\hat{\Lambda}_\alpha\mathbf{D} &\succ \mathbf{D}^T\mathbf{D}\mathbf{M}^T\mathbf{B}^T\mathbf{B}\mathbf{M}\mathbf{D}^T\mathbf{D} \\ &= \mathbf{M}^T\mathbf{B}^T\mathbf{B}\mathbf{M}\end{aligned} \tag{52}$$

where we assume $\Lambda_\alpha = \mathbf{D}^T\hat{\Lambda}_\alpha\mathbf{D}$. It is known that we need $\Lambda_\alpha \succ \mathbf{M}^T\mathbf{B}^T\mathbf{B}\mathbf{M}$ in order to fulfill the condition. Λ_α can be found as before using the methods discussed in Section 2.1 and Appendix 6.1. We can easily compute the $\tilde{\Lambda}_\alpha$ once Λ_α is found, as it is equivalent to applying the non-overlapping transformation to Λ_α. Because $p(\mathbf{z}|\hat{\mathbf{w}})$ and $p\left(\hat{\mathbf{w}}|\mathbf{y}, \hat{\mathbf{S}}, \nu^2\right)$ are Gaussian functions of $\hat{\mathbf{w}}$, when $\mathbf{z}$ is given (typically as a previous estimate for $\mathbf{w}$), the approximation model can be rearranged into a Gaussian form as

Table 1: BLUR, Noise Variance and BSNR (dB)

Exp.	BLUR	ν^2	BSNR
1	9×9 uniform	31.10	20
2	9×9 uniform	0.31	40
3	9×9 uniform	0.03	50
4	$h_{ij} = 1/(1 + i^2 + j^2), i, j = -7, \ldots, 7$	2	31.85
5	$h_{ij} = 1/(1 + i^2 + j^2), i, j = -7, \ldots, 7$	8	25.85

$$\overline{p}\left(\hat{\mathbf{w}}|\mathbf{y}, \mathbf{z}, \hat{\mathbf{S}}, \nu^2\right) = \mathcal{N}\left(\hat{\mathbf{w}}|\hat{\overline{\mu}}, \hat{\overline{\Sigma}}\right) \tag{53}$$

with

$$\hat{\overline{\mu}} = \nu^{-2}\hat{\overline{\Sigma}}[\hat{\Lambda}_\alpha \mathbf{D}\mathbf{z} - \mathbf{D}\mathbf{M}^T\mathbf{B}^T(\mathbf{B}\mathbf{M}\mathbf{z} - \mathbf{y})] \tag{54}$$

$$\hat{\overline{\Sigma}} = (\nu^{-2}\hat{\Lambda}_\alpha + \hat{\mathbf{S}})^{-1} \tag{55}$$

Now (54) and (55) are computationally tractable, since $\hat{\overline{\Sigma}}$ is diagonal. Then we can obtain the tree-structured VBMM Algorithm in Algorithm 2. This uses the same VB continuation strategy as described in Section 2.2 where we impose Gamma distributions for both inverse signal variance $\hat{\mathbf{s}}$ and its rate parameter $\hat{\mathbf{b}}$:

$$p(\hat{\mathbf{s}}|a, \hat{\mathbf{b}}) = \prod_{i=1}^{G} \frac{\hat{b}_i^a}{\Gamma(a)} \hat{s}_i^{a-1} \exp(-\hat{b}_i \hat{s}_i) \tag{56}$$

and

$$p(\hat{\mathbf{b}}|k, \theta) = \prod_{i=1}^{G} \frac{\theta^k}{\Gamma(k)} \hat{b}_i^{k-1} \exp(-\theta \hat{b}_i) \tag{57}$$

This extension effectively provides a framework which incorporates a wavelet tree structure in a variational Bayesian derivation. Note that G now represents the number of (overlapping) groups of complex coefficients, and the g_i are in general larger than before so that now $P = \sum_{i=1}^{G} g_i$.

4. Results

We present a set of experiments to evaluate the performance of these proposed image restoration algorithms. Three applications including image deconvolution, image superresolution and compressive sensing (CS) magnetic resonance imaging (MRI) reconstruction are studied. We have used both

the coefficient-sparse VBMM (VC) in Algorithm 1 and the tree-structured VBMM in Algorithm 2 with "parent+1child" grouping (V1) and "parent+ 4children" grouping (V4) for all these experiments. The DT $\mathbb{C}$WT is chosen as our redundant sparsifying transform because it has good sparsity inducing properties. Since DT $\mathbb{C}$WT produces complex coefficients with circularly symmetric pdfs, 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 VC, $G = \frac{P+6}{4}$ groups for V1 and $G = \frac{P+6}{10}$ groups for V4, where $P = 4(MN - \frac{MN}{4^{lev}})$ for both V1 and V4. In all experiments, we set the level of decomposition $lev = 4$. In all experiments, we set $\epsilon = 1$ in the $\mathbf{D}$ matrix for simplicity. However, as stated in Section 3, varying ϵ may further improve the results.

4.1. Image Deconvolution

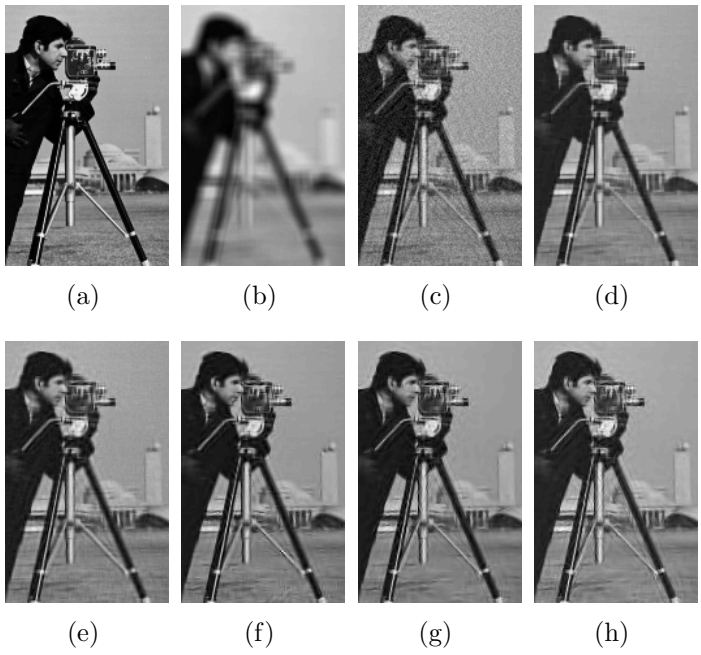


Figure 5: Deconvolution results after 200 iterations on Exp. 2, on Cameraman, BSNR: 40 dB. **(a)** Original. **(b)** Blurred. **(c)** Wiener filter, ISNR=4.512 dB. **(d)** MLTL, ISNR=7.594 dB. **(e)** MSIST, ISNR=7.648 dB. **(f)** VBMM–VC, ISNR=8.127 dB. **(g)** VBMM–V1, ISNR=8.221 dB. **(h)** VBMM–V4, ISNR=8.318 dB.

For image deconvolution, the linear operator $\mathbf{B}$ becomes a convolution

matrix. We compare and show that VBMM and its tree-structured extensions outperform recently developed SIST algorithms including multilevel thresholded Landweber (MLTL) [32] and modified subband-adaptive iterative shrinkage/thresholding (MSIST) [34]. Here we assume $\mathbf{B}$ is known but it is clear that Bayesian inspired schemes can also be applied for blind deconvolution where both blur kernel estimation and image restoration are performed [44].

Table 2: Average ISNR (dB) results for Exp. 1-Exp. 5 over 30 noise realizations.

iterations		10	30	50	70	100
Exp.1	MLTL	**3.054**	3.179	3.197	3.223	3.227
	MSIST	2.584	2.990	3.191	3.308	3.403
	VBMM-VC	2.731	3.282	3.491	3.582	3.646
	VBMM-V1	2.953	3.554	3.666	3.710	3.738
	VBMM-V4	2.957	**3.618**	**3.715**	**3.744**	**3.758**
Exp.2	MLTL	6.818	7.292	7.393	7.477	7.569
	MSIST	7.086	7.410	7.516	7.571	7.613
	VBMM-VC	7.202	7.656	7.857	7.967	8.061
	VBMM-V1	**7.630**	7.993	8.109	8.163	8.200
	VBMM-V4	7.568	**8.012**	**8.138**	**8.197**	**8.242**
Exp.3	MLTL	7.825	9.531	10.256	10.565	10.871
	MSIST	8.760	10.290	10.601	10.669	10.683
	VBMM-VC	10.148	10.656	10.879	10.996	11.085
	VBMM-V1	10.167	10.749	10.939	11.042	11.137
	VBMM-V4	**10.191**	**10.861**	**11.059**	**11.164**	**11.259**
Exp.4	MLTL	7.095	7.136	7.154	7.160	7.256
	MSIST	6.888	6.890	6.890	6.891	6.891
	VBMM-VC	6.769	7.111	7.190	7.211	7.216
	VBMM-V1	7.214	7.220	7.226	7.231	7.236
	VBMM-V4	**7.242**	**7.325**	**7.326**	**7.327**	**7.330**
Exp.5	MLTL	4.891	4.921	4.954	5.041	5.086
	MSIST	4.842	4.854	4.855	4.855	4.855
	VBMM-VC	4.784	5.192	5.259	5.277	5.281
	VBMM-V1	**5.387**	5.456	5.467	5.473	5.477
	VBMM-V4	5.202	**5.536**	**5.562**	**5.570**	**5.574**

An important issue of implementing VBMM is whether or not to use sparsity-by-analysis algorithms where we project $\mathbf{w}$ into the range space of $\mathbf{M}^T$ every iteration before computing $\mathbf{z}$ [45]. This is due to the fact that the sparsity-by-synthesis algorithm seems not to converge towards the ground-truth solution as shown in Fig. 6 (b).

Because the reconstructed image $\mathbf{x} = \mathbf{M}\mathbf{w}$ is only a function of the range-space component of $\mathbf{w}$, we can eliminate the null-space component of $\mathbf{w}$ by replacing $\mathbf{w}$ by $\mathbf{M}^T\mathbf{x}$. When we do this, we obtain results which converge closer to the ground-truth solution and where the ISNR is non-decreasing. This is equivalent to regularization based on the sparsity of $\mathbf{M}^T\mathbf{M}\mathbf{w} = \mathbf{P}\mathbf{w}$

rather than based on the sparsity of $\mathbf{w}$ where $\mathbf{P} = \mathbf{M}^T\mathbf{M}$ is the range-space projection matrix. By encouraging sparsity of $\mathbf{Pw}$, rather than $\mathbf{w}$, we reduce visually unpleasant artifacts in the final solution. Therefore we use the sparsity-by-analysis algorithm for the remainder of our results.

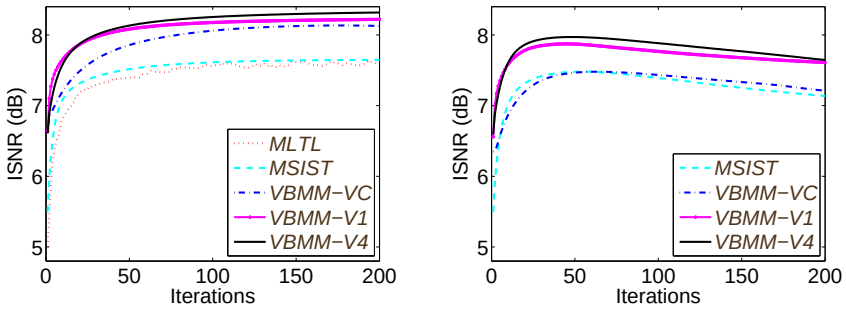


Figure 6: ISNR results using (a) analysis algorithm (b) synthesis algorithm, on Cameraman, 9×9 uniform blur, BSNR=40 dB.

Five experiments were performed as shown in Table 1, where we convolved the Cameraman image with two different blur kernels: 9×9 uniform blur (Exp. 1, 2, 3) and 15×15 cylindrical blur as $h_{ij} = \frac{1}{1+i^2+j^2}, i, j = -7, \ldots, 7$ (Exp. 4, 5). 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) $=10\log_{10}(\frac{\|\mathbf{y}-\mathbf{x}_r\|^2}{\|\mathbf{M}\mathbf{z}-\mathbf{x}_r\|^2})$ was used to evaluate the relative performance of different algorithms, where $\mathbf{z}$ is the current estimate of wavelet coefficients $\mathbf{w}$. An under-regularized Wiener filter $\mathbf{x}_0=(\mathbf{H}^T\mathbf{H}+10^{-3}\nu^2\mathbf{I})^{-1}\mathbf{H}^T\mathbf{y}$ was used to estimate the initial $\mathbf{x}_0$ and hence $\mathbf{z}^{(0)} = \mathbf{M}^T\mathbf{x}_0$ [34]. In the experiment, we set hyperparameters $a = \theta = 10^{-6}$ and adjusted k to control the sparsity where k should satisfy $0 < k < \frac{g_i}{2}$. We've ensured the matrix $\mathbf{\Lambda}_\alpha$ for the VC, V1, and V4 experiments was the same for each test scenario.

Deconvolution results are shown in Fig. 5 for visual comparison. It can be seen (after enlargement) that fewer artifacts are observed using the proposed VBMM method and its tree-structured extensions. Table 2 summarizes the ISNR values obtained from Exp. 1-Exp. 5 over 30 noise realizations. It is found that incorporating VBMM with the group sparse penalty, particularly

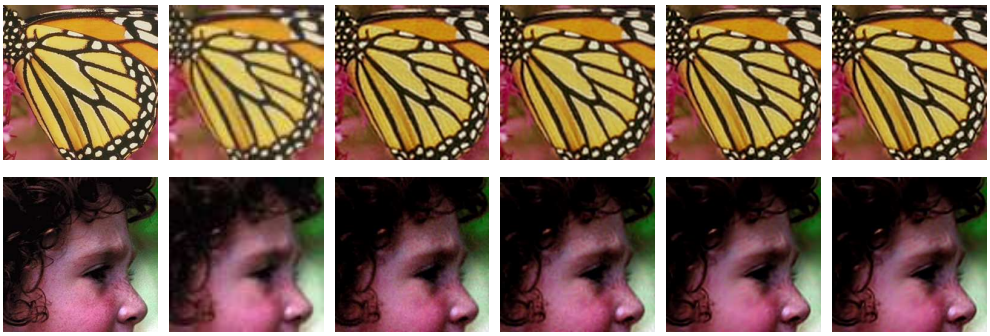


Figure 7: Superresolution results for Butterfly and Girl image, downsampled by a factor of 3 with $\nu = 5$. Left to right: original image, LR image, MSIST [46], VBMM-VC, VBMM-V1 and VBMM-V4.

the "parent+4children" grouping strategy, leads to a faster convergence rate and better ISNR results. Fig. 6 (a) compares the ISNR results of Exp. 2 obtained using different image deconvolution techniques over 200 iterations.

4.2. Image Superresolution

In the case of image superresolution, the linear operator $\mathbf{B}$ becomes $\mathbf{TH}$, where $\mathbf{T}$ is a matrix that represents the subsampling operation and $\mathbf{H}$ is the blurring matrix. We compare our VBMM method and its extensions with MSIST-based image superresolution algorithm as described in [46]. In the experiments, seven test images including Butterfly, Flower, Girl, Hat, Leaves, Parrot and Bike were used[1].

We convolved the test images with a 7×7 Gaussian kernel with standard deviation (std.) $\sigma_h = 1.6$ and down-sample them with a factor of 3 in each direction. White Gaussian noise was then added to generate noisy low resolution (LR) images, and two noise levels $\nu = 1$ (average BSNR=34.280 dB) and $\nu = 5$ (average BSNR=20.312 dB) were tested. The peak signal-to-noise ratio (PSNR) $=10 \log_{10}(\frac{\max\{\mathbf{x_r}\}^2}{\|\mathbf{Mz}-\mathbf{x}_r\|^2})$ and the structural similarity (SSIM) index [49] were used to evaluate the relative performance of different algorithms. We performed the power iteration method in (62) to determine the matrix Λ_α. The inverse noise variance β was estimated in this case using (41). We set hyperparameters $a = k = \theta = d = 10^{-6}$, and used c (typically around 10^5 to be comparable with $M = 2^{16}$) to control the strength of sparsity.

[1]Available online at http://www-sigproc.eng.cam.ac.uk/Main/GZ243.

Table 3: PSNR (dB) and SSIM results of image superresolution, after 50 iterations

Noise std.		$\nu = 1$				$\nu = 5$			
Picture	Measure	MSIST	VC	V1	V4	MSIST	VC	V1	V4
butterfly	PSNR	23.574	23.592	23.791	**23.820**	22.731	22.760	23.203	**23.214**
	SSIM	0.8300	0.8351	0.8482	**0.8483**	0.7516	0.8053	0.8119	**0.8121**
flower	PSNR	27.019	27.136	27.272	**27.291**	25.772	26.064	26.277	**26.333**
	SSIM	0.8135	0.8197	0.8270	**0.8271**	0.7317	0.7633	0.7692	**0.7731**
girl	PSNR	32.831	33.103	33.077	**33.134**	30.526	31.017	31.395	**31.408**
	SSIM	0.8495	0.8568	0.8561	**0.8569**	0.7721	0.7895	0.7976	**0.7982**
hat	PSNR	28.498	28.549	28.761	**28.770**	26.993	27.377	27.850	**27.879**
	SSIM	0.8286	0.8379	0.8444	**0.8448**	0.7659	0.7925	**0.8013**	0.8008
leaves	PSNR	23.376	23.536	23.679	**23.700**	22.387	22.694	23.034	**23.082**
	SSIM	0.8279	0.8322	0.8492	**0.8495**	0.7516	0.7946	0.8115	**0.8117**
parrots	PSNR	27.882	28.010	28.106	**28.156**	25.913	26.774	27.438	**27.494**
	SSIM	0.8845	0.8960	0.8993	**0.8995**	0.8212	0.8575	0.8681	**0.8683**
bike	PSNR	22.278	22.302	22.541	**22.553**	21.648	21.751	22.016	**22.033**
	SSIM	0.7440	0.7443	0.7562	**0.7584**	0.6743	0.6921	0.7097	**0.7098**

Table 3 summarizes the PSNR and SSIM results after 50 iterations. Selected superresolution results with $\nu = 5$ are shown in Fig. 7 for visual comparison. It is shown that in most cases, VBMM with "parent+4children" grouping leads to the best results in terms of both PSNR value and SSIM index. We acknowledge that current state-of-the art algorithms can give better superresolution or deconvolution results than our methods. This is mainly due to the fact that they employ dictionary learning methods [47, 48], which are computationally demanding. We have not included them as this topic is beyond the scope of this paper.

4.3. MRI Image Reconstruction

In this section, we show the performance of our proposed methods on CS MRI Reconstruction. In the MRI imaging problem, $\mathbf{B} = \mathbf{F}$ represents a partial n-point Fourier transform with M rows and n columns. The sampling rate is controlled by $M/n \leq 1$, and the scanning time is reduced as this

Figure 8: Test images: Brain; Chest; Shoulder; Cardiac and the sampling mask with sampling rate 20%.

ratio becomes smaller. We compare our VBMM algorithm and its extensions with MSIST-based MRI and a recently developed CS MRI reconstruction algorithm Wavelet Tree Sparsity MRI (WatMRI) [43].

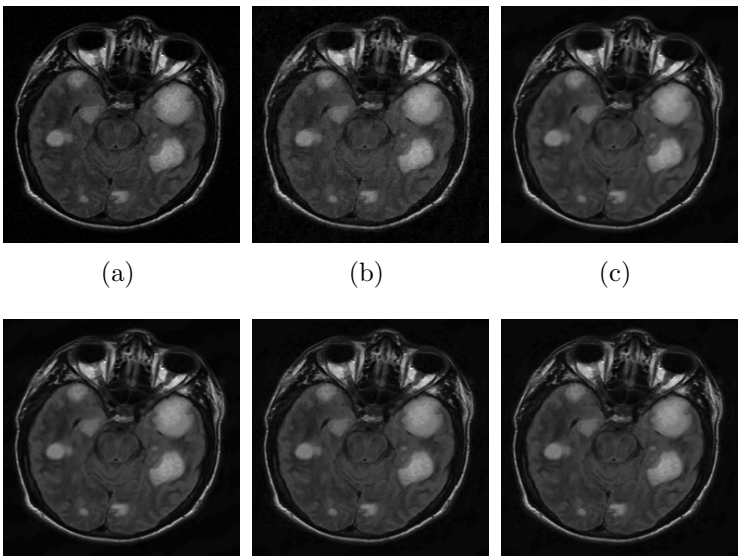


Figure 9: CS MRI reconstruction of Brain image after 100 iterations. (a) original image. (b) WatMRI, SNR: 17.443 dB. (c) MSIST, SNR: 20.355 dB. (d) VBMM-VC, SNR: 20.988 dB. (e) VBMM-V1, SNR: 21.866 dB. (f) VBMM-V4, SNR: 21.969 dB.

Following the same experimental setting as in [43], we have randomly chosen more Fourier coefficients from low frequencies and less from high frequencies with a mean sampling rate of 20 %. Four test images including brain, chest, shoulder and cardiac (Fig. 8) were used. For all test cases, white Gaussian noise with std. $\nu = 0.1$ (average BSNR=60.461 dB) was

Table 4: SNR (dB) results of CS MRI reconstruction after 100 iterations, noise std. $\nu = 0.1$, sampling rate is 0.2.

SNR (dB)	WatMRI	MSIST	VBMM-VC	VBMM-V1	VBMM-V4
Brain	17.443	20.355	20.988	21.866	**21.969**
Chest	16.588	20.044	21.244	21.787	**21.917**
Shoulder	22.193	24.181	25.439	27.654	**27.743**
Cardiac	18.701	19.753	20.344	21.277	**21.407**

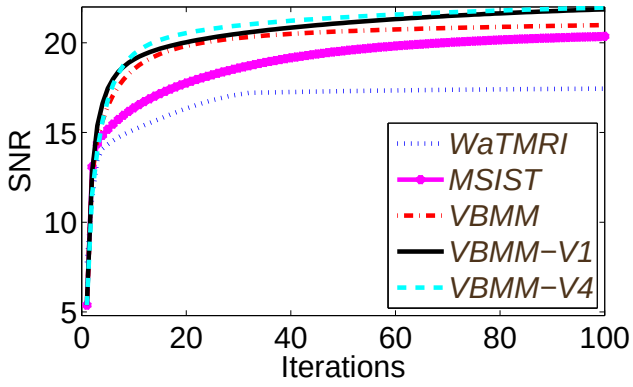


Figure 10: SNR results for CS MRI reconstruction over 100 iterations, on Chest Image.

added. The signal-to- noise ratio (SNR) was chosen to determine the relative performance. We performed the power iteration method in (62) to determine the matrix Λ_α, and estimated the inverse noise variance β using (41). In the experiments, we set hyperparameters $a = k = \theta = d = 10^{-6}$, and used c (typically around 10^5) to control the strength of sparsity. Table 4 shows the SNR results of CS MRI reconstruction where we let all algorithms terminate after 100 iterations. The reconstruction results of Brain image are shown in Fig. 9 for visual comparison and the convergence plot is shown in Fig. 10. It is shown that similar to image superresolution, VBMM with "parent+4children" grouping leads to a faster convergence rate and better SNR results.

From the above three applications, we find that the VBMM method with its tree-structured extensions are effective for various image restoration problems. In particular, VBMM-V4 demonstrates a more robust performance compared with VBMM-VC and VBMM-V1 in most cases. Although we have mainly discussed three applications in this paper, our method is likely to be applicable to many other application areas.

5. Conclusion

The Sparse Bayesian Learning (SBL) framework is a powerful approach for exploring sparse solutions to large inverse problems, and it can be applied effectively to Gaussian scale mixture models which are often employed to model wavelet coefficients. This motivates the use of SBL methods for wavelet based regularization. In this paper, we propose a wavelet-regularized

image restoration algorithm (VBMM) which is a combination of hierarchical Bayesian estimation, using variational Bayesian (VB) inference, with a subband-adaptive majorization minimization (MM) method for efficiently finding maxima of posterior distributions. In addition, we show that VBMM can incorporate tree-structured group-sparse models, that are appropriate for tight-frame (redundant) wavelet transforms. We then consider three image restoration problems including image deconvolution, image superresolution and compressive sensing MRI reconstruction to demonstrate the good performance of our proposed methods.

6. Appendices

6.1. Selection of Gain Coefficients in Λ_α (Sections 2.1 and 3)

One of the keys to the SIST algorithms is the determination of the Λ_α matrix so that the surrogate function majorizes the original objective. To be specific, α_j must be carefully chosen for each subband j to be the smallest α_j which ensures $\bar{J}_\alpha(\mathbf{w}, \mathbf{z}) > J(\mathbf{w})$ in (11) for all $\mathbf{z} \neq \mathbf{w}$.

6.1.1. Subband Adaptive Thresholding for Blurring Kernel

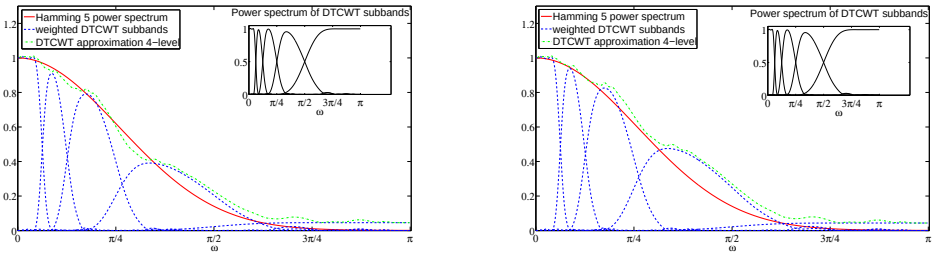
Various subband-adaptive methods can be applied to determine the diagonal entry α_j for j^{th} subband coefficients [32, 33, 34]. In general, the design of a subband adaptive algorithm for a deconvolution problem aims to approximate the blurring kernel $\mathbf{H}$ based on the following bound [34]:

$$\mathbf{u}^T(\Lambda_\alpha - \mathbf{M}^T\mathbf{H}^T\mathbf{H}\mathbf{M})\mathbf{u} > 0 \tag{58}$$

where $\mathbf{B} = \mathbf{H}$ in this case, and the inequality holds for any arbitrary vector $\mathbf{u} \neq 0$. When the forward wavelet transform $\mathbf{W}$ is an orthonormal basis, the following approximation is used to calculate Λ_α [32, 33, 34]:

$$\alpha_j = \rho(\mathbf{M}_j^T\mathbf{H}^T\mathbf{H}\mathbf{M}_j) \tag{59}$$

where ρ denotes the spectral radius and $\mathbf{M}_j$ denotes the inverse wavelet transform in the given subband j. However, (59) typically fails for wavelets that have significant leakage across subbands. To obtain a satisfactory approximation, we then need to increase the value of α_j in (59) to account for the energy leakage. Systematic ways of achieving this are reported in [33] and [34].



(a) (b)

Figure 11: Weighted power spectrum of 1-D DT ℂWT subbands, (a) selection of α_j using (59), (b) α_j are increased to account for energy leakage.

In Fig. 11, we illustrate the effect of the subband-adaptive MM technique for DT ℂWT. In this case, 'Hamming (5)' is chosen as the blurring kernel. Because

$$\mathbf{w}^T\Lambda_\alpha\mathbf{w} - \mathbf{w}^T\mathbf{M}^T\mathbf{H}^T\mathbf{H}\mathbf{M}\mathbf{w} = \mathbf{x}^T(\sum_j \alpha_j\mathbf{W}_j^T\mathbf{W}_j)\mathbf{x} - \mathbf{x}^T\mathbf{H}^T\mathbf{H}\mathbf{x} \qquad (60)$$

where $\mathbf{W}_j$ denotes the forward wavelet transform in the given subband j, we can plot $\alpha_j\mathbf{W}_j^T\mathbf{W}_j$ as the power spectrum of DT ℂWT subbands weighted by α_j and compare it with the power spectrum of $\mathbf{H}^T\mathbf{H}$. In Fig. 11 (a), α_j is calculated for each subband of the DT ℂWT using (59). It is shown that part of the weighted power spectrum of DT ℂWT subbands is below 'Hamming (5)' power spectrum due to the crossband energy leakage. This makes $\Lambda_\alpha - \mathbf{M}^T\mathbf{H}^T\mathbf{H}\mathbf{M} < 0$ which will affect the convergence. To solve the problem, we can increase α_j so that a safe upper bound to the power spectrum of $\mathbf{H}^T\mathbf{H}$ is provided as shown in Fig. 11 (b).

6.1.2. Subband Adaptive Thresholding for Non-Blurring Kernel

Note that the techniques discussed above are mainly only applicable for approximating blurring kernels. When $\mathbf{B}$ is not a convolution matrix, $\mathbf{M}^T\mathbf{B}^T\mathbf{B}\mathbf{M}$ may not be near diagonal and cannot be simply computed in the frequency domain. In this case, we need to use the power iteration method to determine the spectral radius (largest eigenvalue) of $\mathbf{M}_j^T\mathbf{B}^T\mathbf{B}\mathbf{M}_j$.

From the Rayleigh quotient, we know that if a nonzero vector $x \in \mathbb{R}^n$ is an eigenvector for $A \in \mathbb{R}^{n\times n}$ with corresponding eigenvalue λ, we have [50]:

$$r(x) = \frac{x^T Ax}{x^T x} = \frac{\lambda x^T x}{x^T x} = \lambda \tag{61}$$

where $r(x)$ denotes the Rayleigh quotient of x whose value actually minimizes the function $f(\alpha) = \|\alpha x - Ax\|_2$ over all real numbers α. Therefore, if x is treated as an eigenvector of A, we obtain the best estimate for an eigenvalue of A.

In practice, we use the power iteration method to find the eigenvector which corresponds to the dominant eigenvalue. This is suitable for subband-adaptive thresholding, since we only need to compute the largest eigenvalue of $\mathbf{M}_j^T \mathbf{B}^T \mathbf{B} \mathbf{M}_j$. The method of power iteration can be stated as [50]:

$$v^k = \frac{Av^{k-1}}{\|Av^{k-1}\|} \qquad \forall \;\; k = 1 \ldots K \tag{62}$$

with an initialization v^0 as a random vector and $\|v^k\| = 1$. Vector v^k converges to the dominant eigenvector after a number of iterations as all other eigenvector components have iteration gain < 1 and thus decay to zero. Then we can use the Rayleigh quotient to find the corresponding dominant eigenvalue.

6.2. Maximize $\mathcal{L}(q(\xi))$ (Section 2.2)

To maximize $\mathcal{L}(q(\xi))$ in (23) with respect to each of the factors in turn, let us firstly consider a simplified case where $q(\xi)$ is factorized into two disjoint groups as $q_1(\xi_1)$ and $q_2(\xi_2)$. Substituting (24) into (21) with two disjoint groups, we obtain two steps:

(i) Consider maximizing $\mathcal{L}(q(\xi))$ with respect to $q_1(\xi_1)$.

$$\mathcal{L}(q(\xi)) = \int q_1(\xi_1) \ln \tilde{p}(\xi_1, \mathbf{y}) d\xi_1 - \int q_1(\xi_1) \ln q_1(\xi_1) d\xi_1 + \text{const} \tag{63}$$

where

$$\ln \tilde{p}(\xi_1, \mathbf{y}) = \langle \ln p(\xi, \mathbf{y}) \rangle_{\xi_2} + \text{const} \tag{64}$$

and

$$\langle \ln p(\xi, \mathbf{y}) \rangle_{\xi_2} = \int \ln p(\xi_1, \xi_2, \mathbf{y}) q_2(\xi_2) d\xi_2 \tag{65}$$

is the expectation with respect to $q(\xi)$ over ξ_2.

When $\ln q_1(\xi_1) = \ln \tilde{p}(\xi_1, \mathbf{y})$, $\mathcal{L}(q(\xi))$ is maximized with respect to $q_1(\xi_1)$.

(ii) Similarly if we consider maximizing $\mathcal{L}(q(\xi))$ with respect to $q_2(\xi_2)$.

$$\mathcal{L}(q(\xi)) = \int q_2(\xi_2) \ln \tilde{p}(\xi_2, \mathbf{y}) d\xi_2 - \int q_2(\xi_2) \ln q_2(\xi_2) d\xi_2 + \text{const} \tag{66}$$

where

$$\ln \tilde{p}(\xi_2, \mathbf{y}) = \langle \ln p(\xi, \mathbf{y}) \rangle_{\xi_1} + \text{const} \tag{67}$$

and

$$\langle \ln p(\xi, \mathbf{y}) \rangle_{\xi_1} = \int \ln p(\xi_1, \xi_2, \mathbf{y}) q_1(\xi_1) d\xi_1 \tag{68}$$

is the expectation with respect to $q(\xi)$ over ξ_1.

When $\ln q_2(\xi_2) = \ln \tilde{p}(\xi_2, \mathbf{y})$, $\mathcal{L}(q(\xi))$ is maximized with respect to $q_2(\xi_2)$.

By sequentially updating these two steps, the KL divergence will then be minimized. Now consider the general case where $q(\xi)$ can be factorized into more than two groups. It can be shown that for each factor $q_j(\xi_j)$,

$$\mathcal{L}(q(\xi)) = \int q_j(\xi_j) \ln \tilde{p}(\xi_j, \mathbf{y}) d\xi_j - \int q_j(\xi_j) \ln q_j(\xi_j) d\xi_j + \text{const} \tag{69}$$

where $\ln \tilde{p}(\xi, \mathbf{y}) = \langle \ln p(\xi, \mathbf{y}) \rangle_{q(\xi \setminus \xi_j)} + \text{const}$ and $\langle \ln p(\xi, \mathbf{y}) \rangle_{q(\xi \setminus \xi_j)}$ denotes the expectation of $\ln p(\xi_j, \mathbf{y})$ with respect to $q(\xi)$ over all variables ξ_i for $i \neq j$ [40]. $\mathcal{L}(q(\xi))$ is maximized when $\ln q_j(\xi_j) = \ln \tilde{p}(\xi_j, \mathbf{y})$.

Although further investigations on the convergence are required, it is known that convergence to at least a local maximum of $\mathcal{L}(q(\xi))$ is guaranteed because minimizing the KL-divergence from a prior distribution q is a convex optimization problem and the bound is convex with respect to each of the factors [41, 40]. The proof is straightforward. Since $\ln(x) \leq x - 1$ for any $x > 0$, there exists the following inequality for any valid pdfs $q_a(\xi)$ and $q_b(\xi)$ and for any $\eta \in [0, 1]$,

$$\eta \mathrm{KL}(q_a(\xi)) + (1 - \eta) \mathrm{KL}(q_b(\xi)) \geq \mathrm{KL}\big(\eta q_a(\xi) + (1 - \eta) q_b(\xi)\big) \tag{70}$$

7. Reference

[1] G. Wang and J. Zhang and G. Pan, "Solution of inverse problems in image processing by wavelet expansion", *IEEE Trans. on Image Process.*, vol. 4, pp. 579–593, 1995.

[2] M. Afonso and J. Bioucas-Dias and M. Figueiredo, "Fast image recovery using variable splitting and constrained optimization", *IEEE Trans. on Image Process.*, vol. 19, pp. 2345–2356, 2010.

[3] J. Oliveira, J. Bioucas-Dias and M. Figueiredo, "Adaptive total variation image deblurring: A majorization–minimization approach", *Signal Process.*, vol. 89, pp. 1683-1693, 2009.

[4] M. Figueiredo and R. Nowak, "An EM algorithm for wavelet-based image restoration", *IEEE Trans. Image Process.*, vol. 12, pp. 906-916, 2003.

[5] 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, pp. 937-951, 2006.

[6] 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.

[7] R. Coifman, and D. Donoho, *Translation-invariant de-noising*, Springer, 1995.

[8] R. Neelamani, and H. Choi, and R. Baraniuk, "ForWaRD: Fourier-wavelet regularized deconvolution for ill-conditioned systems", *IEEE Trans. on Signal Process.*, vol. 52, pp. 418-433, 2004.

[9] M. Vetterli, "Wavelets, approximation, and compression", *IEEE Signal Process. Mag.*, vol. 18, pp. 59-73, 2011.

[10] T. Chang and C. Kuo, "Texture analysis and classification with tree-structured wavelet transform", *IEEE Trans. on Image Process.*, vol. 2, pp. 429-441, 1993.

[11] N. Nguyen and P. Milanfar, "An efficient wavelet-based algorithm for image superresolution", *IEEE Proc. Int. Conf. Image Process.*, vol. 2, pp. 351-354, 2000.

[12] M. Shensa, "The discrete wavelet transform: wedding the a trous and Mallat algorithms", *IEEE Trans. on Signal Process.*, vol. 40, pp. 2464-2482, 1992.

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

[14] E. Candes, L. Demanet, D. Donoho and L. Ying, "Fast discrete curvelet transforms", *Multiscale Modeling & Simulation*, vol. 5, pp. 861-899, 2006.

[15] M. Do and M. Vetterli, "The contourlet transform: an efficient directional multiresolution image representation", *IEEE Trans. on Image Process.*, vol. 14, pp. 2091-2106, 2005.

[16] 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.

[17] S. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 11, pp. 674-693, 1989.

[18] J. Fadili and L. Boubchir, "Analytical form for a Bayesian wavelet estimator of images using the Bessel K form densities", *IEEE Trans. on Image Process.*, vol. 14, pp. 231-240, 2005.

[19] M. Miller and N. Kingsbury, "Image denoising using derotated complex wavelet coefficients", *IEEE Trans. on Image Process.*, vol. 17, pp. 1500-1511, 2008.

[20] 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.

[21] 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.

[22] 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.

[23] Y. Rakvongthai, A. Vo and S. Oraintara, "Complex Gaussian scale mixtures of complex wavelet coefficients", *IEEE Trans. on Signal Process.*, vol. 58, pp. 3545-3556, 2010.

[24] M. Tipping, "Bayesian inference: An introduction to principles and practice in machine learning", *advanced lectures on machine Learning*, pp. 41-62, 2004.

[25] J. Palmer, D. Wipf, K. Kreutz-delgado and B. Rao, "Variational EM algorithms for non-Gaussian latent variable models", *Advances in Neural Information Processing Systems 18*, pp. 1059-1066, 2006.

[26] 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.

[27] Z. Zhang and B. Rao, "Extension of SBL algorithms for the recovery of block sparse signals with intra-block correlation", *IEEE Trans. on Signal Process.*, vol. 61, pp. 2009-2015, 2013.

[28] S. Babacan, S. Nakajima and M. Do, "Bayesian group-sparse modeling and variational inference", *IEEE Trans. on Signal Process.*, vol. 62, pp. 2906-2921, 2014.

[29] N. Rao, R. Nowak, S. Wright and N. Kingsbury, "Convex approaches to model wavelet sparsity patterns", *Proc. IEEE ICIP 2011*, pp. 1917-1920, 2011.

[30] G. Zhang, T. Roberts, N. Kingsbury, "Image Deconvolution Using Tree-Structured Bayesian Group Sparse Modeling", *IEEE ICIP Session TEC-P10*, Paris, 2014.

[31] 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.

[32] C. Vonesch and M. Unser, "A fast multilevel algorithm for wavelet-regularized image restoration", *IEEE Trans. on Image Process.*, vol. 18, pp. 509-523, 2009.

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

[34] Y. Zhang and N. Kingsbury, "Improved bounds for subband-Adaptive iterative shrinkage/thresholding algorithms", *IEEE Trans. Image Process.*, vol. 22, pp. 1373-1381, 2013.

[35] Y. Zhang and N. Kingsbury, "Fast L0-based sparse signal recovery", *IEEE International Workshop on Machine Learning for Signal Processing*, pp. 403-408, 2010.

[36] D. Tzikas, A. Likas and N. Galatsanos, "The variational approximation for Bayesian inference: Life after the EM algorithm", *IEEE Signal Process. Mag.*, vol. 25, pp. 131-146, 2008.

[37] D. Wipf, B. Rao and S. Nagarajan, "Latent variable Bayesian models for promoting sparsity", *IEEE Trans. on Information Theory*, vol. 57, pp. 6236-6255, 2011.

[38] J. O'Ruanaidh and W. Fitzgerald, "Numerical Bayesian methods applied to signal processing", *Springer-Verlag New York*, vol. 5, pp. 6-22, 1996.

[39] M. Figueiredo, "Adaptive sparseness using Jeffreys prior", *Advances in Neural Information Processing Systems*, pp. 697–704, 2001.

[40] C. Bishop and N. Nasrabadi. *Pattern recognition and machine learning*, vol. 1, springer New York, 2006.

[41] S. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge university press, 2004.

[42] R. Chartrand and W. Yin, "Iteratively reweighted algorithms for compressive sensing", *Proc. IEEE ICASSP 2008*, pp. 3869-3872, 2008.

[43] C. Chen and J. Huang, "Compressive sensing MRI with wavelet tree sparsity", *Advances in Neural Information Processing Systems*, pp. 1115-1123, 2012.

[44] Q. Shan, J. Jia, and A. Agarwala, "High-quality motion deblurring from a single image", *ACM Transactions on Graphics (TOG)*, vol. 27, pp. 73, 2008

[45] I. Selesnick and M. Figueiredo, "Signal restoration with overcomplete wavelet transforms: comparison of analysis and synthesis priors", *Proc. of SPIE*, vol. 7446 (Wavelets XIII), August 2-4, 2009.

[46] Y. Zhang and N. Kingsbury, "Restoration of images and 3D data to higher resolution by deconvolution with sparsity regularization", *Proc. IEEE ICIP 2010*, pp. 1685-1688, 2010.

[47] J. Yang, J. Wright, T. S. Huang and Y. Ma, "Image super-resolution via sparse representation", *IEEE Trans. on Image Process.*, vol.19, pp. 2861-2873, 2010.

[48] S. Villena, M. Vega, S. D. Babacan, R. Molina and A. K. Katsaggelos, "Bayesian combination of sparse and non-sparse priors in image super resolution", *Digital Signal Process.*, vol. 23, pp. 530-541, 2013.

[49] Z. Wang, A. C. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error measurement to structural similarity", *IEEE Trans. Image Process.*, vol. 13, pp. 600-612, 2004.

[50] M. Panju, "Iterative methods for computing eigenvalues and eigenvectors", *arXiv preprint arXiv:1105.1185*, 2011.

Ganchi Zhang received the B.Eng. degree in electrical engineering from the University of Strathclyde, Glasgow, U.K. and the M.Phil. degree in industrial engineering from the University of Cambridge, Cambridge, U.K., in 2011 and 2012, respectively. He is currently working towards the Ph.D. degree in the Signal Processing and Communications Laboratory, Department of Engineering, University of Cambridge, Cambridge, U.K.. His research interests include image enhancement, wavelet-based techniques, compressive sensing and Bayesian inference.

Nick Kingsbury received the Honours and Ph.D. degrees in electrical engineering from the University of Cambridge, Cambridge, U.K., in 1970 and 1974, respectively. From 1973 to 1983, he was a Design Engineer and subsequently a Group Leader with Marconi Space and Defence Systems, Portsmouth, U.K., specializing in digital signal processing and coding theory. Since 1983, he has been a Lecturer in communications systems and image processing with the University of Cambridge, where he became a Professor of signal processing in 2007 and where he is the Head of the Signal Processing and Communications Research Group. Since 1983, he has been a Fellow of Trinity College, Cambridge. His current research interests include image analysis and enhancement techniques, object recognition, motion analysis, and registration methods. He has developed the dual-tree complex wavelet transform and is especially interested in the application of wavelet frames to the analysis of images and 3-D data sets.