Fast L0-based Image Deconvolution with Variational Bayesian Inference and Majorization-Minimization

Ganchi Zhang and Nick Kingsbury
Signal Processing Group, Dept. of Engineering, University of Cambridge, Cambridge, CB2 1PZ, UK

***Abstract*—In this paper, we propose a new wavelet-based image deconvolution algorithm to restore blurred images based on a Gaussian scale mixture model within the variational Bayesian framework. Our sparsity-regularized model approximates an l_0 norm by reweighting an l_2 norm iteratively. We derive a hierarchial Bayesian estimation with the use of subband adaptive majorization-minimization which simplifies computation of the posterior distribution, and has been shown to find good solutions in the non-convex search space. The proposed method is flexible enough to incorporate group-sparse optimization.**

I. Introduction

Image deconvolution can usually be modeled as a linear inverse problem where the objective is to estimate the sharp image $\mathbf{x}$ from the blurred image $\mathbf{y}$:

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

where $\mathbf{H}$ is a $M \times M$ convolution matrix, and $\mathbf{n}$ is Gaussian noise with zero mean and variance ν^2. In general this is an ill-posed problem and therefore regularization methods are typically applied to stabilize the solution. Wavelet-based regularization methods are good for image restoration problems because the wavelet transform of natural images tends to be sparse [1]. The discrete wavelet transform exhibits a characteristic signal-dependent structure and represents this structure in a compact manner [2]. To well capture the statistical dependencies, the Gaussian Scale Mixture (GSM) has been widely applied to model wavelet coefficients whose energies are not randomly distributed [2].

However, wavelet-based image deconvolution is challenging because convolution operators are not simply represented in the wavelet domain [1]. Furthermore, wavelet-based regularization leads to large high-dimensional optimization problems because wavelet priors are often non-differentiable and sometimes even non-convex [3]. To address these problems, several authors have proposed the use of majorization-minimization (MM) techniques together with sparsity-based regularizers to alternate iteratively between a Landweber update and wavelet thresholding [3]–[6].

In this paper, we propose a wavelet-regularized image deconvolution algorithm (VBMM) which is a combination of hierarchical Bayesian estimation, using variational Bayesian (VB) inference, with a subband-adaptive MM method for efficiently finding maxima of posterior distributions. A significant contribution of our approach is to cause both of these iterative methods to converge simultaneously. In addition, we show that VBMM can efficiently incorporate group-sparse models, that are appropriate for tight-frame (redundant) wavelet transforms. We illustrate this with dual-tree complex wavelets (DT $\mathbb{C}$WT) which are well matched to VBMM because of their good sparsity versus redundancy tradeoff, high computational efficiency, and ability to decorrelate typical blur kernels. The latter property is a key to fast VBMM convergence.

The paper is organized as follows. Section II discusses the advantages of VB inference. Section III describes the key formulations of our statistical model. Section IV shows our continuation strategy and presents the proposed algorithm. Simulation results are shown in Section V.

II. Advantages of Variational Bayesian Inference

From a Bayesian perspective, many practical methods for sparse signal recovery are equivalent to performing maximum *a posteriori* (MAP) estimation which generates MAP point estimates using a sparsity-inducing prior distribution [8]. For instance, the l_1-norm approach to regression corresponds to performing MAP estimation using a Laplacian prior [2]. However 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 [7], and it has been shown in [8] that VB inference can effectively smooth out local minima and help to ensure that a near-global minimum solution is found.

III. Model Formulations

In this section we present the formulations of our statistical model. 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 $N > M$ [3]. The linear model in (1) then becomes

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

and the resulting likelihood of the data can be shown to be

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

A GSM model is now employed to model the wavelet coefficients. Inspired from [9], 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 formed from the vector $\mathbf{s}$ 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 [9]; 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 circular 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 is repeated g_i times. 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 squared form as

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

with

$$\mu = \nu^{-2}\Sigma\mathbf{M}^T\mathbf{H}^T\mathbf{y} \tag{7}$$

$$\Sigma = \left(\nu^{-2}\mathbf{M}^T\mathbf{H}^T\mathbf{H}\mathbf{M} + \mathbf{S}\right)^{-1} \tag{8}$$

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, as N is often $\sim 10^7$ or more. Here we adopt the MM technique from [4], together with the recent subband-adaptive MM from [5] to derive our fast algorithm.

To keep a Bayesian viewpoint, we now introduce the following approximation model for the posterior distribution:

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

where

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

Note that taking the logarithm of both sides of (9) will give a similar surrogate function to that proposed in [4]:

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

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 of a vector α whose elements α_j may be optimized independently for each subspace/subband j of $\mathbf{M}$, such that $\Lambda_\alpha - \mathbf{M}^T\mathbf{H}^T\mathbf{H}\mathbf{M}$ is 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 $\mathbf{w} = \mathbf{z}$ [5], and hence produces monotonicity of the decay of $J(\mathbf{w})$ [3]. Because $p\left(\mathbf{z}|\mathbf{w}\right) \propto \mathcal{N}\left(\mathbf{w}|\mathbf{z}, \nu^2(\Lambda_\alpha - \mathbf{M}^T\mathbf{H}^T\mathbf{H}\mathbf{M})^{-1}\right)$ and $p\left(\mathbf{w}|\mathbf{y}, \mathbf{S}, \nu^2\right)$ are Gaussian functions of $\mathbf{w}$, the approximation model $\overline{p}\left(\mathbf{w}, \mathbf{z}|\mathbf{y}, \mathbf{S}, \nu^2\right)$ is also a Gaussian distribution and, when $\mathbf{z}$ is given, can be rearranged into the form:

$$\overline{p}\left(\mathbf{w}|\mathbf{y}, \mathbf{z}, \mathbf{S}, \nu^2\right) = \mathcal{N}\left(\mathbf{w}|\overline{\mu}, \overline{\Sigma}\right) \tag{12}$$

with

$$\overline{\mu} = \nu^{-2}\overline{\Sigma}[(\Lambda_\alpha - \mathbf{M}^T\mathbf{H}^T\mathbf{H}\mathbf{M})\mathbf{z} + \mathbf{M}^T\mathbf{H}^T\mathbf{y}] \tag{13}$$

$$\overline{\Sigma} = (\nu^{-2}\Lambda_\alpha + \mathbf{S})^{-1} \tag{14}$$

where $\overline{\Sigma}^{-1}$ is now purely diagonal and easy to invert. This gives the subband-adaptive MM technique, whose convergence rate is improved by keeping the (Λ_α-$\mathbf{M}^T\mathbf{H}^T\mathbf{H}\mathbf{M}$) term small.

IV. Continuation Strategy

In this section, we describe the continuation strategy for our Bayesian framework. In the above approximation model $\overline{p}\left(\mathbf{w}|\mathbf{y}, \mathbf{z}, \mathbf{S}, \nu^2\right)$, it is required to estimate the hidden variable $\mathbf{z}$ as well as the inverse signal variance $\mathbf{S}$. Here we keep the noise variance ν^2 as a user parameter in order to be able to adjust the regularization strength. Note that although it can be estimated via Bayesian inference, its estimate can be inaccurate because of the difficulty of accurately separating broadband signal components from noise. For $\mathbf{S}$, or more conveniently $\mathbf{s}$, we impose a Gamma distribution because it is conjugate to the Gaussian distribution and can strongly encourage sparsity [7].

$$p(\mathbf{s}|a, \mathbf{b}) = \prod_{i=1}^{G}\frac{b_i^a}{\Gamma(a)}s_i^{a-1}\exp(-b_i s_i) \tag{15}$$

The rate parameter $\mathbf{b} = [b_1 \ldots b_G]^T$ has a strong influence on $\mathbf{s}$. So, as suggested in [10], we further assume $\mathbf{b}$ is a vector associated with a Gamma prior:

$$p(\mathbf{b}|k, \theta) = \prod_{i=1}^{G}\frac{\theta^k}{\Gamma(k)}b_i^{k-1}\exp(-\theta b_i) \tag{16}$$

The complete graphical model with hierarchical priors is shown in Fig. 1. In fact, the model we adopt is a 3-layer hierarchical prior, similar to the model in [10] except that [10] uses a Gamma distribution to estimate signal variance σ^2. As a result, the posterior of hidden variables now becomes

$$\begin{aligned} p\left(\mathbf{w}, \mathbf{z}, \mathbf{s}, \mathbf{b}|\mathbf{y}\right) &= \frac{p(\mathbf{w}, \mathbf{z}, \mathbf{s}, \mathbf{b}, \mathbf{y})}{p(\mathbf{y})} \\ &= \frac{p\left(\mathbf{y}|\mathbf{w}, \beta\right)p\left(\mathbf{w}|\mathbf{S}\right)p(\mathbf{z}|\mathbf{w})p(\mathbf{s}|a, \mathbf{b})p(\mathbf{b}|k, \theta)}{p\left(\mathbf{y}\right)} \end{aligned} \tag{17}$$

where $\beta \equiv \nu^{-2}$. Note that the exact Bayesian inference of (17) cannot be performed as the marginal likelihood $p\left(\mathbf{y}\right)$ is intractable [7]. To approximate the posterior $p(\xi|\mathbf{y})$ where $\xi = \{\mathbf{w}, \mathbf{z}, \mathbf{s}, \mathbf{b}\}$, we adopt the VB approximation, which provides a distribution $q(\xi)$ to approximate $p(\xi|\mathbf{y})$ [7] [9]. To be specific, $q(\xi)$ is determined by minimizing the Kullback-Leibler (KL) divergence between $q(\xi)$ and $p(\xi|\mathbf{y})$ as

$$\text{KL}(q(\xi)\|p(\xi|\mathbf{y})) = -\int q(\xi)\ln\left(\frac{p(\xi|\mathbf{y})}{q(\xi)}\right)d\xi \tag{18}$$

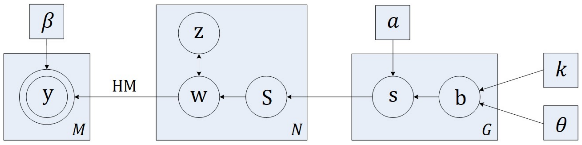


Fig. 1. The graphic model of linear regression with hierarchical priors. $\mathbf{y}$ and $\mathbf{z}$ are Gaussian distributions, $\mathbf{w}$ is a GSM, $\mathbf{s}$ and $\mathbf{b}$ are Gamma distributions.

We note that $\mathrm{KL}(q(\xi)\|p(\xi|\mathbf{y})) \geq 0$, and equality holds only when $q(\xi) = p(\xi|\mathbf{y})$ [7]. To find $q(\xi)$, we use the mean-field approximation which assumes the posterior independence between $\mathbf{w}, \mathbf{z}, \mathbf{s}$ (and hence $\mathbf{S}$) and $\mathbf{b}$ [9], such that

$$q(\xi) = q(\mathbf{w}, \mathbf{z}, \mathbf{s}, \mathbf{b}) \approx q(\mathbf{w})q(\mathbf{z})q(\mathbf{s})q(\mathbf{b}) \tag{19}$$

Based on this factorization, the distribution of each variable $q(\lambda)$, $\lambda \in \xi$, which minimizes (18) can be optimized as

$$\begin{aligned}\ln q(\lambda) &= \langle \ln p(\xi|\mathbf{y})\rangle_{q(\xi\setminus\lambda)} \\ &= \langle \ln p(\xi, \mathbf{y})\rangle_{q(\xi\setminus\lambda)} + \text{const}\end{aligned} \tag{20}$$

where $\langle\cdot\rangle_q$ denotes expectation over q and $\xi \setminus \lambda$ means the set of ξ with λ removed. By sequentially calculating $q(\lambda)$, we obtain the following updating rules.

(i) *Optimize* $\ln q(\mathbf{w})$ *using (5), (9) and (12)*

$$\begin{aligned}\ln q(\mathbf{w}) &= \langle \ln p(\xi, \mathbf{y})\rangle_{q(\mathbf{z})q(\mathbf{s})q(\mathbf{b})} + \text{const} \\ &= \langle \ln(p(\mathbf{y}|\mathbf{w}, \beta)p(\mathbf{w}|\mathbf{S})p(\mathbf{z}|\mathbf{w}))\rangle + \text{const} \\ &= -\frac{1}{2}\mathbf{w}^T\overline{\Sigma}^{-1}\mathbf{w} + \mathbf{w}^T\overline{\Sigma}^{-1}\overline{\mu} + \text{const}\end{aligned} \tag{21}$$

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

$$\mathbf{w}^{(t+1)} = \overline{\mu}^{(t)} \tag{22}$$

where $\overline{\mu}^{(t)}$ is computed using the current estimates of $\mathbf{S}^{(t)}$ and $\mathbf{z}^{(t)}$ as in (13) and (14):

$$\overline{\Sigma}^{(t)} = \Big(\beta\Lambda_\alpha + \mathbf{S}^{(t)}\Big)^{-1} \tag{23}$$

$$\overline{\mu}^{(t)} = \beta\overline{\Sigma}^{(t)}[\Lambda_\alpha\mathbf{z}^{(t)} - \mathbf{M}^T\mathbf{H}^T(\mathbf{HMz}^{(t)} - \mathbf{y})] \tag{24}$$

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

$$\begin{aligned}\ln q(\mathbf{z}) &= \langle \ln p(\xi, \mathbf{y})\rangle_{q(\mathbf{w})q(\mathbf{s})q(\mathbf{b})} + \text{const} \\ &= \langle \ln p(\mathbf{z}|\mathbf{w})\rangle + \text{const} \\ &= -\frac{1}{2}\mathbf{z}^T\Sigma_{\mathbf{z}}\mathbf{z} + \mathbf{z}^T\Sigma_{\mathbf{z}}\langle\mathbf{w}\rangle + \text{const}\end{aligned} \tag{25}$$

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

$$\mathbf{z}^{(t+1)} = \langle\mathbf{w}\rangle = \mathbf{w}^{(t+1)} \tag{26}$$

(iii) *Optimize* $\ln q(\mathbf{s})$ *using (4) and (15)*

$$\begin{aligned}&\ln q(\mathbf{s}) = \langle \ln p(\xi, \mathbf{y})\rangle_{q(\mathbf{w})q(\mathbf{z})q(\mathbf{b})} + \text{const} \\ &= \langle \ln(p(\mathbf{w}|\mathbf{S})p(\mathbf{s}|a, \mathbf{b}))\rangle + \text{const} \\ &= \sum_{i=1}^{G}\Big(\big(a + \frac{g_i}{2} - 1\big)\ln s_i - \big(\frac{\langle|\mathbf{w}_i|^2\rangle}{2} + \langle b_i\rangle\big)s_i\Big) + \text{const}\end{aligned} \tag{27}$$

This is the exponent of the product of G Gamma distributions [7]. Thus the mean of s_i for $i = 1 \dots G$, occurs when

$$s_i^{(t+1)} = \frac{g_i + 2a}{\Big(\|\overline{\mu}_i^{(t)}\|^2 + \mathrm{tr}[\overline{\Sigma}_i^{(t)}]\Big) + 2b_i^{(t)}} \tag{28}$$

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.

(iv) *Optimize* $\ln q(\mathbf{b})$ *using (15) and (16)*

$$\begin{aligned}&\ln q(\mathbf{b}) = \langle \ln p(\xi, \mathbf{y})\rangle_{q(\mathbf{w})q(\mathbf{z})q(\mathbf{s})} + \text{const} \\ &= \langle \ln(p(\mathbf{s}|a, \mathbf{b})p(\mathbf{b}|k, \theta))\rangle + \text{const} \\ &= \sum_{i=1}^{G}\Big(\big(a + k - 1\big)\ln b_i - \big(\langle s_i\rangle + \theta\big)b_i\Big) + \text{const}\end{aligned} \tag{29}$$

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

$$b_i^{(t+1)} = \frac{a + k}{s_i^{(t+1)} + \theta} \tag{30}$$

The procedure of iteratively updating $q(\lambda)$ can be seen as an alternating minimization of KL divergence in (18), which is repeated until the KL divergence converges [9]. Similar to the analysis in [10], if we marginalize $p(\mathbf{w}, \mathbf{s}, \mathbf{b})$ over $\mathbf{b}$ and $\mathbf{s}$, we can obtain the prior pdf $p(\mathbf{w})$ for the ith group as

$$p(\mathbf{w}_i) = \frac{\theta^{\frac{g_i}{2}}}{\sqrt{2\pi}}\frac{\Gamma(a+k)\Gamma(a+\frac{g_i}{2})}{\Gamma(a)\Gamma(k)}\Big(\frac{\theta|\mathbf{w}_i|^2}{2}\Big)^{k-\frac{g_i}{2}}F \tag{31}$$

where $F = U(a + k; k + 1 - \frac{g_i}{2}; \frac{\theta|\mathbf{w}_i|^2}{2})$ is a confluent hypergeometric function. By calculating the negative log likelihood, the optimization function of our model is found to be

$$\Omega(\mathbf{w}) = \arg\min_{\mathbf{w}}\ \Big(\frac{1}{2}\|\mathbf{y} - \mathbf{HMw}\|^2 + \nu^2\sum_{i\in G}\phi(\mathbf{w}_i)\Big) \tag{32}$$

where $\phi(\mathbf{w}_i) = -\log p(\mathbf{w}_i)$. It can be seen from (28) that the proposed method approximates the l_0 norm by reweighting the l_2 norm iteratively, which can be regarded as a relaxation of Iterative Reweighted Least Squares (IRLS) studied in [11].

V. Experimental Results

In this section, we present a set of experiments to evaluate our proposed VBMM algorithm. We show that the performance of VBMM is better than a closely related and recently developed image deconvolution algorithm: modified subband-adaptive iterative shrinkage/thresholding (MSIST) in [6].

For the wavelet basis, we chose the DT $\mathbb{C}$WT because it has a good frequency selectivity and is almost shift-invariant [12]. For typical blur $\mathbf{H}$, the DT $\mathbb{C}$WT can compress most of the energy of $\mathbf{M}^T\mathbf{H}^T\mathbf{HM}$ into the leading diagonal or near diagonal terms, which ensures that Λ_α provides a good approximation to $\mathbf{M}^T\mathbf{H}^T\mathbf{HM}$. Because for the DT $\mathbb{C}$WT there is energy leakage to adjacent subbands, we computed Λ_α=$2\rho(\mathbf{M}^T\mathbf{H}^T\mathbf{HM})$ for each subband with thresholding to a limiting value of 1 when $\alpha \geq 1$, which accounts for spectral leakage. Note that the DT $\mathbb{C}$WT is also chosen for evaluating

MSIST in [6]. Because the DT $\mathbb{C}$WT produces complex wavelet coefficients, we rearranged them in $G=\frac{N}{2}$ groups of $g_i=2$ as described in Section III. The standard test image, Cameraman, was used in the experiments for comparative purposes. We convolved the 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)$=10\log_{10}\frac{\|\mathbf{Hx}_r-\overline{\mathbf{Hx}_r}\|^2}{M\nu^2}$ was used to define the noise level. $\mathbf{x}_r$ is the original image and $\overline{\mathbf{Hx}_r}$ is the mean of $\mathbf{Hx}_r$. The improvement in signal-to-noise ratio (ISNR) $=10\log_{10}(\frac{\|\mathbf{y}-\mathbf{x}_r\|^2}{\|\mathbf{Mw}-\mathbf{x}_r\|^2})$ was used to evaluate each estimate $\mathbf{w}$. The initial estimation of $\mathbf{x}_r$ was achieved by a Wiener-type filter $\mathbf{x}_0=(\mathbf{H}^T\mathbf{H}+10^{-3}\nu^2\mathbf{I})^{-1}\mathbf{H}^T\mathbf{y}$ [6]. For each group i, the initial estimation of weight was $s_i=\frac{g_i+2a}{\|\mathbf{w}_i\|^2+2b}$. In the experiment, we set hyperparameters $b^{(0)}=\theta=10^{-6}$ and $k=0.5$ which were found to be optimal. The shape parameter a can control the initial sparsity pattern of the Gamma distribution. Typically using a small a will converge faster but to a slightly low ISNR compared to a larger a as shown in Fig. 2. Although the noise variance was known in the experiment, we found that for VBMM, a bigger β (by$\sim$2:1) provides a better performance as it allows the algorithm to put less emphasis on the regularization.

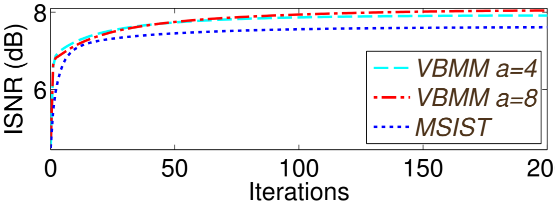
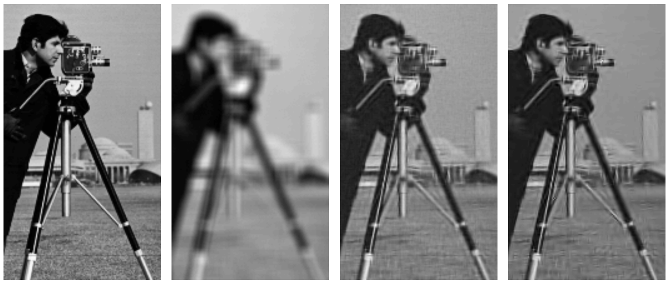


Fig. 2. ISNR results over 200 iterations, on Cameraman, BSNR: 40 dB.

Fig. 2 compares the ISNR of VBMM to MSIST over 200 iterations when the BSNR of the observation is 40dB. Deconvolution results are shown in Fig. 3 for visual comparison. The results show that VBMM reaches a faster convergence rate and better ISNR results than MSIST. The computation time for 200 iterations was 10.62 seconds for VBMM and 13.98 seconds for MSIST on a Core i7 PC with 3.40 GHz Intel Processor. We then considered another two noise levels, BSNR=20 dB, 50 dB. Table I shows the average ISNR results of VBMM over 30 noise realizations compared with the results of MSIST reported in [6]. VBMM requires fewer iterations to reach a given quality of recovery and consistently outperforms MSIST at all three noise levels.

TABLE I
AVERAGE ISNR RESULTS OVER 30 NOISE REALIZATIONS, 'M' STANDS FOR MSIST, 'V' STANDS FOR VBMM

BSNR	20 dB		40 dB		50 dB	
Method	M	V	M	V	M	V
10 iters	2.584	2.731	7.011	7.107	8.760	10.148
30 iters	2.990	3.282	7.348	7.531	10.290	10.656
50 iters	3.191	3.491	7.449	7.730	10.601	10.879
70 iters	3.308	3.582	7.506	7.842	10.669	10.996
100 iters	3.403	3.646	7.553	7.939	10.683	11.085



(a) Original (b) Blurred (c) MSIST, ISNR: 7.611 dB (d) VBMM a=8, ISNR: 8.042 dB

Fig. 3. Deconvolution results, on Cameraman, BSNR: 40 dB.

VI. Conclusion

Here we have proposed the VBMM image deconvolution algorithm, based on hierarchical Variational Bayesian inference combined with subband adaptive Majorization-Minimization for fast convergence. Our model shows how VB approximates the l0-norm on wavelet coefficients by reweighting the l2-norm iteratively. Experimental results confirm the performance of the method. We have considered group-sparse optimization of complex wavelet coefficients, which can be further extended for tree-structured wavelet modeling.

References

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

[2] V. Cevher and P. Indyk and L. Carin and R. Baraniuk, "Sparse signal recovery and acquisition with graphical models", *IEEE Signal Process. Mag.*, vol. 27,pp. 92-103, Nov. 2010.

[3] 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.

[4] 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.

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

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

[7] 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.

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

[9] S. Babacan, S. Nakajima and M. Do, "Bayesian group-sparse modeling and variational inference", unpublished. [Online]. Avaliable: http://www.dbabacan.info/papers/VBGS_final.pdf.

[10] N. Pedersen, D. Shutin, C. Manchón, and B. Fleury, "Sparse estimation using Bayesian hierarchical prior modeling for real and complex models", *preprint arXiv:1108.4324v2*, 2012.

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

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