Model based variational Bayesian compressive sensing using heavy tailed sparse prior

Zahra Sadeghigol, Mohammad Hossein Kahaei, Farzan Haddadi

School of Electrical Engineering, Iran University of Science and Technology, Tehran, Narmak 16846-13114, Iran

A R T I C L E   I N F O
Article history:
Received 17 February 2015
Received in revised form
13 September 2015
Accepted 14 September 2015
Available online 3 November 2015

Keywords:
Bayesian compressive sensing
Hidden Markov tree
Statistical signal modelling
Variational Bayes inference

A B S T R A C T
In this paper, a novel multiscale model-based Bayesian compressive sensing is investigated using variational Bayesian inference in the complex wavelet domain. This model preserves the structural information by two-state signal/noise Hidden Markov Tree (HMT). Tree structured hierarchical Generalized Double Pareto (GDP) distribution is used to model the sparsity of the signal. Using the Variational Bayes (VB) inference procedure a closed-form solution is obtained for model parameters. Experimental results in compressive sensing application show that the reconstruction error and CPU time of the proposed algorithm is lower compared to the other well-known algorithms.

1. Introduction

The demand for the digitally rich communications multimedia is escalating specially in the form of image and visual medium. In order to store these digital products, the data should be compressed as much as possible. The most recent image compression standard, JPEG2000 [1] and the former JPEG [2], are usually used for this purpose. The conventional compression standards extract a myriad of samples from the underlying image, then compress them. During the compression process, many of the collected samples should be discarded, which seems wasteful. Moreover, there are some limitations for the real time implementation such as the acquisition time and the large amount of stored data. These deficits can be answered in Compressive Sensing (CS) theory [3,4]. In the CS framework, a low dimensional vector containing linear combinations of the demonstrated signal coefficients is extracted. This can be written as

\[ y = \Phi w + n \] (1)

in which \( y_{N \times 1} \) is the CS measurement vector, \( \Phi_{N \times M} \) is the random projection matrix. In CS applications, \( N \ll M \). This property reduces the dimension of the underlying signal which is shown by \( w_{M \times 1} \) in (1). \( n_{N \times 1} \) is assumed to be additive white Gaussian i.i.d. random process. Much effort has been devoted to solving the underdetermine equation (1). Some of these studies are concentrated on the optimization techniques. The related reconstruction problem is given by \( \ell_1 \)-regularized formulation [5]:

\[ w = \arg \min_w \| y - \Phi w \|_2^2 + \rho \| w \|_1 \] (2)

where the scaler \( \rho \) denotes the model parameter that controls the relative importance given to the error term and the sparseness term.

1.1. Related literature

Optimization techniques attract much interest in the existing algorithms for solving (1). Several CS inversion algorithms can be found in the literature including linear programming [6], greedy algorithms [7,8], and reweighted
norm algorithms [9,10]. These methods are restricted to analyzing an optimization problem and are based on a point estimate of \( w \). Their main drawback is that they do not consider the uncertainty of the underlying signal in the reconstruction process. To overcome this challenge, the Bayesian framework is developed [11,12]. These approaches estimate the signal by assuming an appropriate prior on its transform coefficients. Choosing a suitable transform domain, an appropriate prior and a powerful inference are the main factors in Bayesian CS. In general, a Bayesian compressive sensing system includes the stages depicted in Fig. 1.

A Bayesian method has been presented in [13] that apply Bernoulli-Gaussian distribution as the prior in Fourier transform domain. However, Fourier domain can not display time-frequency structure of an image explicitly. Another drawback of this work is assuming a Gaussian prior. However this distribution can not illustrate the heavy-tailed behavior of the signal coefficients histogram. In [14], a beta-Bernoulli process has also been proposed for semisupervised hyperspectral unmixing problem in the compressive sensing framework. It used Gibbs sampling as a hierarchical Bayesian inference to determine an appropriate dictionary for image recovery. The transform domain in [14] is discrete wavelet which exhibits some fundamental drawbacks which will be mentioned later.

The multiplicative perturbation problem in compressive sensing has been considered in [15] as a probabilistic model. A variational Bayesian expectation maximization technique is devised to recover the underlying signal and perturbation. In another work, [16], Double Lomax distribution is considered as the prior pdf. It is categorized in the group of the automatic relevance determination priors with non-log-concave property for the strong sparsity promotion. Variational Bayes is used as the inference procedure in [16]. The method uses discrete wavelet transform which suffers from some known drawbacks [23]. Moreover, it does not take advantage of statistical relationship between wavelet coefficients such as inter-scale or intra-scale dependencies.

As another technique, multiple description coding has been presented in [17], for compressive sensing in high packet loss transmission. Also, two-dimensional discrete wavelet transform (DWT) is applied for sparse representation. DWT coefficients re-sampled towards equal importance of information. Exploiting the intra-scale and inter-scale correlation of DWT, the CS recovery algorithm is developed for the decoder side.

A Bayesian Expectation Maximization (EM) algorithm for reconstructing Markov-tree sparse signals via belief propagation has been appeared in [18]. This signal reconstruction algorithm is based on an EM iteration which maximizing the posterior distribution of the signal and its state variables given the noise variance. The missing data was compensated for in the EM iteration such that the complete-data posterior distribution corresponds to a hidden Marcov tree (HMT) probabilistic graphical model.

In two recent technique [19,20], two model-based compressive sensing schemes using Markov chain Monte Carlo and variational Bayes inference are introduced, respectively. They used spike-slab prior for two-state HMT modelling of signal and noise. Both use DWT as transform domain and Gaussian pdf as prior.

In all the above references, DWT or Fourier transform is used as the transform domain. Though, these transforms can not represent sufficient orientations in the image. DWT suffers from multiple weaknesses which are discussed in the sequel. Another drawback in some of the above algorithms is utilizing the Gaussian distribution as the prior, while the transform coefficients histogram is strongly non-Gaussian. In this paper, a novel multiscale model based on the three level hyper prior is developed to extract structural information efficiently.

1.2. Contribution

To address the above-mentioned drawbacks, in this paper a novel fully automated algorithm and an original hierarchical Bayesian technique are proposed. In this work, the new GDP-HMT is utilized to model the signal sparsity. Based on this model, the complete VB inference is designed to derive the closed form posteriors for all unknown parameters. These are shown in the 3rd and 5th blocks of Fig. 2 which illustrates the block diagram of the proposed algorithm. Generalized double Pareto distribution is a scale mixture of Laplace and normal distributions and also, it is an interface between the Laplace and Normal-Jeffreys priors. While it peaks at zero such as the Laplace density, it has a Student’s t-like heavy-tailed behavior [21].

The paper is organized as follows. In Section 2, dual tree-complex wavelet transform is elaborated. Generalized Double Pareto Mixture Hidden Markov Tree Model is described in Section 3. Section 4 is contributed to prior assignment. Variational Bayes inference is developed in Section 5. The last two sections include experimental results and conclusion.

---

Fig. 1. Bayesian compressive sensing system.

Fig. 2. Block diagram of the proposed Generalize Double Pareto HMT algorithm.
2. Dual tree-complex wavelet

Transform as Sparse Transform Discrete Wavelet Transform (DWT) is often used as a sparsifying transform [7–20,22]. Although, DWT suffers from fundamental difficulties [23]:

1. Oscillations of coefficients around singularities,
2. Shift variance,
3. Aliasing caused by down-sampling, and
4. Poor directional selectivity.

DT-CWT avoids the aforementioned problems [24,25]. The relation between the imaginary and real parts of the DT-CWT is given by [23]:

\[
\Psi_c(t) = \Psi_r(t) + j\Psi_i(t)
\]

(3)

where \(\Psi_r(t)\) is a real even function and \(\Psi_i(t)\) is an imaginary odd function. Since these two parts are the Hilbert transform of each other, \(\Psi_r(t)\) is an analytical function. The DT-CWT can detect singularities much more accurately than the DWT and the directional selectivity of the complex wavelet is much better than the real DWT. Fig. 3 shows the filter bank implementation for the DT-CWT.

3. Generalized Double Pareto Mixture Hidden Markov Tree Model

In the following section, the proposed GBMG-HMT model is introduced.

3.1. Generalized Double Pareto distribution

In spite of a highly non-Gaussian statistics of the signal coefficients, in many studies such as [11,19,20] a Gaussian distribution has been considered as its model. To remedy this disagreement, here we propose the generalized double Pareto density:

\[
f(\theta|\zeta, \alpha) = \frac{1}{2\zeta}\left(1 + \frac{\theta}{\alpha\zeta}\right)^{-(\alpha+1)}
\]

(4)

where \(\zeta > 0\) is the scale parameter and \(\alpha > 0\) is the shape parameter. The mean and variance for the generalized double Pareto distribution are

\[
E = \frac{\zeta}{1 - \frac{1}{\alpha}}
\]

(5)

\[
V = \zeta^2\left(1 - \frac{1}{\alpha}\right)^2\left(1 - \frac{2}{\alpha}\right)
\]

(6)

If \(\alpha \to \infty\), (4) becomes an exponential density with mean \(\zeta\) and variance \(\zeta^2\). The special case of (4) is obtained when \(\alpha = \zeta = 1\), so that \(f(\theta) = 1/(2(1 + \theta)^2)\). This form is referred to as the standard double Pareto. The standard double Pareto distribution is the same as Laplace density near zero, leading to similar sparse shrinkage properties of small transform coefficients by maximum a posteriori estimation. Also, it has Cauchy-like tail which results in avoiding over-shrinkage away from the origin. Prior (4) can be synthesized as a scale mixture of Gaussian pdf to reduce computational complexity.

Proposition 1. Let \(\theta \sim N(0, \tau)\), \(\tau \sim \text{Exp}(\lambda^2/2)\), and \(\lambda \sim \text{Gamma}(\alpha, \eta)\), where \(\alpha > 0\) and \(\eta > 0\). The resulting marginal density for \(\theta\) is \(\text{GDP}(\zeta = \eta/\alpha, \alpha)\). This three-level hierarchical prior model is utilized in the proposed algorithm [21].

Proposition 2. Given the representation in Proposition 1, \(\theta \sim \text{GDP}(\zeta = \eta/\alpha, \alpha)\) implies

1. \(f(\theta) \propto 1/|\theta|\) for \(\alpha = 0\) and \(\eta = 0\),
2. \(f(|\theta|\lambda) = (\lambda^2/2)\exp(-\lambda^2/\theta)\) for \(\alpha \to \infty, \alpha/\eta = \lambda'\) and \(0 < \lambda' < \infty\).

To justify the first part of the Proposition 2, if \(\alpha = \zeta = 0\), the \(\lambda\) distribution will be \(1/2 \sim \text{Gamma}(0,0)\). Therefore the distribution of \(\tau\) can be shown to be

\[
\tau \sim \int p(\tau|\lambda)p(\lambda)\,d\lambda = \frac{1}{\tau}
\]

(7)

therefore distribution \(\tau\) is proportional to \(1/\tau\). In this situation, prior of \(\theta\) can be obtained by the Normal-Jeffreys. Proposition 2 shows that the prior in (4) is an interface

![Image](image.png)

Fig. 3. Filter bank for the dual-tree CWT.
between two limiting cases: Laplace and Normal–Jeffrey’s priors [21].

Theorem 1. Generalized Double Pareto prior in Eq. (7) is log-convex.

\[
\ln \text{GDP} = - \ln (2\zeta) - (\alpha + 1) \ln \left( 1 + \frac{|\theta|}{\alpha \zeta} \right)
\]

The second derivative of it, can be shown as

\[
\frac{\partial^2 (\ln \text{GDP})}{\partial \theta^2} = \frac{\alpha + 1}{\alpha \zeta + |\theta|^2} \begin{cases} > 0 & \alpha > 0 \\ = 0 & \alpha = 0 \\ < 0 & \alpha < 0 \end{cases}
\]

Therefore, GDP prior has log-convex property for any \( \alpha > 0 \). If \( \alpha \rightarrow \infty \) then the second derivative is equal to zero. The motivation for the GDP distribution is discussed in detail in [21].

Incorporation of dependencies of the transform coefficient locations into model improves the CS reconstruction [26]. In the proposed Tree Structured Complex Wavelet-Generalized Double Pareto-Hidden Markov Tree-VB (TSCW-GDP-HMT), the two-state HMT is used to represent the inter-scale statistical dependency of complex wavelet coefficients. Specifically, the high and low states are described by the GDP prior and an impulse function at zero, respectively. To find the variational Bayes solution, the GDP prior plus impulse function is modelled by multiplication of GDP and a Bernoulli random variable. It is vital that the selected model is conjugate to the Gaussian likelihood function which leads to a closed form posterior distribution [11]. Three hyper parameters in this model makes the histogram fitting problem more flexible. In Fig. 4, histogram fitting of the Gaussian, best fit Gaussian and GDP distribution are shown. It is clear that GDP is the best fit. GDP parameters are estimated by the Maximum Likelihood (ML) method (see Appendix A).

3.2. Generalized Double Pareto HMT Model

Suitable signal model improves the performance of CS significantly by incorporating dependencies among signal coefficients’ [26]. In this paper, the proposed generalized double Pareto HMT model is included in a hierarchical Bayesian framework.

Dual tree-complex wavelet transform can be represented as

\[
x = \Psi w
\]

where \( x_{M \times 1} \) is the signal vector, \( \Psi \) is an \( M \times M \) matrix containing the DT-CWT basis vectors, and \( w_{M \times 1} \) is the vector of DT-CWT coefficients. DT-CWT coefficients can be represented in a tree structure. The coefficients at the first scale \( s = 1 \) are root nodes of the tree and the coefficients at the \( s > 1 \) are leaf nodes. Each coefficient at scale \( s \) which is named as parent, has four children in the corresponding scale \( s + 1 \). There is a dependency between parent and children coefficients. If the parent coefficient is small, the corresponding children coefficients will also be small. However, if it is large, the corresponding children coefficients may be large or small. This statistics can be demonstrated by hidden Markov tree structure. In the proposed HMT model, each DT-CWT coefficient may be drawn from GDP/spike distribution which is dependent on the hidden state: high or low. High state is modelled by GDP/spike distribution corresponding to large coefficients. Also, low state is modelled by spike distribution devoted to small coefficients. The spike function is constructed by conditional Bernoulli distribution which is zero when its corresponding parent is small and one when its parent is large. This model can be illustrated as

\[
w_i \sim \text{GDP}(\theta_i | \zeta_s, \alpha) \times \text{Bernoulli}(\pi_i)
\]

where \( w_i \) is wavelet coefficient and \( \pi_i \) is its parent. \( \theta_i \) is an auxiliary variable carrying parent information. In fact, \( \theta_i \) is the original wavelet coefficients which is added to parent information while the final value of wavelet coefficient is \( w_i \).

4. Prior assignment

To infer the posterior distribution of complex wavelet coefficients in the TSCW-GDP-HMT algorithm, the prior of the CS observation and \( i \)th complex wavelet coefficient, \( i = 1, 2, \ldots, M \) is proposed as

\[
yw, \lambda_n \sim \mathcal{N}(\Phi w, \lambda_n^{-1} I)
\]

\[
w = \Theta \odot z
\]

\[
\theta_{ij} \sim \text{GDP}(\lambda_s, \tau_s)
\]

\[
\tau_s \sim \text{Exp}(\lambda_s^2/2)
\]

\[
\lambda_s \sim \text{Gamma}(\alpha_s, \eta_s)
\]

\[
z_{ij} \sim \text{Bernoulli}(\pi_{si})
\]

where \( \lambda_n \) denotes the noise precision parameter, \( \Theta \in \mathbb{R}^M \) is the vector of signal coefficients, \( z \in \mathbb{R}^M \) is the vector of zero/one indicators, and \( \odot \) denotes Hadamard product. Also, the shape parameter \( \alpha_s \) and the scale parameter \( \eta_s \) are learned from the data. Accordingly, the priors of the
unknown parameter $\pi_{s,i}$ and $\lambda_n$ in the above equations are given by

$$\pi_{s,i} = \begin{cases} 
\pi^0 & \text{if } s = 0, 1 \\
\pi^{10} & \text{if } 2 \leq s \leq L, z_{pah(i)} = 0 \\
\pi^1 & \text{if } 2 \leq s \leq L, z_{pah(i)} = 1 
\end{cases}$$

(18)

$$\lambda_n \sim \text{Gamma}(a_0, b_0)$$

(19)

$$\pi^t \sim \text{Beta}(\pi^t(0), f^t(0)), s = 0, 1,$$

(20)

$$\pi^{10} \sim \text{Beta}(\pi^{10}(0), f^{10}(0)), s = 2, \ldots, L$$

(21)

$$\pi^1 \sim \text{Beta}(\pi^1(0), f^1(0)), s = 2, \ldots, L$$

(22)

where $(s, i)$ is the index of the $i$th elements at level $s$, for $i = 1, \ldots, M_i$ and $M_i$ is the total number of coefficients at this scale. Also, $\pi^t, \pi^{10}$, and $\pi^1$ represent the mixing weights generated by the beta priors which model the structural information based on the scale and parent dependencies. $\pi^0$ determines the fact that when a parent node is zero, usually its children coefficients are zero as well [19]. The wavelet tree structural information incorporates the parent-children relation and the prior represents the propagation of small coefficients across the scales in a statistical sense. The priors selected in (18)–(22) are the same as those in [19].

5. Variational Bayes inference procedure

The posterior pdf is estimated in this paper via the variational Bayes inference. The base of the Bayesian inference is the Bayes rule on the posterior

$$p(\theta, \tau, \lambda, z, \lambda_n, \pi^t, \pi^{10}, \pi^1 | y) = \frac{p(\theta, \tau, \lambda, z, \lambda_n, \pi^t, \pi^{10}, \pi^1, y)}{p(y)}$$

(23)

Since the integral of $p(y)$ is usually intractable, approximated inferences are proposed. The Variational Bayesian (VB) inference [27–29] is incorporated in TSCW-GDP-HMT algorithm, which approximates the posterior by pursuing a factorizable distribution, established by minimizing the Kullback–Leibler divergence between the posterior and its approximation. In the TSCW-GDP-HMT algorithm, the approximation posterior distribution $q(\theta, \tau, \lambda, z, \lambda_n, \pi^t, \pi^{10}, \pi^1)$ is factorized as

$q(\theta)q(\tau)q(\lambda)q(z)q(\lambda_n)q(\pi^t)q(\pi^{10})q(\pi^1)$. To compute $q(\theta)$,

$$\ln \hat{q}(\theta) \propto \langle \ln p(y(\theta, \tau, \lambda_n), \lambda_n) \rangle_{z, \lambda_n, \tau}$$

(24)

where the angle bracket $\langle \rangle$ represents the expectation with respect to the random variable in its subscript. $\hat{q}(\theta)$ is a Gaussian random variable with parameters:

$$\lambda_\theta = (\lambda_n) \Phi^T \Phi (\hat{z}) + \frac{1}{(\hat{\tau})}$$

(25)

$$\mu_\theta = 2(\lambda_n) \langle z \rangle \Phi \lambda_\theta^{-1}$$

(26)

Proof. Appendix B.

To compute the posterior of $\tau$, it can be shown that

$$\ln \hat{q}(\tau) \propto \langle \ln \left( \frac{1}{\sqrt{2\pi \tau}} \exp \left( -\frac{1}{2} \left( \frac{\theta^2}{\tau} + \lambda^2 \tau \right) \right) \right) \rangle$$

(27)

Therefore, $\tau$ is distributed as Generalized Inverse Gaussian (GIG):

$$\tau = \text{GIG} \left( \langle \lambda^2 \rangle, \langle \theta^2 \rangle, \frac{1}{2} \right)$$

(28)

Proof. Appendix C.

The relation between the moments of different orders can be derived by moments analysis of GIG distribution as follows:

$$\langle \tau \rangle = < \tau^{-1} > - 1 + < \lambda^2 > - 1$$

(29)

$$\langle \tau^{-1} \rangle = \frac{1}{\langle \theta^2 \rangle}$$

(30)

By substituting $\alpha = -1$ (see Appendix E), $\lambda$ will be a Gaussian distribution which its parameters can be obtained as

$$\mu_\lambda = -\eta \theta$$

$$\alpha_\lambda = \langle \lambda \rangle$$

(31)

$\mu_\lambda$ is its mean and $\alpha_\lambda$ denotes its precision.

Proof. Appendix E.

Since Gamma distribution is conjugate to the precision of Gaussian distribution, its posterior is still a Gamma distribution with respect to $\lambda_n$

$$\hat{q}(\lambda_n) = \Gamma(\lambda_n | A_{\lambda_n}, B_{\lambda_n})$$

$$\frac{1}{\lambda_n} = A_{\lambda_n} + x_n$$

$$B_{\lambda_n} = \frac{1}{2} \text{tr} \left( (y - \Phi \omega) (y - \Phi \omega)^T \right) + b_{\lambda_n}$$

(32)

The update equations for posterior pdf of (17) and (20)–(22) are repeated given in [19] which are in Appendix F for the sake of completeness.

6. Experimental results

In this section, the proposed multiscale TSCW-GDP-HMT algorithm is evaluated in two scenarios by several test images. The first scenario investigates the performance of the proposed algorithm in noise free CS measurements. While in the second scenario, the additive white Gaussian noise is considered.
In all algorithms presented in this section, the matrix $\Phi \in \mathbb{R}^{N \times M}$ is drawn randomly from $\mathcal{N}(0, 1/N)$ with i.i.d components. $\Phi$ may also be demonstrated with any other constructions as in [3,5]. Test images are of size $128 \times 128$. The Mexican-hat wavelet with three levels of decomposition is used in all experiments to provide a fair comparison, but any discrete wavelet transform can be employed. Scaling coefficients are given by an $8 \times 8$ matrix. In the relevant literature, scaling coefficients are usually assumed to be known and measured separately. All the algorithms addressed below are to estimate the complex wavelet coefficients of size $(128^2 - 8^2)$ by a limited number of CS measurements.

6.1. Comparisons to conventional CS algorithms

The performance of TSCW-GDP-HMT is compared with seven recently developed CS reconstruction algorithms including: L1-LS [30], SPGL1 [31,32], SL0 [33], BCS [11], TSW-CS-MCMC [19], TSW-CS-VB [20] and Basis Pursuit (BP) [6]. The source code of L1-LS, SPGL1 and SL0 algorithms are available online at http://dsp.rice.edu/cs For L1-LS, SPGL1 and SL0, the packages ell – 1LS, SPGL1 and SL0 are used. For the BCS, TSW-CS-MCMC and TSW-CS-VB, the packages bcs_ver0.1, tswcs (for wavelet) and bcs_vb at http://people.ee.duke.edu/lecarin/BCS.html are utilized, respectively. The required algorithm parameters for the above methods are set according to their default setups. All softwares are written in MATLAB, and run on a PC with 3.6 GHz CPU and 16 GB of memory. Note that the conventional Basis Pursuit (BP) method [6] have been also evaluated. However, due to the large errors and CPU time of this method, the corresponding results are not presented here.

To evaluate the proposed multiscale TSCW-GDP-HMT algorithm, the Microsoft database [34] is used. Five image classes which contain flowers, cows, buildings, urban, and office are considered here. Ten images from each set of these classes are randomly selected which are depicted in Fig. 5. The reconstruction error is defined as $\|x - \tilde{x}\|^2$ and presented as a function of CS measurements, where $x$ is the underlying image and $\tilde{x}$ is the reconstructed image using a particular CS algorithm.

The results are shown in (Figs. (6) and 7) for the first “flowers” image in the top-left side of Fig. 5. Each point in (Figs. (7) and 8) is the average of 10 trials. As is evident, by employing a powerful structure for the wavelet coefficients, the relative reconstruction error reduces significantly, even for a small number of measurements. The VB procedure decreases the required CPU time considerably, with a comparatively small cost of recovery quality. The VB inference, generally converges after 40
iterations. However MCMC needs at least 300 iterations to converge. Therefore VB is the choice in the applications that CPU time is the main concern and the reconstruction error is not as important as computational time. Fig. 6 shows that the TSCW-GDP-HMT algorithm has a lower relative reconstruction error compared with all other ones for CS measurements greater than 2800. As mentioned before, the reason is that VB inference is an approximate method which needs more CS measurements to error levels superior to MCMC’s. It has the minimum CPU time against all the other algorithms.

Table 1 includes the average performance of all algorithms, based on all 50 images in Fig. 5. In section a of this table where the CS measurement number is 3000 which is a relatively small number, the TSCW-GDP-HMT algorithm considerably outperforms the other algorithms. Moreover, in section b where the CS measurement number is 6000 (a relatively large number), the mean of reconstruction error in TSCW-GDP-HMT is lower than all the other ones.

6.2. Performance with noisy measurements

In this scenario, the CS measurement contains additive white Gaussian noise \(\eta\):

\[
y = y + \eta.
\]

where \(y\) is the noise free CS measurements. Each element of \(\eta\) is sampled from an i.i.d. Gaussian distribution \(\mathcal{N}(0, \lambda_n^{-1})\) with precision \(\lambda_n^{-1}\). Therefore the signal to noise ratio (SNR) is defined as

\[
\text{SNR (dB)} = 20 \log_{10} \left( \frac{\|y\|_2^2}{\|\eta\|_2^2} \right).
\]

Fig. 9 represents the relative reconstruction error versus SNR when the number of CS measurements is 4000. The proposed TSCW-GDP-HMT outperforms all the other algorithms in noisy scenario. It is less sensitive to noise, and especially performs better than the other algorithms when SNR is low. This fact is attributed to the natural property of Bayesian regression model that assumes noisy measurements as input signal.

<table>
<thead>
<tr>
<th>Class</th>
<th>Algorithm</th>
<th>Proposed</th>
<th>TSW-VB</th>
<th>TSW-MCMC</th>
<th>FAST-BCS</th>
<th>L1-LS</th>
<th>SPGL1</th>
<th>SLO</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) (N=3000)</td>
<td>Flowers Mean</td>
<td>0.1525</td>
<td>0.2320</td>
<td>0.2483</td>
<td>0.3335</td>
<td>0.3256</td>
<td>0.3256</td>
<td>0.3183</td>
</tr>
<tr>
<td></td>
<td>STD</td>
<td>0.0741</td>
<td>0.0835</td>
<td>0.0798</td>
<td>0.1245</td>
<td>0.0982</td>
<td>0.0982</td>
<td>0.0990</td>
</tr>
<tr>
<td></td>
<td>Cows Mean</td>
<td>0.1056</td>
<td>0.1518</td>
<td>0.1548</td>
<td>0.1920</td>
<td>0.3740</td>
<td>0.2091</td>
<td>0.2124</td>
</tr>
<tr>
<td></td>
<td>STD</td>
<td>0.0386</td>
<td>0.0495</td>
<td>0.0436</td>
<td>0.0787</td>
<td>0.2874</td>
<td>0.0700</td>
<td>0.0643</td>
</tr>
<tr>
<td></td>
<td>Buildings Mean</td>
<td>0.0980</td>
<td>0.1275</td>
<td>0.1345</td>
<td>0.1502</td>
<td>0.2788</td>
<td>0.1725</td>
<td>0.1751</td>
</tr>
<tr>
<td></td>
<td>STD</td>
<td>0.0115</td>
<td>0.0553</td>
<td>0.0510</td>
<td>0.0660</td>
<td>0.1179</td>
<td>0.0729</td>
<td>0.0691</td>
</tr>
<tr>
<td></td>
<td>Urban Mean</td>
<td>0.1148</td>
<td>0.1583</td>
<td>0.1607</td>
<td>0.1924</td>
<td>0.2682</td>
<td>0.2154</td>
<td>0.2066</td>
</tr>
<tr>
<td></td>
<td>STD</td>
<td>0.0245</td>
<td>0.0336</td>
<td>0.0317</td>
<td>0.0473</td>
<td>0.1477</td>
<td>0.0438</td>
<td>0.0450</td>
</tr>
<tr>
<td></td>
<td>Office Mean</td>
<td>0.0998</td>
<td>0.1487</td>
<td>0.1663</td>
<td>0.1761</td>
<td>0.2446</td>
<td>0.2022</td>
<td>0.2104</td>
</tr>
<tr>
<td></td>
<td>STD</td>
<td>0.0271</td>
<td>0.0362</td>
<td>0.0309</td>
<td>0.0523</td>
<td>0.0336</td>
<td>0.0436</td>
<td>0.0412</td>
</tr>
<tr>
<td>(b) (N=6000)</td>
<td>Flowers Mean</td>
<td>0.0945</td>
<td>0.1532</td>
<td>0.1793</td>
<td>0.2052</td>
<td>0.2454</td>
<td>0.1825</td>
<td>0.1679</td>
</tr>
<tr>
<td></td>
<td>STD</td>
<td>0.0624</td>
<td>0.0786</td>
<td>0.0677</td>
<td>0.0964</td>
<td>0.0802</td>
<td>0.0768</td>
<td>0.0788</td>
</tr>
<tr>
<td></td>
<td>Cows Mean</td>
<td>0.0771</td>
<td>0.1146</td>
<td>0.1156</td>
<td>0.1383</td>
<td>0.1869</td>
<td>0.1255</td>
<td>0.1184</td>
</tr>
<tr>
<td></td>
<td>STD</td>
<td>0.0295</td>
<td>0.0445</td>
<td>0.0319</td>
<td>0.0409</td>
<td>0.0389</td>
<td>0.0476</td>
<td>0.0462</td>
</tr>
<tr>
<td></td>
<td>Buildings Mean</td>
<td>0.0628</td>
<td>0.0837</td>
<td>0.1016</td>
<td>0.1010</td>
<td>0.2773</td>
<td>0.0931</td>
<td>0.0833</td>
</tr>
<tr>
<td></td>
<td>STD</td>
<td>0.0326</td>
<td>0.0440</td>
<td>0.0396</td>
<td>0.0437</td>
<td>0.1173</td>
<td>0.0445</td>
<td>0.417</td>
</tr>
<tr>
<td></td>
<td>Urban Mean</td>
<td>0.0776</td>
<td>0.1112</td>
<td>0.1205</td>
<td>0.1309</td>
<td>0.1592</td>
<td>0.1199</td>
<td>0.1100</td>
</tr>
<tr>
<td></td>
<td>STD</td>
<td>0.0159</td>
<td>0.0257</td>
<td>0.0231</td>
<td>0.0270</td>
<td>0.0884</td>
<td>0.0267</td>
<td>0.0265</td>
</tr>
<tr>
<td></td>
<td>Office Mean</td>
<td>0.0567</td>
<td>0.0629</td>
<td>0.1208</td>
<td>0.0990</td>
<td>0.3075</td>
<td>0.0948</td>
<td>0.0806</td>
</tr>
<tr>
<td></td>
<td>STD</td>
<td>0.0213</td>
<td>0.0365</td>
<td>0.0307</td>
<td>0.0252</td>
<td>0.0827</td>
<td>0.0249</td>
<td>0.0236</td>
</tr>
</tbody>
</table>
7. Conclusion

Using Bayesian modelling, a novel three level hierarchical model leading to a higher degree of sparsity, has been proposed as TSCW-GDP-HMT algorithm which can model and estimate the signal coefficients efficiently. Using this framework, the statistical relation between the complex wavelet coefficients can be captured efficiently by a two-state Markovian structure which is particularly developed for variational Bayes inference. Employing this model in the inversion part, can reduce the reconstruction error considerably. The sparse complex wavelet coefficients are estimated along with the unknown model parameters in the proposed TSCW-GDP-HMT algorithm, simultaneously. Extensive simulation results based on natural images confirmed the effectiveness and superiority of the proposed algorithm.

Appendix A

The Likelihood function is denoted by

\[
\mathcal{L} = -\log 2\zeta - (\alpha + 1) \log \left(1 + \frac{|\theta|}{\alpha \zeta}\right),
\]

\[
\frac{\partial \mathcal{L}}{\partial \zeta} = -\frac{1}{\zeta} - (\alpha + 1) \frac{|\theta|}{\alpha \zeta^2} = 0,
\]

\[
\hat{\zeta} = \frac{1}{k} \sum_{i=1}^{k} |\theta_i|,
\]

\[
\frac{\partial \mathcal{L}}{\partial \alpha} = \log \left(1 + \frac{|\theta|}{\alpha \zeta}\right) + (\alpha + 1) \frac{|\theta|}{\alpha^2 \zeta^2} = 0
\]

By using the Maclaurin series, the above equation can be written as

\[
\log \left(1 + \frac{|\theta|}{\alpha \zeta}\right) = \frac{|\theta|}{\alpha \zeta} - \frac{|\theta|^2}{2 \alpha^2 \zeta^2}
\]

\[
\hat{\alpha} = \left(2\zeta^3 + \zeta^2 |\theta|\right) + \sqrt{\left(2\zeta^3 + \zeta^2 |\theta|\right)^2 - 16\zeta^5 |\theta|^2}
\]

Appendix B

From (24), the wavelet coefficient \( \theta \) can be derived as

\[
\ln \hat{q}(\theta) \propto \ln p(y(\theta, z, \lambda_n) p(\theta|0, \tau))_{x, t} + \lambda_n
\]

\[
\propto -\frac{\lambda_n}{2} |y - \Phi(\theta \odot z)|^2 - \frac{\theta^2}{2\tau}
\]

\[
\propto \langle \lambda_n \frac{\phi(\theta \odot z)}{2} + \left( -\frac{\lambda_n}{2}\phi^2 - \frac{1}{2\tau}\right)\theta^2 \rangle_{x, t}
\]

Fig. 8. Examples of reconstructed flower image using different CS inversion algorithm. (a) Original Image. (b) TSW-CS-MCMC, Err=0.0887. (c) RVM, Err=0.0878. (d) SPGL1, Err=0.0659. (e) TSCW-GDP, Err=0.0242.

Fig. 9. Comparison of the relative reconstruction error given number of measurements N=4000, with noisy measurements.
\[ \alpha \left( \frac{\lambda_0}{2} \right) \text{exp} \left( -\frac{1}{2} \left( \frac{\theta^2}{\tau} + \lambda^2 \tau \right) \right) \]

\[ \kappa_1(\lambda) = \kappa_0(\lambda) \left( \frac{1}{\sqrt{\lambda}} \right) \]

\[ E[\tau^1] = \kappa_{1.5}(\lambda^2) \left( \frac{\lambda^2}{\theta^2} \right)^{1/2} \]

\[ \kappa_1(\lambda) = \kappa_0(\lambda) \left( 1 + \frac{1}{\sqrt{\lambda^2}} \right) \]

\[ \langle \tau \rangle = \left( 1 + \frac{1}{\sqrt{\lambda^2}} \right) \left( \frac{\lambda^2}{\theta^2} \right)^{1/2} \]

\[ \langle \tau \rangle = \frac{1}{\sqrt{\lambda^2}} + \frac{1}{\theta^2} - 1 \]

**Appendix E**

From (30), by using variational Bayes inference for \( \lambda \)

\[ \ln \hat{q}_A(\lambda) \propto \ln \left( \frac{\lambda^{\alpha + 1}}{\tau} \text{exp} \left( -\frac{\lambda^2 \tau}{2} - \eta \lambda \right) \right) \]

\[ \propto \ln \left( \frac{\lambda^{\alpha + 1}}{\tau} \right) - \frac{\lambda^2 \tau}{2} - \eta \lambda \]

The above equations consists of logarithmic, square and linear terms, thus does not assign to any standard distribution. However, when \( \alpha \) is set to \(-1\), noninformative hyper-prior is suggested for it. \( \hat{q}(\lambda) \) becomes the Gaussian distribution and the following parameters for this distribution is derived:

\[ \mu_\lambda = -\frac{\eta}{\langle \tau \rangle} \]

\[ \alpha_\lambda = \langle \tau \rangle \]

**Appendix F**

\[ q(z_{s,i}) = \text{Bernoulli}(z_{s,i} | p_{s,i}) \]

\[ p_{s,i} = \left[ 1 + \exp \left( \langle \ln (1 - \pi_{s,i}) \rangle - \langle \ln (\pi_{s,i}) \rangle \right) \right]^{-1} \left( \frac{\langle \lambda_0 \rangle}{2} \right) \left( \Phi(a_0 \lambda_i \Phi(z_i \lambda_i)) \right) \]

\[ q(\pi) = \prod_{s=0}^{s=2} \beta(\pi^0 | e^0, f^0) \prod_{s=1}^{s=L} \left( \beta(\pi^s | e^0, f^0) \beta(\pi^1 | e^1, f^1) \right) \]
\[ f^{(0)} = f_0^{(1)} + \sum_{i=1}^L \left( \langle z_{pa}(s, i) \rangle (1 - \langle z(s, i) \rangle) \right), \quad s = 2, \ldots, L. \]  

(47)

References


[34] Available at [http://research.microsoft.com/en-us/projects/objectclassrecognition/].