Variational Bayesian Compressed Sensing for Sparse and Locally Constant Signals

Shunsuke Horii*
* Waseda University, Tokyo, Japan
E-mail: s.horii@aoni.waseda.jp

***Abstract*—In this paper, we deal with the signal recovery problem in compressed sensing, that is, the problem of estimating the original signal from its linear measurements. Recovery algorithms can be mainly classified into two types, optimization based algorithms and statistical modeling based algorithms. Basis pursuit (BP) or basis pursuit denoising (BPDN) is one of the most widely used optimization based recovery algorithms, that minimizes the ℓ_1 norm of the signal or its coefficients in some basis under the constraint that its linear transform is equal to or close to the observation signal. There are various extensions of those algorithms depending on the problem structure. When the original signal is an image, the objective function is often the sum of the ℓ_1 norm of the coefficients of the signal in some basis and a total variation (TV) of the image. It can be considered that it requires the image to be sparse in both the specific transform domain and finite differences at the same time. In this paper, we propose a statistical model that represents those sparsities and the signal recovery algorithm based on the variational method. One of the advantages of the statistical approach is that we can utilize the posterior information of the original signal and it is known that it can be used to construct the compressed sensing measurements adaptively. The proposed recovery algorithm and adaptive construction of the compressed sensing measurements are validated on numerical experiments.**

I. Introduction

In this paper, we consider the following linear model,

$$\boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{x} \in \mathbb{R}^n$ is an original signal and it is linearly transformed by measurement matrix $A \in \mathbb{R}^{m\times n}$ and then a noise $\boldsymbol{\epsilon}$ is added. As a result, we obtain an observation signal $\boldsymbol{y} \in \mathbb{R}^m$. We assume that the noise vector follows multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{0}, s^{-1}I_m)$, where I_m is an $m \times m$ identity matrix and s is the precision parameter of the noise and $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ denotes multivariate Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ. In the problem of compressed sensing, it is assumed that $\boldsymbol{x}$ is sparse in some basis. More precisely, let $n \times n$ matrix W represent a sparsity inducing linear transform such as the wavelet transform and the linear transform $\boldsymbol{z} = W\boldsymbol{x}$ is assumed to be sparse. Recovery algorithms estimate the original signal $\boldsymbol{x}$ from the observation signal $\boldsymbol{y}$ and existing algorithms can be mainly classified into two types, optimization based algorithms and statistical modeling based algorithms.

Thanks to No. 16K00417 of Grant-in-Aid for Scientific Research Category (C), Japan Society for the Promotion of Science for funding.

Basis pursuit denoising (BPDN) is one of the most widely used optimization based recovery algorithms and the recovered signal is obtained by solving the following constrained optimization problem [1][1]:

$$\begin{array}{ll} \text{minimize} & ||W\boldsymbol{x}||_1 \\ \text{s.t.} & ||\boldsymbol{y} - A\boldsymbol{x}||_2 < \delta, \end{array} \tag{2}$$

where δ is the threshold parameter and it is usually set below the expected noise level.

Let $\boldsymbol{x}$ represent some kind of gray scale 2-D image. That is, let $x_{j,k}$ represent the signal value of (j,k) pixel of $J \times K$ image and $\boldsymbol{x}$ is a vector formed by rearranging them. In such a case, total variation (TV) is used as a sparsity inducing transform. TV for $\boldsymbol{x}$ is defined as

$$TV(\boldsymbol{x}) = \sum_{j,k} \sqrt{|x_{j+1,k} - x_{j,k}|^2 + |x_{j,k+1} - x_{j,k}|^2}. \tag{3}$$

An anisotropic version is sometimes used since it may sometimes be easier to minimize and it is defined as

$$TV(\boldsymbol{x}) = \sum_{j,k} \left(|x_{j+1,k} - x_{j,k}| + |x_{j,k+1} - x_{j,k}|\right). \tag{4}$$

Hereafter, $TV(\cdot)$ denotes the anisotropic version of total variation. By adding TV penalty, the estimated values of adjacent pixels tend to take similar values. It is often useful to combine other sparsity inducing linear transform with TV penalty [2] [3]. This is considered as requiring the image to be sparse in both the specific transform domain and finite differences at the same time. In this case, the recovered signal is obtained by solving the following optimization problem:

$$\begin{array}{ll} \text{minimize} & ||W\boldsymbol{x}||_1 + \lambda TV(\boldsymbol{x}) \\ \text{s.t.} & ||\boldsymbol{y} - W\boldsymbol{x}||_2 < \delta, \end{array} \tag{5}$$

where λ trades sparsity in the domain of W with finite differences sparsity. This optimization problem can be seen as the optimization problem in graph guided fused lasso [4].

In statistical modeling approach, it models the probability distribution of the original signal $\boldsymbol{x}$. It is well known that solving (2) is equivalent to finding a maximum a posterior (MAP) estimator of $\boldsymbol{x}$ assuming Laplace prior [5],

$$p(\boldsymbol{x}) \propto \prod_{i=1}^{n} \exp\left(-\frac{\alpha}{2}|\boldsymbol{w}_i^T\boldsymbol{x}|\right), \tag{6}$$

[1] In general, it is assumed that $W = I_n$.

where $\boldsymbol{w}_i$ is the i-th row of W. It is beneficial to consider the problem of finding not only the MAP estimator but also the posterior distribution of the original signal since its information is useful to many decision making problems, such as adaptive compressed sensing [5]. However, it is very difficult to find the posterior distribution of the original signal since it does not have an analytic form. Figueiredo used the fact that the Laplace distribution can be expressed as a Gaussian scale mixture and developed an estimator using the EM algorithm [6]. Similar model is assumed in [5] [7] [8] [9], but type-Ⅱ ML estimator is used in [5], variational method is used in [7] [8], and Gibbs sampling is used in [9]. Similarly, Gibbs sampling for group lasso, fused lasso, and elastic net is proposed in [10] and variational method for group lasso is proposed in [11].

In this paper, we propose a hierarchical prior model that expresses the following properties:

- The original signal is sparse in the transform domain with the linear transform W.
- The adjacent pixels of the original signal tend to take similar values. In other words, the original signal is locally almost constant over 2-D grid.

It is a statistical model that corresponds to the optimization problem (5) and it is also a generalization of the model proposed in [10]. We also propose an approximation algorithm to find the posterior distribution of the original signal based on the variational method. As an output of the algorithm, not only the estimator of the original signal but also an approximate covariance matrix of its posterior distribution is obtained. As in [5] [8], this information can be used to design the measurement matrix adaptively.

The rest of the paper is organized as follows. In Section 2, we describe the hierarchical model for the signal that is sparse in some specific transform domain and locally almost constant over a predefined graph. In Section 3, we establish an approximation algorithm for computing the posterior distribution of the original signal based on the variational method. We briefly review how the information of the posterior distribution can be used for the adaptive design of the measurements in Section 4. Some performance analysis of the proposed algorithm and adaptive design based on numerical experiments are made in Section 5. We conclude the paper in Section 6.

II. Hierarchical Model For Sparse and Locally Constant Signal and Observation Model

Solving the optimization problem (5) is equivalent to finding the MAP estimator with the assumption that the prior distribution of $\boldsymbol{x}$ is the following distribution and the parameters are appropriately set.

$$p(\boldsymbol{x}) \propto \prod_{i=1}^{n} \exp\left(-\frac{\alpha}{2}|\boldsymbol{w}_i^T\boldsymbol{x}|\right) \prod_{(j,k)\in E} \exp\left(-\frac{\beta}{2}|x_j - x_k|\right), \tag{7}$$

where E is the set of pairs (j,k) such that x_j and x_k are adjacent pixels if $(j,k) \in E$. Unfortunately, if the prior distribution (7) is assumed, it is very difficult to calculate the posterior distribution even if we resort to approximation methods. Instead of that, we assume the following hierarchical prior distribution. Conditioned on the parameters $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_n) \in \mathbb{R}^n$, $\boldsymbol{\nu} = (\nu_{jk})_{(j,k)\in E} \in \mathbb{R}^{|E|}$, we assume that $\boldsymbol{x}$ follows the following distribution.

$$p(\boldsymbol{x}|\boldsymbol{\tau},\boldsymbol{\nu}) \propto \prod_{i=1}^{n} \exp\left(-\frac{(\boldsymbol{w}_i^T\boldsymbol{x})^2}{2\tau_i}\right) \prod_{(j,k)\in E} \exp\left(-\frac{(x_j - x_k)^2}{2\nu_{jk}}\right). \tag{8}$$

This is equivalent to assume that $p(\boldsymbol{x}|\boldsymbol{\tau},\boldsymbol{\nu})$ is the multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{0}, S_{\boldsymbol{\tau},\boldsymbol{\nu}}^{-1})$, where $S_{\boldsymbol{\tau},\boldsymbol{\nu}}$ is the matrix defined by

$$S_{\boldsymbol{\tau},\boldsymbol{\nu}} = W^T \mathrm{diag}(\tau_1^{-1}, \ldots, \tau_n^{-1}) W + L_{\boldsymbol{\nu}}, \tag{9}$$

and $L_{\boldsymbol{\nu}}$ is the matrix whose (j,k) element is given by

$$(\boldsymbol{L}_{\boldsymbol{\nu}})_{j,k} = \begin{cases} \displaystyle\sum_{(j',k')\in N(j)} \nu_{j'k'}^{-1} & \text{if } j = k \\ -\nu_{jk}^{-1} & \text{if } (j,k) \in E \\ 0 & \text{otherwise,} \end{cases} \tag{10}$$

where $N(j) = \{k \mid (j,k) \in E \text{ or } (k,j) \in E\}$. We further assume that $\boldsymbol{\tau}$ and $\boldsymbol{\nu}$ follow the following distributions.

$$p(\boldsymbol{\tau}|a_\tau, b_\tau, \rho_\tau) = \prod_{i=1}^{n} \mathrm{GIG}(\tau_i|a_\tau, b_\tau, \rho_\tau), \tag{11}$$

$$p(\boldsymbol{\nu}|a_\nu, b_\nu, \rho_\nu) = \prod_{(j,k)\in E} \mathrm{GIG}(\nu_{jk}|a_\nu, b_\nu, \rho_\nu), \tag{12}$$

where $\mathrm{GIG}(\cdot|a,b,\rho)$ denotes the generalized inverse Gaussian distribution with parameters a, b, ρ. The probability density function of the generalized inverse Gaussian distribution is given by

$$\mathrm{GIG}(x|a,b,\rho) \propto x^{\rho-1} \exp\left(-\frac{1}{2}(ax + bx^{-1})\right). \tag{13}$$

As a special case, the generalized inverse Gaussian distribution coincides with the exponential distribution when $b \to 0$ and $\rho = 1$. In such a case, the marginal distribution of $\boldsymbol{x}$ is given by

$$p(\boldsymbol{x}|a_\tau, a_\nu) \propto \prod_{i=1}^{n} \exp\left(-\frac{\sqrt{a_\tau}}{2}|x_i|\right) \prod_{(j,k)\in E} \exp\left(-\frac{\sqrt{a_\nu}}{2}|x_j - x_k|\right). \tag{14}$$

Thus, the proposed model (8) (11) (12) is an extension of (7). Another important case is when $a \to 0$ and $\rho < 0$, and in this case, the generalized inverse Gaussian distribution coincides with the inverse gamma distribution. Consequently, the proposed model includes various models in the past studies.

- When $W = I_n$, $\boldsymbol{\nu} = \boldsymbol{0}$, $a_\tau \to 0$, and $\rho_\tau < 0$, the proposed model coincides with the model in [5] [7] [9].

- When $W = I_n$, $E = \{(1,2),(2,3),\ldots,(n-1,n)\}$, $a_\tau, a_\nu \to 0$, and $\rho_\tau, \rho_\nu < 0$, the proposed model coincides with the model in [10].

For the observation signal $\boldsymbol{y}$, we assume the model (1) and the conditional distribution of $\boldsymbol{y}$ conditioned on $\boldsymbol{x}$ and s is given by

$$p(\boldsymbol{y}|\boldsymbol{x}, s) = \mathcal{N}(\boldsymbol{y}|A\boldsymbol{x}, s^{-1}I_m). \tag{15}$$

For the precision parameter s of the noise, we assume the gamma distribution so that it is conjugate prior for $p(\boldsymbol{y}|\boldsymbol{x}, s)$, so

$$p(s|k_s, \theta_s) = \mathrm{Ga}(s|k_s, \theta_s) \propto s^{k_s-1}\exp(-\theta_s s) \tag{16}$$

In summary, the following joint distribution is obtained.

$$\begin{aligned} p(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\tau}, \boldsymbol{\nu}, s) = p(\boldsymbol{y}|\boldsymbol{x}, s)p(\boldsymbol{x}|\boldsymbol{\tau}, \boldsymbol{\nu})\cdot \\ p(\boldsymbol{\tau}|a_\tau, b_\tau, \rho_\tau)p(\boldsymbol{\nu}|a_\nu, b_\nu, \rho_\nu)p(s|k_s, \theta_s). \end{aligned} \tag{17}$$

In this paper, $k_s, \theta_s, a_\tau, a_\nu, b_\tau, b_\nu, \rho_\tau, \rho_\nu$ are treated as hyper-parameters.

III. Variational Inference

Given the joint distribution (17), what we want is the posterior distribution $p(\boldsymbol{x}|\boldsymbol{y})$, however, we have to perform a complex integral calculation to find the posterior and it is very hard. In this paper, we give an approximation algorithm based on the variational Bayesian method [12]. Let $\boldsymbol{\xi} = (\boldsymbol{x}, \boldsymbol{\tau}, \boldsymbol{\nu}, s)$ and the variational Bayesian method finds an approximation distribution $q(\boldsymbol{\xi})$ that approximates $p(\boldsymbol{\xi}|\boldsymbol{y})$. More specifically, the goal is to find $q(\boldsymbol{\xi})$ that minimizes the Kullback-Leibler divergence $\mathrm{KL}(q(\boldsymbol{\xi})||p(\boldsymbol{\xi}|\boldsymbol{y}))$:

$$q^*(\boldsymbol{\xi}) = \operatorname*{argmin}_{q(\boldsymbol{\xi})} \int q(\boldsymbol{\xi}) \ln \frac{q(\boldsymbol{\xi})}{p(\boldsymbol{\xi}|\boldsymbol{y})} d\boldsymbol{\xi} \tag{18}$$

$$= \operatorname*{argmin}_{q(\boldsymbol{\xi})} \int q(\boldsymbol{\xi}) \ln \frac{q(\boldsymbol{\xi})}{p(\boldsymbol{\xi}, \boldsymbol{y})} d\boldsymbol{\xi}. \tag{19}$$

However, it is difficult to minimize (19) for arbitrary probability distributions. In this paper, we limit the optimization distributions to $q(\boldsymbol{\xi})$ that can be factorized as follows.

$$q(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\nu}, s) = q(\boldsymbol{\beta})q(\boldsymbol{\tau}, \boldsymbol{\nu})q(s). \tag{20}$$

For $\boldsymbol{\xi}_k \in \boldsymbol{\xi}$, the variational Bayes method minimizes (19) by updating $q(\boldsymbol{\xi}_k)$ sequentially. With the distribution $q(\boldsymbol{\xi} \setminus \boldsymbol{\xi}_k)$ of $\boldsymbol{\xi} \setminus \boldsymbol{\xi}_k$ fixed, the update equation of $q(\boldsymbol{\xi}_k)$ is given as follows [12].

$$\ln q^*(\boldsymbol{\xi}_k) = E_{q(\boldsymbol{\xi}\setminus\boldsymbol{\xi}_k)}[\ln p(\boldsymbol{y}, \boldsymbol{\xi})] + \text{const.} \tag{21}$$

In the following, we describe concrete update equation of each $q(\boldsymbol{\xi}_k)$. To keep the description concise, for functions $f(\boldsymbol{\xi}_k)$ of $\boldsymbol{\xi}_k$, the expectation taken by $q(\boldsymbol{\xi}_k)$ at the point is written as $\langle f(\boldsymbol{\xi}_k)\rangle$.

A. Update equation of $q(\boldsymbol{x})$

From (21), the update equation of $q(\boldsymbol{x})$ is

$$\ln q^*(\boldsymbol{x}) = E_{q(\boldsymbol{\xi}\setminus\boldsymbol{x})}[\ln(p(\boldsymbol{y}|\boldsymbol{x}, s)p(\boldsymbol{x}|\boldsymbol{\tau}, \boldsymbol{\nu}))] + \text{const.}. \tag{22}$$

Using the model assumption that $p(\boldsymbol{y}|\boldsymbol{x}, s)$ and $p(\boldsymbol{x}|\boldsymbol{\tau}, \boldsymbol{\nu})$ are Gaussian distributions, we obtain

$$q^*(\boldsymbol{x}) = \mathcal{N}(\bar{\boldsymbol{x}}, \Sigma_{\boldsymbol{x}}), \tag{23}$$

$$\bar{\boldsymbol{x}} = \langle s\rangle\ \Sigma_{\boldsymbol{x}} A^T \boldsymbol{y}, \tag{24}$$

$$\Sigma_{\boldsymbol{x}} = \left(\langle s\rangle\ A^T A + \langle S_{\boldsymbol{\tau},\boldsymbol{\nu}}\rangle\right)^{-1}. \tag{25}$$

B. Update equation of $q(\boldsymbol{\tau}, \boldsymbol{\nu})$

From (21), the update equation of $q(\boldsymbol{\tau}, \boldsymbol{\nu})$ is

$$\begin{aligned} &\ln q^*(\boldsymbol{\tau}, \boldsymbol{\nu}) = \\ &E_{q(\boldsymbol{\xi}\setminus\boldsymbol{\tau},\boldsymbol{\nu})}[\ln(p(\boldsymbol{x}|\boldsymbol{\tau}, \boldsymbol{\nu})p(\boldsymbol{\tau}|a_\tau, b_\tau, \rho_\tau)p(\boldsymbol{\nu}|a_\nu, b_\nu, \rho_\nu))] \\ &+ \text{const.}. \end{aligned} \tag{26}$$

From (7), (11), and (12), without loss of generality, we can assume that $q(\boldsymbol{\tau}, \boldsymbol{\nu})$ is decomposed as follows.

$$q(\boldsymbol{\tau}, \boldsymbol{\nu}) = \prod_{i=1}^{n} q(\tau_i) \prod_{(j,k)\in E} q(\nu_{jk}). \tag{27}$$

By arranging the terms in (26) that include τ_i, ν_{jk}, we obtain

$$q^*(\tau_i) = \mathrm{GIG}\left(a_\tau, b_\nu + \left\langle(\boldsymbol{w}_i^T\boldsymbol{x})^2\right\rangle, \rho_\tau - \frac{1}{2}\right), \tag{28}$$

$$q^*(\nu_{jk}) = \mathrm{GIG}\left(a_\nu, b_\nu + \left\langle(x_j - x_k)^2\right\rangle, \rho_\nu - \frac{1}{2}\right). \tag{29}$$

In order to update $q(\boldsymbol{x})$, we need the expected values $\left\langle\tau_j^{-1}\right\rangle, \left\langle\nu_{jk}^{-1}\right\rangle$. We describe the analytic forms of these variables for some special cases that are used in the experiments.

1. $b_\tau, b_\nu \to 0, \rho_\tau = \rho_\nu = 1$:

$$\left\langle\tau_i^{-1}\right\rangle = \frac{\sqrt{a_\tau}}{\sqrt{\left\langle(\boldsymbol{w}_i^T\boldsymbol{x})^2\right\rangle}}, \tag{30}$$

$$\left\langle\nu_{jk}^{-1}\right\rangle = \frac{\sqrt{a_\nu}}{\sqrt{\left\langle(x_j - x_k)^2\right\rangle}}. \tag{31}$$

2. $a_\tau, a_\nu \to 0, \rho_\tau, \rho_\nu < 0$:

$$\left\langle\tau_i^{-1}\right\rangle = \frac{\frac{1}{2} - \rho_\tau}{\frac{1}{2}\left(b_\tau + \left\langle(\boldsymbol{w}_i^T\boldsymbol{x})^2\right\rangle\right)}, \tag{32}$$

$$\left\langle\nu_{jk}^{-1}\right\rangle = \frac{\frac{1}{2} - \rho_\nu}{\frac{1}{2}(b_\nu + \left\langle(x_j - x_k)^2\right\rangle)}. \tag{33}$$

C. Update equation of $q(s)$

From (21), the update equation of $q(s)$ is

$$\ln q^*(s) = E_{q(\boldsymbol{\xi}\setminus s)}[\ln(p(\boldsymbol{y}|\boldsymbol{x}, s)p(s|k_s, \theta_s))] + \text{const.}. \tag{34}$$

From the assumption that $p(s|k_s, \theta_s)$ is gamma distribution, we obtain

$$q^*(s) = \mathrm{Ga}\left(k_s + \frac{m}{2}, \theta_s + \frac{1}{2}\left\langle||\boldsymbol{y} - A\boldsymbol{x}||_2^2\right\rangle\right), \tag{35}$$

$$\left\langle||\boldsymbol{y} - A\boldsymbol{x}||_2^2\right\rangle = ||\boldsymbol{y} - A\bar{\boldsymbol{x}}||_2^2 + \mathrm{Tr}(A^T A\Sigma_{\boldsymbol{x}}). \tag{36}$$

The expected value $\langle s \rangle$ is given by

$$\langle s \rangle = \frac{k_s + \frac{m}{2}}{\theta_s + \frac{1}{2}\langle ||\boldsymbol{y} - A\boldsymbol{x}||_2^2 \rangle}. \quad (37)$$

IV. Adaptive Compressed Sensing

In this section, we consider the problem of how to design a new measurement or projection $\boldsymbol{a}_*$. That is, assuming that the measurement signal $\boldsymbol{y}$ for measurement matrix A is obtained and (approximate) posterior distribution $p(\boldsymbol{x}|\boldsymbol{y})$ is calculated, we consider the situation that we can choose the next measurement $\boldsymbol{a}_*$, which will be added to the measurement matrix A. The content in this section is a brief review of the contents in articles [5] [8]. We adopt the expected differential entropy of posterior distribution as an evaluation criterion for $\boldsymbol{a}_*$. This is defined as follows:

$$h(\boldsymbol{a}_*) = E_{p(\boldsymbol{y}_*|\boldsymbol{y},A,\boldsymbol{a}_*)} H(p(\boldsymbol{x}|\boldsymbol{y}, A, \boldsymbol{a}_*, y_*)), \quad (38)$$

where H denotes the differential entropy function. Assuming that $p(\boldsymbol{x}|\boldsymbol{y})$ is multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$, we have

$$h(\boldsymbol{a}_*) = -\ln\left(1 + s\boldsymbol{a}_*^T \Sigma \boldsymbol{a}_*\right) + \text{const.} \quad (39)$$

Therefore, if we can freely design $\boldsymbol{a}_*$, the optimal solution is to set the eigenvector of Σ with largest eigenvalue[2]. If we have to select $\boldsymbol{a}_*$ from a set of vectors $\tilde{A}$, the optimal solution is to select $\boldsymbol{a}_*$ that maximizes $\boldsymbol{a}_* \Sigma \boldsymbol{a}_*$ from the set $\tilde{A}$.

In our algorithm, the posterior distribution is approximated with multivariate Gaussian and it outputs the covariance matrix $\Sigma_{\boldsymbol{x}}$ (see (25)), this matrix can be directly used to design $\boldsymbol{a}_*$.

V. Experiments

A. Comparison with optimization based algorithm

The main objective of the experiments here is to examine the utility of our proposed method by showing that it properly works for the compressed sensing problem of a sparse image. We consider the problem to recover the original image of the 64×64 Shepp-Logan phantom image from its compressed and noisy observation (Fig. 1). Due to the computational cost, we divide the image into $2 \times 2 = 4$ regions and set each image as an original signal (therefore, $n = 1024$). Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_4$ denote these original signals. The original signals are multiplied by measurement matrix A and contaminated by Gaussian noise $\boldsymbol{\epsilon}$. The measurement matrix $A \in \mathbb{R}^{192 \times 1024}$ is constructed by drawing i.i.d. from the standard Gaussian distribution $\mathcal{N}(0, 1)$, and then each row of A is normalized to unit magnitude. The precision parameter of the noise vector is $s = 1.0\text{E}6$. Daubechies 4 wavelet [13] is used for W.

For the proposed algorithm, we set the parameters as follows. For the hyperparameters k_s, θ_s of the precision parameter of the noise, we set $k_s = 1.0\text{E}16$ and $\theta_s = 1.0\text{E}10$. Considering the correspondence with the optimization problem (5), we should set $b_\tau, b_\nu \to 0$ and $\rho_\tau = \rho_\nu = 1$. However, in such a case, we have an implementation issue in updating $\left\langle \tau_i^{-1} \right\rangle$ and $\left\langle \nu_{jk}^{-1} \right\rangle$ according to (30) and (31) since their denominators tend to take 0. Therefore, we consider the case where $a_\tau, a_\nu \to 0$ and $\rho_\tau, \rho_\nu < 0$. The values of $b_\tau, b_\nu, \rho_\tau, \rho_\nu$ are determined so that they maximize $p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{16}|b_\tau, b_\nu, \rho_\tau, \rho_\nu)$[3].

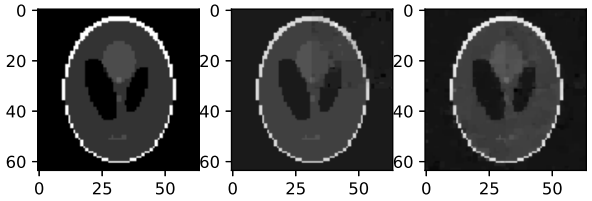

Fig. 1. Recovery results of Shepp-Logan image. Left: Original Image. Center: Recovered image based on the proposed algorithm. Right: Recovered image based on the optimization method (5).

TABLE I
PSNR [DB] AND SSIM OF RECOVERED IMAGES

method	Proposed	Optimization
MSE	5.03E-4	6.55E-4
PSNR	39.01	37.85
SSIM	0.9712	0.9671

For the optimization problem (5), the parameter δ is set to $1.0\text{E} - 3$, which is equal to the standard deviation of the noise. The value of the parameter λ is determined as follows. Considering the corresponding prior distribution (14), one way to determine the value of λ is to set $\lambda = \sqrt{a_\nu}/\sqrt{a_\tau}$. Thus, we consider the prior distribution (8) (11) (12) with $b_\tau, b_\nu \to 0$ and $\rho_\tau = \rho_\nu = 1$ and find the values of a_τ, a_ν that maximize $p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{16}|a_\tau, a_\nu)$, and then set $\lambda = \sqrt{a_\nu}/\sqrt{a_\tau}$. It is one of the advantages of the statistical approach that we can statistically determine the values of hyper parameters in this way.

Table I shows mean squared error (MSE), peak signal-to-noise ratio (PSNR) [dB] and structural similarity (SSIM) of the recovered images. Recovered images are shown in Fig. 1. From these results, we can see that the proposed algorithm is competitive with the optimization based method and the proposed method shows slightly better performance. However, more important point is that we can obtain the information of the posterior information of the unknown signal. We will see this in the next experiment.

B. Adaptive compressed sensing

As discussed in Section IV, the posterior information of $\boldsymbol{x}$ can be used to design the measurement matrix A. In this experiment, we study the performance of the design based on the posterior information. In this experiment, we divide the image into $4 \times 4 = 16$ regions and set each image as an original signal (therefore, $n = 256$). The initial 32 measurements are constructed by using the standard Gaussian distribution as in the previous experiment. The remaining 64 measurements

[2] We assume that $||\boldsymbol{a}_*|| = 1$.

[3] In practice, these parameters should be determined by using similar images obtained in advance.

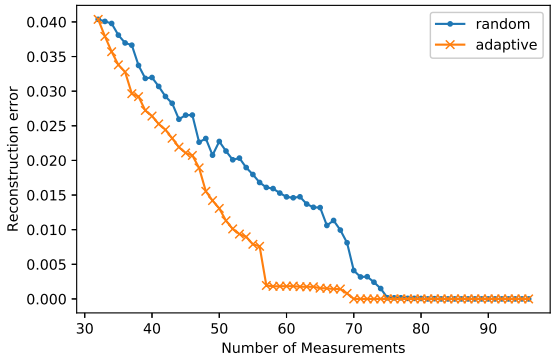

Fig. 2. Comparison of adaptive design and random design. The horizontal axis shows the number of measurements and the vertical axis shows the reconstruction error in terms of MSE.

are sequentially constructed by finding the eigenvector of the covariance matrix of the approximate posterior distribution with the maximum eigenvalue and then normalized to unit magnitude. It is compared to the random construction. Fig. 2 shows the reconstruction error (MSE) of the optimized design and random design. As with the result in [5], we can see that the reconstruction error of the optimized design is significantly smaller than that of the random design. This result indicates the effectiveness of using the information of the posterior distribution.

VI. Conclusions

In this paper, we proposed a hierarchical modeling for image signals that are sparse in a specific transform domain and finite differences at the same time. The proposed model includes various models of past studies as special cases. As an application of the proposed model, we considered the compressed sensing problem and developed an estimation algorithm for the original signal based on the variational inference method. Experiments results showed that the proposed method is comparative with the optimization based method. We also showed that the adaptive design for the compressed sensing problem based on the information of the posterior distribution works effectively.

Although the proposed scheme is powerful, one of the main drawbacks of it is its computational cost. In the algorithm, it requires the inversion of a matrix and it is not practical for very high dimensional problems. Constructing a reduced computational complexity algorithm is a future work.

References

[1] Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic decomposition by basis pursuit. *SIAM review*, 43(1):129–159, 2001.

[2] Yaakov Tsaig and David L Donoho. Extensions of compressed sensing. *Signal processing*, 86(3):549–571, 2006.

[3] Michael Lustig, David Donoho, and John M Pauly. Sparse mri: The application of compressed sensing for rapid mr imaging. *Magnetic resonance in medicine*, 58(6):1182–1195, 2007.

[4] Laurent Jacob, Guillaume Obozinski, and Jean-Philippe Vert. Group lasso with overlap and graph lasso. In *Proceedings of the 26th annual international conference on machine learning*, pages 433–440. ACM, 2009.

[5] Shihao Ji, Ya Xue, and Lawrence Carin. Bayesian compressive sensing. *IEEE Transactions on Signal Processing*, 56(6):2346–2356, 2008.

[6] Mário AT Figueiredo. Adaptive sparseness for supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 25(9):1150–1159, 2003.

[7] Christopher M Bishop and Michael E Tipping. Variational relevance vector machines. In *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence*, pages 46–53. Morgan Kaufmann Publishers Inc., 2000.

[8] Matthias W Seeger and Hannes Nickisch. Large scale bayesian inference and experimental design for sparse linear models. *SIAM Journal on Imaging Sciences*, 4(1):166–199, 2011.

[9] Trevor Park and George Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

[10] Minjung Kyung, Jeff Gill, Malay Ghosh, George Casella, et al. Penalized regression, standard errors, and bayesian lassos. *Bayesian Analysis*, 5(2):369–411, 2010.

[11] S Derin Babacan, Shinichi Nakajima, and Minh N Do. Bayesian group-sparse modeling and variational inference. *IEEE transactions on signal processing*, 62(11):2906–2921, 2014.

[12] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[13] Arne Jensen and Anders la Cour-Harbo. *Ripples in mathematics: the discrete wavelet transform*. Springer Science & Business Media, 2001.