NON-CONVEX PRIORS IN BAYESIAN COMPRESSED SENSING

S. Derin Babacan[1], *Luis Mancera*[2], *Rafael Molina*[2], *Aggelos K. Katsaggelos*[1]

[1]Dept. Electrical Engineering and Computer Science
Northwestern University
Evanston, IL 60208-3118
phone: (847) 491-7164 fax: (847) 491-4455
email: sdb@northwestern.edu, aggk@eecs.northwestern.edu

[2]Depto. Ciencias de la Computacion e IA
Universidad de Granada
18071 Granada, Spain
phone: +(34) 958 248 565, fax: + (34) 958 243 317
email: {mancera,rms}@decsai.ugr.es

ABSTRACT

We propose a novel Bayesian formulation for the reconstruction from compressed measurements. We demonstrate that high-sparsity enforcing priors based on l_p-norms, with $0 < p \leq 1$, can be used within a Bayesian framework by majorization-minimization methods. By employing a fully Bayesian analysis of the compressed sensing system and a variational Bayesian analysis for inference, the proposed framework provides model parameter estimates along with the unknown signal, as well as the uncertainties of these estimates. We also show that some existing methods can be derived as special cases of the proposed framework. Experimental results demonstrate the high performance of the proposed algorithm in comparison with commonly used methods for compressed sensing recovery.

1. INTRODUCTION

Compressed sensing (CS) proposes new techniques to acquire signals from a reduced number of samples. The theory of CS indicates that, if a signal is *compressible* in some basis, i.e., most of the energy of the coefficients in that basis is concentrated in relatively few coefficients, then high-accuracy recovery is possible even with fewer samples than the dimension of the signal [1]. Thus, the traditional sensing and compression phases of signal acquisition are merged into a single phase. The traditional decoding is replaced by recovery algorithms making use of the compressibility assumption.

During the last years, many recovery algorithms for compressed sensing problem have been proposed. A main class of algorithms is based on l_1 minimization via linear programming [1, 2, 3]. Another alternative is minimizing l_0-norms, either approximated by smooth functions (e.g., [4]), or directly utilized by iterative hard-thresholding methods [5]. Also, greedy methods have been widely proposed [6, 7], which approximate the signal by incrementally selecting the bases best describing the part not yet represented. These methods are computationally more efficient than global optimization methods, but generally at the expense of a decreased reconstruction accuracy. Finally, minimization of non-convex l_p-norms ($0 \leq p < 1$) has been shown to provide a potentially better recovery than l_1 norms [8]. These algorithms are known as *iteratively re-weighted least squares* (IRLS) methods. Early work on IRLS methods utilized l_p norms with $p > 1$ [9, 10], and extensions to non-convex optimization frameworks were proposed in [11, 12, 13]. A similar re-weighting approach is utilized for l_1-norms in [14].

A main issue of these methods is that the physical meaning of the model parameters is generally obscure. Recently, several algorithms have been developed within the Bayesian framework [15, 16, 17], with the advantage of systematic modeling of the unknown signal along with the model parameters, which results in fully-automated algorithms simultaneously estimating all required parameters. However, analytical difficulties within Bayesian inference limit the class of sparsity priors to Gaussian-based priors [16]. Those methods utilizing a more general class of sparsity priors generally resort to sampling algorithms for inference [18], which are generally computationally less efficient.

In this paper, we propose a novel Bayesian framework for CS recovery using non-convex l_p-norms. With a majorization-minimization approach, we demonstrate that Bayesian inference can be performed without resorting to sampling approaches. Specifically, we employ a variational Bayesian inference which provides distribution estimates of the unknowns, and therefore allows the calculation of the uncertainties of the estimates. The proposed algorithm simultaneously optimally estimates the unknown signal along with all needed parameters. We also show that existing IRLS methods are special cases of the proposed formulation. Finally, we demonstrate with experimental results that our algorithm compares favorably to commonly used CS recovery algorithms.

2. BAYESIAN MODELING

The CS acquisition system can be modeled as

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

where $\mathbf{x}$ is a $N \times 1$ sparse vector, $\mathbf{y}$ is the $M \times 1$ observation vector, $\mathbf{n}$ is the $M \times 1$ independent, Gaussian, zero-mean noise vector with variance equal to β^{-1}, and $\mathbf{\Phi}$ is the $M \times N$ measurement matrix, with $M < N$. A general form of the reconstruction problem is given by

$$\hat{\mathbf{x}} = \underset{\mathbf{x}}{\operatorname{argmin}}\{\|\mathbf{y} - \mathbf{\Phi}\mathbf{x}\|_2^2 + \tau\|\mathbf{x}\|_p^p\}, \tag{2}$$

where p is generally chosen to be within the interval $[0,2]$. We utilize a hierarchical Bayesian framework to model the components of the compressed acquisition system in (1). We first define the joint distribution $p(\mathbf{x}, \alpha, \beta, \mathbf{y})$ of all unknown and observed quantities, which we factorize as

$$\mathrm{p}(\mathbf{x}, \alpha, \beta, \mathbf{y}) = \mathrm{p}(\mathbf{y}|\mathbf{x}, \beta)\,\mathrm{p}(\mathbf{x}|\alpha)\,\mathrm{p}(\alpha)\,\mathrm{p}(\beta). \tag{3}$$

This work has been partially supported by the Spanish research programme Consolider Ingenio 2010: MIPRCV (CSD2007-00018) and the Ministerio de Educacion y Ciencia under contract TIN2007-65533.

In the first stage, the observation noise is modeled using the *conditional* distribution $\mathrm{p}(\mathbf{y}|\mathbf{x},\beta)$ and the unknown signal $\mathbf{x}$ is modeled by a sparsity prior $\mathrm{p}(\mathbf{x}|\alpha)$. These distributions depend on model parameters β and α, called *hyperparameters*, which are modeled in the second stage through *hyperpriors* $\mathrm{p}(\alpha)$ and $\mathrm{p}(\beta)$.

2.1 Observation and signal model

We model the conditional probability $p(\mathbf{y}|\mathbf{x},\beta)$ as

$$\mathrm{p}(\mathbf{y}|\mathbf{x},\beta) \propto \beta^{\frac{N}{2}} \exp\left[-\frac{\beta}{2}\|\mathbf{y}-\mathbf{\Phi}\mathbf{x}\|_2^2\right]. \tag{4}$$

The signal is assumed to be sparse, which is modeled using a Generalized Gaussian prior given by

$$\mathrm{p}(\mathbf{x}|\alpha) \propto \frac{1}{Z_x(\alpha)} \exp\left[-\alpha \sum_i |x_i|^p\right], \tag{5}$$

with $Z_x(\alpha)$ the partition function normalizing the distribution. The partition function $Z_x(\alpha)$ can be calculated using

$$\int_0^\infty \exp[-\alpha u^p] du = \frac{1}{p}\int_0^\infty \exp[-\alpha v] v^{\frac{1-p}{p}} dv \propto \alpha^{-\frac{1}{p}},$$

with $u^p = v$, which results in $Z_x(\alpha) = c\alpha^{-\frac{N}{p}}$, with c a constant. The final form of the sparsity prior is given by

$$\mathrm{p}(\mathbf{x}|\alpha) = c\,\alpha^{\frac{N}{p}} \exp\left[-\alpha \sum_i |x_i|^p\right]. \tag{6}$$

Note that this probability distribution uses a single hyperparameter α for all signal coefficients, whereas existing Bayesian methods generally employ independent distributions on each signal coefficient [15, 16, 17], where each distribution is modeled using a separate hyperparameter. However, as will be shown in Sec. 3, we introduce an additional variable which will separately enforce adaptivity for each coefficient. Note also that a *maximum a posteriori* (MAP) formulation with the distributions in (4) and (5) results in the same inverse problem shown in (2), using $\tau = \frac{\alpha}{\beta}$.

2.2 Model for hyperparameters

In order to simplify the inference procedure, in Bayesian models, hyperprior distributions are generally chosen to be conjugate distributions, i.e., they have the same form as the product of the conditional distribution and the priors. Therefore, we utilize conjugate Gamma hyperpriors on both hyperparameters α and β. In addition to being conjugate, the Gamma distribution includes the uniform distribution as a limiting case, in which case the hyperparameters are estimated only depending on the observations. The distributions $\mathrm{p}(\alpha)$ and $\mathrm{p}(\beta)$ are then expressed as

$$\mathrm{p}(\alpha) = \Gamma(\alpha|a_\alpha^0, b_\alpha^0) = \frac{(b_\alpha^0)^{a_\alpha^0}}{\Gamma(a_\alpha^0)} \alpha^{a_\alpha^0 - 1} \exp\left[-\alpha b_\alpha^0\right], \tag{7}$$

$$\mathrm{p}(\beta) = \Gamma(\beta|a_\beta^0, b_\beta^0) = \frac{(b_\beta^0)^{a_\beta^0}}{\Gamma(a_\beta^0)} \beta^{a_\beta^0 - 1} \exp\left[-\beta b_\beta^0\right], \tag{8}$$

with a_α^0, a_β^0 the shape parameters and b_α^0, b_β^0 the scale parameters, respectively. In this work, these parameters are assigned small values (e.g., 10^{-3}) to obtain vague hyperpriors which make the hyperparameter estimates to rely more on the observations than on prior knowledge. The means and variances of α and β are given respectively by

$$\mathrm{Mean}[\alpha] = <\alpha> = \frac{a_\alpha}{b_\alpha},\ \mathrm{Var}[\alpha] = \frac{a_\alpha}{(b_\alpha)^2}. \tag{9}$$

$$\mathrm{Mean}[\beta] = <\beta> = \frac{a_\beta}{b_\beta},\ \mathrm{Var}[\beta] = \frac{a_\beta}{(b_\beta)^2}. \tag{10}$$

Combining the distributions at both stages of the hierarchical model defined in (4), (5) and (7)-(8), we obtain the joint distribution in (3).

3. INFERENCE PROCEDURE

The Bayesian inference is based on the posterior distribution

$$\mathrm{p}(\alpha,\beta,\mathbf{x}|\mathbf{y}) = \frac{\mathrm{p}(\alpha,\beta,\mathbf{x},\mathbf{y})}{\mathrm{p}(\mathbf{y})}. \tag{11}$$

However, approximation methods are needed because $\mathrm{p}(\mathbf{y})$ cannot be computed. In this work, we incorporate a variational Bayesian approach for the inference (as in [19, 20, 21]), which approximates the posterior distribution $\mathrm{p}(\alpha,\beta,\mathbf{x}|\mathbf{y})$ by an analytically tractable distribution $\mathrm{q}(\alpha,\beta,\mathbf{x})$, found by minimizing the Kullback-Leibler (KL) divergence between the posterior distribution and its approximation, given by

$$\begin{aligned} & C_{KL}(\mathrm{q}(\alpha,\beta,\mathbf{x}) \parallel \mathrm{p}(\alpha,\beta,\mathbf{x}|\mathbf{y})) \\ &= \int\int\int \mathrm{q}(\alpha,\beta,\mathbf{x}) \log\left(\frac{\mathrm{q}(\alpha,\beta,\mathbf{x})}{\mathrm{p}(\alpha,\beta,\mathbf{x}|\mathbf{y})}\right) d\alpha d\beta d\mathbf{x} \\ &= \int\int\int \mathrm{q}(\alpha,\beta,\mathbf{x}) \log\left(\frac{\mathrm{q}(\alpha,\beta,\mathbf{x})}{\mathrm{p}(\alpha,\beta,\mathbf{x},\mathbf{y})}\right) d\alpha d\beta d\mathbf{x} + \mathrm{const.} \end{aligned} \tag{12}$$

The KL divergence is always non negative and equal to zero only when $\mathrm{q}(\alpha,\beta,\mathbf{x}) = \mathrm{p}(\alpha,\beta,\mathbf{x}|\mathbf{y})$. It is generally assumed that the distribution $\mathrm{q}(\alpha,\beta,\mathbf{x})$ can be factorized. In this work we use the following factorization

$$\mathrm{q}(\alpha,\beta,\mathbf{x}) = \mathrm{q}(\alpha)\,\mathrm{q}(\beta)\,\mathrm{q}(\mathbf{x}) \tag{13}$$

Unfortunately, the KL divergence in (12) cannot be calculated with the sparsity prior in (6). Therefore, we resort to a majorization-minimization approach, where we search a bound of the prior which can be used for further Bayesian analysis. Let us consider the weighted arithmetic and geometric mean inequality given by

$$a^{\frac{p}{2}} b^{1-\frac{p}{2}} \leq \frac{p}{2}a + (1-\frac{p}{2})b, \tag{14}$$

with $0 \leq p \leq 2$, and nonnegative numbers a and b. Assuming $b > 0$, $p > 0$ and dividing both sides by $b^{1-\frac{p}{2}}$ we obtain

$$a^{\frac{p}{2}} \leq \frac{p}{2} \frac{a + \frac{2-p}{p}b}{b^{1-p/2}}. \tag{15}$$

Let us also define the following functional

$$\mathbf{M}(\alpha,\mathbf{x},\mathbf{v}) = c\,\alpha^{\frac{N}{p}} \exp\left[-\frac{\alpha p}{2}\sum_i \left(\frac{(x_i)^2 + \frac{2-p}{p} v_i}{(v_i)^{1-p/2}}\right)\right], \tag{16}$$

where $\mathbf{v} \in (R^+)^N$ is a vector with components v_i, and c is the constant in (6). Using (15) with $a = (x_i)^2$ and $b = v_i$ in $\mathbf{M}(\alpha,\mathbf{x},\mathbf{v})$, and comparing it with the prior in (6), it is clear that

$$\mathrm{p}(\mathbf{x}|\alpha) \quad \geq \quad \mathbf{M}(\alpha,\mathbf{x},\mathbf{v}). \tag{17}$$

Since the bounding functional $\mathbf{M}(\alpha,\mathbf{x},\mathbf{v})$ has a quadratic form, Bayesian inference can analytically be carried out by majorizing the prior $\mathrm{p}(\mathbf{x}|\alpha)$ by this functional. Using (17), a lower bound of the joint probability distribution in (3) can be found, that is,

$$\begin{aligned} \mathrm{p}(\alpha,\beta,\mathbf{x},\mathbf{y}) &\geq \mathrm{p}(\alpha)\,\mathrm{p}(\beta)\,\mathbf{M}(\alpha,\mathbf{x},\mathbf{v})\,\mathrm{p}(\mathbf{y}|\mathbf{x},\beta) \\ &= \mathbf{F}(\alpha,\beta,\mathbf{x},\mathbf{v}), \end{aligned} \tag{18}$$

which leads to the following upper bound of the KL divergence in (12)

$$\begin{aligned} & C_{KL}(\mathrm{q}(\alpha,\beta,\mathbf{x}) \parallel \mathrm{p}(\alpha,\beta,\mathbf{x}|\mathbf{y})) \\ & \leq C_{KL}(\mathrm{q}(\alpha,\beta,\mathbf{x}) \parallel \mathbf{F}(\alpha,\beta,\mathbf{x},\mathbf{v})) + \mathrm{const.} \end{aligned} \tag{19}$$

Note that, since

$$\begin{aligned} & C_{KL}(\mathrm{q}(\alpha,\beta,\mathbf{x}) \parallel \mathrm{p}(\alpha,\beta,\mathbf{x}|\mathbf{y})) \\ & \leq \min_{\mathbf{v}} C_{KL}(\mathrm{q}(\alpha,\beta,\mathbf{x}) \parallel \mathbf{F}(\alpha,\beta,\mathbf{x},\mathbf{v})) + \mathrm{const}, \end{aligned} \tag{20}$$

the upper bound $C_{KL}(\mathrm{q}(\alpha,\beta,\mathbf{x}) \parallel \mathbf{F}(\alpha,\beta,\mathbf{x},\mathbf{v}))$ can be made tighter by minimizing it iteratively with respect to both $\mathrm{q}(\alpha,\beta,\mathbf{x})$ and $\mathbf{v}$, which results in a decreasing sequence of upper bounds, and also in closer approximations of the signal prior $\mathrm{p}(\mathbf{x}|\alpha)$ by the bounding functional $\mathbf{M}(\alpha,\mathbf{x},\mathbf{v})$.

Based on this, we replace the minimization of the KL divergence in (12) by its upper bound given in (19). Note, however, that (19) cannot be analytically minimized with respect to all $\mathrm{q}(\cdot)$ distributions and the vector $\mathbf{v}$ at the same time, and so an alternating minimization procedure has to be employed as follows. Let us denote by $\Theta = \{\mathbf{x},\alpha,\beta\}$ the set of all unknowns, and by Θ_θ the set Θ with θ removed. Then, for each unknown $\theta \in \Theta$, the posterior $\mathrm{q}(\theta)$ can be computed by holding $\mathrm{q}(\Theta_\theta)$ constant and solving

$$\mathrm{q}(\theta) = \underset{\mathrm{q}(\theta)}{\mathrm{argmin}}\, C_{KL}(\mathrm{q}(\Theta_\theta)\mathrm{q}(\theta) \parallel \mathbf{F}(\alpha,\beta,\mathbf{x},\mathbf{v})). \tag{21}$$

The standard solution of variational Bayesian analysis [21, 19] can then be used for (21), which results in

$$\mathrm{q}(\theta) = \mathrm{const} \times \exp\left(\mathbf{E}_{\mathrm{q}(\Theta_\theta)}\left[\,\log \mathbf{F}(\alpha,\beta,\mathbf{x},\mathbf{v})\right]\right), \tag{22}$$

where

$$\mathbf{E}_{\mathrm{q}(\Theta_\theta)}\left[\,\log \mathbf{F}(\alpha,\beta,\mathbf{x},\mathbf{v})\right] = \int \log \mathbf{F}(\alpha,\beta,\mathbf{x},\mathbf{v})\mathrm{q}(\Theta_\theta)d\Theta_\theta.$$

Applying this general solution to each unknown in an alternating fashion results in an iterative procedure, which converges to the best approximation of the true posterior distribution $\mathrm{p}(\mathbf{x},\alpha,\beta|\mathbf{y})$ by distributions of the form $\mathrm{q}(\alpha,\beta,\mathbf{x}) = \mathrm{q}(\alpha)\,\mathrm{q}(\beta)\,\mathrm{q}(\mathbf{x})$. Convergence is guaranteed since the upper bound of the KL divergence in (20) is convex.

We next proceed to give the explicit forms of each $\mathrm{q}(\cdot)$ distribution. In what follows, the means of the distributions will be denoted by $< \cdot > = \mathbf{E}_{\mathrm{q}(\theta)}[\cdot]$, when the corresponding distribution is clear from the context. The distribution $\mathrm{q}(\mathbf{x})$ is calculated from (22) as a N-dimensional multivariate Gaussian distribution $\mathcal{N}(\mathbf{x}|{<}\mathbf{x}{>},\Sigma_{\mathbf{x}})$, whose mean and covariance are given by

$$<\mathbf{x}> = \Sigma_{\mathbf{x}} <\beta> \mathbf{\Phi}^t\,\mathbf{y}, \tag{23}$$

$$\Sigma_{\mathbf{x}} = \left(<\beta> \mathbf{\Phi}^t\mathbf{\Phi} + p<\alpha> \mathbf{W}\right)^{-1} \tag{24}$$

with

$$\mathbf{W} = \mathrm{diag}\left(v_i^{p/2-1}\right), i = 1,...,N. \tag{25}$$

The components v_i of the vector $\mathbf{v}$ can be calculated using

$$v_i = \underset{v_i}{\mathrm{argmin}}\, \frac{<x_i^2> + \frac{2-p}{p} v_i}{v_i^{1-p/2}},$$

which results in the following update

$$v_i = <x_i^2>,\ i = 1,...,N. \tag{26}$$

It is clear that $\mathbf{W}$ in (25) is a weighting matrix, similar to that used in IRLS algorithm [12], which together with x_i^2 provides an estimate of $\parallel \mathbf{x} \parallel_p^p$. However, in [12] the elements of $\mathbf{W}$ are chosen as $(<x_i>^2)^{p/2-1}$, whereas in this work they are equal to $(<x_i^2>)^{p/2-1}$, which is calculated from

$$\begin{aligned} <x_i^2> &= (\mathbf{E}_{q(\mathbf{x})}[x_i])^2 + \mathbf{E}_{q(\mathbf{x})}[\left(x_i - \mathbf{E}_{q(\mathbf{x})}[x_i]\right)^2] \\ &= <x_i>^2 + (\Sigma_{\mathbf{x}})_{ii}, \end{aligned} \tag{27}$$

where $(\Sigma_{\mathbf{x}})_{ii}$ denotes the i^{th} diagonal element of the matrix $\Sigma_{\mathbf{x}}$, and it is the variance of the coefficient x_i. The first term is equivalent to the one used in IRLS algorithms, and the second term incorporates the uncertainty of the estimate $\mathbf{x}$ in the reweighting procedure. Using this information results in significant improvement in the reconstruction performance compared to the IRLS methods. Additionally, the estimated variances can be utilized for designing adaptive measurement systems as in [16].

Finally, from (22), the distributions $\mathrm{q}(\alpha)$ and $\mathrm{q}(\beta)$ are found as Gamma distributions given by

$$q(\alpha) \propto \alpha^{N/p+a_\alpha^0-1} \exp\left[-\alpha\left(\sum_i v_i^{p/2} + b_\alpha^0\right)\right], \tag{28}$$

$$q(\beta) \propto \beta^{N/2+a_\beta^0-1} \exp\left[-\beta\left(\frac{\mathbf{E}_{q(\mathbf{x})}\left(\|\mathbf{y} - \mathbf{\Phi}\mathbf{x}\|_2^2\right)}{2} + b_\beta^0\right)\right]. \tag{29}$$

The means of these distributions are given by

$$<\alpha> = \mathbf{E}_{q(\alpha)}[\alpha] = \frac{N/p + a_\alpha^0}{\sum_i v_i^{p/2} + b_\beta^0}, \tag{30}$$

and

$$<\beta> = \mathbf{E}_{q(\beta)}[\beta] = \frac{N/2 + a_\beta^0}{\mathbf{E}_{q(\mathbf{x})}\left(\|\mathbf{y} - \mathbf{\Phi}\mathbf{x}\|_2^2\right)/2 + b_\beta^0}, \tag{31}$$

The denominator in (31) is calculated using

$$\mathbf{E}_{q(\mathbf{x})}[\|\mathbf{y}-\mathbf{\Phi}\mathbf{x}\|_2^2] = \|\mathbf{y}-\mathbf{\Phi}<\mathbf{x}>\|_2^2 + \text{trace}(\Sigma_{\mathbf{x}}\mathbf{\Phi}^t\mathbf{\Phi}). \tag{32}$$

In summary, the algorithm iterates between (23), (25), (30) and (31) until convergence. The estimate $<\mathbf{x}>$ in (23) can be calculated by standard methods, such as Gaussian elimination. However, explicit calculation of the matrix $\Sigma_{\mathbf{x}}$ is needed in (27) and (32). This is computationally very intense, since $\Sigma_{\mathbf{x}}$ is of size $N \times N$. To increase efficiency and decrease numerical errors, we first calculate the incomplete Cholesky factorization $\Sigma_{\mathbf{x}}^{-1} \approx \mathbf{L}\mathbf{L}^T$ and approximate $\Sigma_{\mathbf{x}}$ by $\left(\mathbf{L}\mathbf{L}^T\right)^{-1}$.

To conclude this section, we investigate the special case of noiseless CS measurements ($\mathbf{y} = \mathbf{\Phi}\mathbf{x}$). From (23) and (24), we see that when $\beta \longrightarrow \infty$, the estimate of $\mathbf{x}$ is given by

$$<\mathbf{x}> = \mathbf{W}^{-1}\mathbf{\Phi}^t\left(\mathbf{\Phi}\mathbf{W}^{-1}\mathbf{\Phi}^t\right)^{-1}\mathbf{y}. \tag{33}$$

Let us further assume that the distribution $\mathrm{q}(\mathbf{x})$ is a degenerate distribution, that is, a distribution which takes the value $<\mathbf{x}>$ with probability one and the rest with probability zero. Then, $<x_i^2> = <x_i>^2$, and therefore

$$\mathbf{W} = \operatorname{diag}\left(|<x_i>|^{p-2}\right). \tag{34}$$

The estimate in (33) combined with (34) coincides with the IRLS algorithm [11], thus [11] is a special case of the proposed method. Moreover, by including an appropriate regularization in (34), [12] can also be shown to be a special case of the proposed method.

4. EXPERIMENTS

In this section we present numerical comparison of our method with some of the state-of-the-art algorithms for CS recovery. We generate sparse vectors $\mathbf{x}$ of size $N = 256$ with 20 nonzero coefficients, which are drawn from a zero-mean Gaussian distribution of variance 1. The $M \times N$ measurement matrices $\mathbf{\Phi}$ are also generated from a zero-mean Gaussian distribution with variance 1, and their columns are scaled to have unit 2-norms. Other choices of both the signal and measurement matrix gave similar results.

In the proposed algorithm, the LS-solution $\left(\mathbf{\Phi}^t\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^t\mathbf{y}$ is used as the initial estimate of $\mathbf{x}$, and the iterations are stopped when the Euclidean divergence from the estimate to the previous iteration is less than 10^{-6}. Finally, the reconstruction error is calculated as $\|\hat{\mathbf{x}}-\mathbf{x}\|_2/\|\mathbf{x}\|_2$, where $\hat{\mathbf{x}}$ and $\mathbf{x}$ are the estimated and true coefficient vectors, respectively. For all methods and experiments, the number of samples M varies from 60 to 120 in steps of 10, and results were averaged over 100 executions of each method. We study both noiseless and noisy observations, using zero-mean white gaussian noise of standard deviation 0.03.

We first study the effect of the variable p on the reconstruction performance of our algorithm, denoted by BCS-lp in the following. Figure 1 shows error rate comparisons between six selected p-values for both noiseless and noisy observations. It is clear that smaller values of p result in lower reconstruction errors for both cases. Note also that the performance increase is logarithmic when decreasing p, so values close to $p = 0.01$ results in similar performance.

Fig. 2 compares the proposed algorithm, using both noiseless and noisy measurements, with respect to a selection of existing CS reconstruction algorithms, namely, BCS [16] and BCS-Laplace [22] (both greedy algorithms); BP [2] and GPSR [3] (global optimization methods minimizing l_1-norm), the IRLS method [12] (minimizing a nonconvex norm and using an optimal regularization strategy as described in [12]) and iterative hard thresholding (IHT) [5] (minimizing l_0 norm). For all algorithms, their MATLAB implementations in the corresponding websites are used, and the required algorithm parameters are set according to their default setups. We chose $p = 0$ for IRLS as it provided the best comparative results.

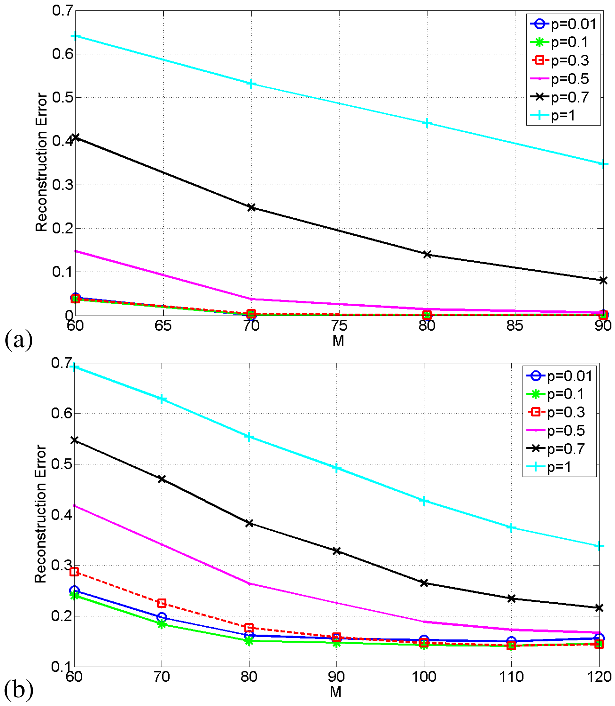

Figure 1: Reconstruction errors obtained by the proposed method with varying the number of measurements M for p-values 0.01, 0.1, 0.3, 0.05, 0.7 and 1. The measurements are noiseless in (a) and Gaussian noise with standard deviation 0.03 is added to the measurements.

It can be observed from Fig. 2 that (a) in the case of noiseless measurements, BCS-lp outperforms other methods in terms of reconstruction error and it achieves perfect reconstruction with fewer number of measurements; and (b) when noise is present, BCS-lp provides the smallest reconstruction error for small number of measurements, but other algorithms are better for higher values. We believe that this is mostly due to numerical errors arising when solving the linear system in (23). Note that a similar behavior can also be observed with IRLS. The performance of GPSR is expected to increase with manually tuning its parameters. An additional advantage of our method is that it does not require parameter-tuning. An interesting observation from Fig. 2(a) is that the reconstruction performance is improved as more heavy-tailed distributions are utilized as sparsity priors.

5. CONCLUSIONS

We have developed a Bayesian framework utilizing nonconvex sparsity priors for compressed sensing reconstruction, through a majorization-minimization approach. By

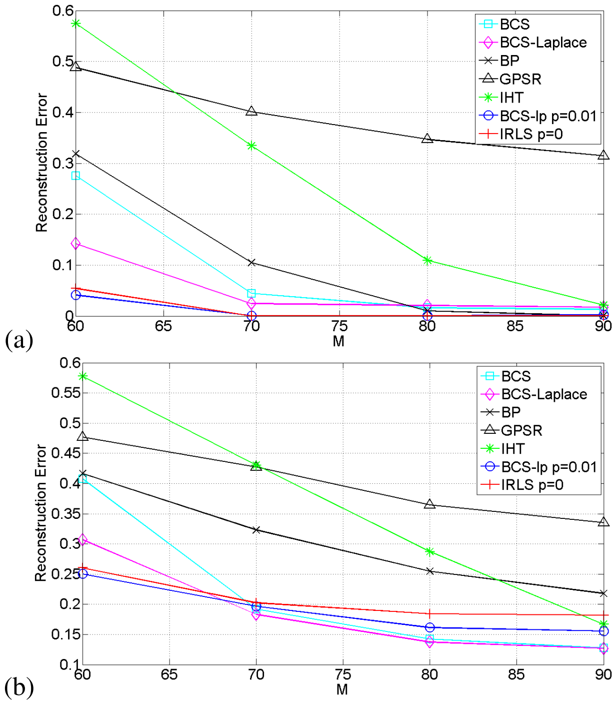

Figure 2: Comparison between a number of CS reconstruction algorithms with varying number of measurements M with (a) noiseless and (b) noisy measurements.

using variational Bayesian analysis, the reconstruction algorithm developed from this framework simultaneously estimates all unknowns and provides distribution estimates, which account for the estimation uncertainties and can be used to ensure the accuracy of the estimation process. We have shown that the proposed formulation is a generalized version of some existing methods, such as reweighted least squares and sparse Bayesian methods, and therefore it can provide potential directions for improvement. Experimental results demonstrate that using non-convex priors our method achieves higher reconstruction accuracy, and also that the unknown signal can be recovered with fewer measurements. We have shown that our method is competitive compared to state-of-the-art methods in terms of reconstruction error.

REFERENCES

[1] E. Candes, J. Romberg, and T. Tao, “Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information,” *IEEE Trans. Inf. Theory*, **52** (2), pp. 489-509, Feb. 2006.

[2] S. S. Chen, D. L. Donoho, and M. A. Saunders, “Atomic decomposition by basis pursuit,” *SIAM J. on Sci. Comp.*, **20** (1), pp. 33-61, 1999.

[3] M. Figueiredo, R. Nowak, and S. J. Wright, “Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems,” *IEEE Trans. on Selected Topics in Signal Processing*, **1** (4), pp. 586-597, December 2007.

[4] J. Trzasko and A. Manduca, “Highly undersampled magnetic resonance image reconstruction via homotopic L0-minimization”, *IEEE Trans. Med. Imaging.*, **28** (1), pp.106–121, Jan. 2009.

[5] T. Blumensath and M. E. Davies, “A simple, efficient and near optimal algorithm for compressed sensing,” in *IEEE Int. Conf. on Acoustics, Speech, and Sig. Proc. (ICASSP09)*, Taipei, Taiwan, 2009.

[6] J. A. Tropp and A. C. Gilbert, “Signal recovery from random measurements via orthogonal matching pursuit,” *IEEE Trans. Inf. Theory*, **53** (12), pp. 4655-4666, Dec. 2007.

[7] T. Blumensath and M. E. Davies, “Gradient pursuits,” *IEEE Trans. Sig. Proc.*, **56** (6), pp. 2370-2382, Jun. 2008.

[8] R. Chartrand, “Exact reconstruction of sparse signals via nonconvex minimization,” *IEEE Sig. Proc. Letters*, **14** (10), pp. 707-710, Oct. 2007.

[9] C. L. Lawson, “Contributions to the theory of linear least maximum approximations,” Ph.D. dissertation, UCLA, 1961.

[10] A. E. Beaton and J. W. Tukey, “The fitting of power series, meaning polynomials, illustrated on bandspectroscopic data,‘” *Technometrics*, **16**, pp. 145-185, 1974.

[11] B. Rao and K. Kreutz-Delgado, “An affine scaling methodology for best basis selection,” *IEEE Trans. on Sig. Proc.*, **47** (1), pp. 187-200, Jan 1999.

[12] R. Chartrand and W. Yin, “Iteratively reweighted algorithms for compressive sensing,” *IEEE Int. Conf. on Acoustics, Speech and Sig. Proc. (ICASSP 2008)*, 31 March - April 4 2008.

[13] I. Daubechies, R. DeVore, M. Fornasier, and S. Gunturk, “Iteratively re-weighted least squares minimization for sparse recovery,” to appear in *Commun. Pure Appl. Math.*, 2009.

[14] E. J. Candes, M. B. Wakin, and S. P. Boyd, “Enhancing sparsity by reweighted l1 minimization,” *J. of Fourier Anal. and App. (Special issue on sparsity)*, **14** (5), pp. 877-905, December 2008.

[15] D. Wipf and B. Rao, “Sparse Bayesian learning for basis selection,” *IEEE Trans. Sig. Proc.*, **52** (8), pp. 2153-2164, Aug. 2004.

[16] S. Ji, Y. Xue, and L. Carin, “Bayesian compressive sensing,” *IEEE Trans. Sig. Proc.*, **56** (6), pp. 2346-2356, Jun. 2008.

[17] S. Ji, D. Dunson, and L. Carin, “Multi-task compressive sensing,” to appear in *IEEE Trans. Sig. Proc.*, 2008.

[18] L. He and L. Carin, “Exploiting structure in wavelet-based Bayesian compressive sensing,” submitted to *IEEE Trans. on Sig. Proc.*, 2009.

[19] C. M. Bishop, “Pattern Recognition and Machine Learning,”. Springer- Verlag, 2006.

[20] S. D. Babacan, R. Molina, and A. Katsaggelos, “Parameter estimation in TV image restoration using variational distribution approximation,” *IEEE Trans. on Im. Proc.*, **17** (3), pp. 326-339, Mar. 2008.

[21] J. Miskin, “Ensemble learning for independent component analysis,” Ph.D. dissertation, University of Cambridge, 2000.

[22] S. Babacan, R. Molina, and A. Katsaggelos, “Fast Bayesian compressive sensing using Laplace priors,” in *IEEE Int. Conf. on Acoustics, Speech, and Sig. Proc. (ICASSP09)*, Taipei, Taiwan, 2009.