Bayesian Compressive Sensing Using Laplace Priors

S. Derin Babacan, *Student Member, IEEE*, Rafael Molina, *Member, IEEE*, and Aggelos K. Katsaggelos, *Fellow, IEEE*

***Abstract*—In this paper, we model the components of the compressive sensing (CS) problem, i.e., the signal acquisition process, the unknown signal coefficients and the model parameters for the signal and noise using the Bayesian framework. We utilize a hierarchical form of the Laplace prior to model the sparsity of the unknown signal. We describe the relationship among a number of sparsity priors proposed in the literature, and show the advantages of the proposed model including its high degree of sparsity. Moreover, we show that some of the existing models are special cases of the proposed model. Using our model, we develop a constructive (greedy) algorithm designed for fast reconstruction useful in practical settings. Unlike most existing CS reconstruction methods, the proposed algorithm is fully automated, i.e., the unknown signal coefficients and all necessary parameters are estimated solely from the observation, and, therefore, no user-intervention is needed. Additionally, the proposed algorithm provides estimates of the uncertainty of the reconstructions. We provide experimental results with synthetic 1-D signals and images, and compare with the state-of-the-art CS reconstruction algorithms demonstrating the superior performance of the proposed approach.**

***Index Terms*—Bayesian methods, compressive sensing, inverse problems, relevance vector machine (RVM), sparse Bayesian learning.**

I. Introduction

COMPRESSIVE sensing (or sampling) (CS) has become a very active research area in recent years due to its interesting theoretical nature and its practical utility in a wide range of applications. Let $\mathbf{f}$ represent the $N\times 1$ unknown signal, which is compressible in a linear basis Ψ (such as a wavelet basis). In other words, $\mathbf{f}=\Psi\mathbf{w}$, where $\mathbf{w}$ is an $N\times 1$ sparse signal, i.e., most of its coefficients are zero. Consider the following acquisition system:

$$\mathbf{y}=\Phi'\mathbf{f}+\mathbf{n} \tag{1}$$

where $M\times 1$ linear measurements $\mathbf{y}$ of the original unknown signal $\mathbf{f}$ are taken with an $M\times N$ measurement matrix $\Phi'=[\phi'_1,\phi'_2,\ldots,\phi'_N]$ and $\mathbf{n}$ represents the acquisition noise. We can also write (1) in terms of the sparse transform coefficients as

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

where $\Phi=\Phi'\Psi$, which is the commonly used notation in the CS literature and will be adopted in the rest of this paper.

According to the theory of compressive sensing, when the number of measurements is small compared to the number of signal coefficients ($M\ll N$), under certain conditions the original signal $\mathbf{f}$ can be reconstructed very accurately by utilizing appropriate reconstruction algorithms [1], [2]. Compressive sensing can be seen as the combination of the conventional acquisition and compression processes: Traditionally, the signal $\mathbf{f}$ is acquired in a lossless manner followed by compression where only the most important features are kept, such as the largest wavelet coefficients. In [1] and [2], it has been shown that since the signal is compressible, it is possible to merge the acquisition and compression processes by performing a reduced number of measurements and recovering the most important features by utilizing an incoherent sampling mechanism, i.e., the sensing basis Φ' and the representation basis Ψ have low coherence. Recent theoretical results show that random sampling matrices exhibit such low coherence with the representation bases. Deterministic designs have also been proposed with slightly reduced performance [3], [4].

There are many applications of compressive sensing, including medical imaging [5] where reducing the number of measurements results in reduced image acquisition time, imaging processes where the cost of taking measurements is high, and sensor networks, where the number of sensors may be limited [6].

Since the number of measurements M is much smaller than the number of unknown coefficients $\mathbf{w}$, the original signal cannot be obtained directly from the measurements. The inversion of (1) or (2) is required, which is an ill-posed problem. Therefore, compressive sensing incorporates a reconstruction mechanism to obtain the original signal. By exploiting the sparsity of $\mathbf{w}$, the inverse problem is regularized constraining the l_0 norm of $\mathbf{w}$, $||\mathbf{w}||_0$, which is equal to the number of nonzero terms in $\mathbf{w}$. An approximation to the original signal is then obtained by solving the following optimization problem:

$$\hat{\mathbf{w}}=\arg\min_{\mathbf{w}}\left\{||\mathbf{y}-\Phi\mathbf{w}||_2^2+\tau||\mathbf{w}||_0\right\}. \tag{3}$$

This optimization problem is NP-hard; therefore, some simplifications are used. The most common one is to use the l_1-norm instead of the l_0-norm, so that the optimization problem becomes

$$\hat{\mathbf{w}}=\arg\min_{\mathbf{w}}\left\{||\mathbf{y}-\Phi\mathbf{w}||_2^2+\tau||\mathbf{w}||_1\right\} \tag{4}$$

where $||\cdot||_1$ denotes the l_1-norm.

Manuscript received September 21, 2008; revised August 24, 2009. First published September 22, 2009; current version published December 16, 2009. This work was supported in part by the "Comisión Nacional de Ciencia y Tecnología" under contract TIC2007-65533 and in part by the Spanish research programme Consolider Ingenio 2010: MIPRCV (CSD2007-00018). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Magdy Bayoumi.

S. D. Babacan and A. K. Katsaggelos are with the Department of Electrical Engineering and Computer Science, Northwestern University, IL 60208-3118 USA (e-mail: sdb@northwestern.edu; aggk@eecs.northwestern.edu).

R. Molina is with the Departamento de Ciencias de la Computación e I.A. Universidad de Granada, Spain (e-mail: rms@decsai.ugr.es).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2009.2032894

A number of methods have been proposed to solve the CS reconstruction problems defined in (3) and (4) or their extensions (for example, formulations utilizing l_p norms for $\mathbf{w}$ with $0 < p \leq 1$). Most of the proposed methods are examples of energy minimization methods, including linear programming algorithms [7], [8] and constructive (greedy) algorithms [9]–[11]. Additionally, sparse signal representation is a very close topic to CS, and many algorithms proposed there can also be applied to the CS reconstruction problem (see [12] and references therein).

The compressive sensing formulation in (3) and (4) can be considered as the application of a deterministic regularization approach to signal reconstruction. However, the problem can also be formulated in a Bayesian framework, which provides certain distinct advantages over other formulations. These include providing probabilistic predictions, automatic incorporation and estimation of model parameters, and estimation of the uncertainty of reconstruction. The latter advantage also facilitates the estimation of the quality of the measurements which can be used to design adaptive measurements [13], [14]. The Bayesian framework was utilized for the compressive sensing problem in [13] and [14]. In [13], the *relevance vector machine* (RVM) proposed in [15] is adapted to the CS problem. Independent Laplace priors are utilized for each coefficient in an expectation-propagation framework in [14], and both signal reconstruction and measurement design problems are considered. However, the resulting algorithm is complicated to implement, and all required parameters are not estimated, but rather left as parameters to be tuned.

In this paper, we also formulate the CS reconstruction problem from a Bayesian perspective. We utilize a Bayesian model for the CS problem and propose the use of Laplace priors on the basis coefficients in a hierarchical manner. As will be shown, our formulation includes the RVM formulation [15] as a special case, but results in smaller reconstruction errors while imposing sparsity to a higher extent. Moreover, we provide a Bayesian inference procedure which results in an efficient greedy constructive algorithm. Our formulation naturally incorporates the advantages of the Bayesian framework, such as providing posterior distributions rather than point estimates, and, therefore, providing an estimate of the uncertainty in the reconstructions, which, for instance, can be used as a feedback mechanism for adapting the data acquisition process. Furthermore, the resulting algorithm is fully automated since all required model parameters are estimated along with the unknown signal coefficients $\mathbf{w}$. This is in contrast to most of the existing methods in the literature which include a number of parameters to be tuned specifically to the data, which is a cumbersome process. We will demonstrate with experimental results that despite being fully automated, the proposed algorithm provides competitive and even higher reconstruction performance than state-of-the-art methods.

The rest of this paper is organized as follows. In Section II, we present the hierarchical Bayesian modeling of the CS problem, the observation model and the prior model on the signal coefficients. In this section, we review existing prior models for sparse learning and show that some of them are special cases of the model used in this paper. In Section III, we apply the evidence procedure to the CS problem and propose an efficient reconstruction algorithm. We present experimental results in Section IV and conclusions are drawn in Section V.

II. Bayesian Modeling

In Bayesian modeling, all unknowns are treated as stochastic quantities with assigned probability distributions. The unknown signal $\mathbf{w}$ is assigned a *prior* distribution $\mathrm{p}(\mathbf{w}|\boldsymbol{\gamma})$, which models our knowledge on the nature of $\mathbf{w}$. The observation $\mathbf{y}$ is also a random process with *conditional* distribution $\mathrm{p}(\mathbf{y}|\mathbf{w},\beta)$, where $\beta = 1/\sigma^2$ is the inverse noise variance. These distributions depend on the model parameters $\boldsymbol{\gamma}$ and β, which are called *hyperparameters*, and additional prior distributions, called *hyperpriors*, are assigned to them.

The Bayesian modeling of the CS reconstruction problem requires the definition of a joint distribution $\mathrm{p}(\mathbf{w},\boldsymbol{\gamma},\beta,\mathbf{y})$ of all unknown and observed quantities. In this paper, we use the following factorization:

$$\mathrm{p}(\mathbf{w},\boldsymbol{\gamma},\beta,\mathbf{y}) = \mathrm{p}(\mathbf{y}|\mathbf{w},\beta)\mathrm{p}(\mathbf{w}|\boldsymbol{\gamma})\mathrm{p}(\boldsymbol{\gamma})\mathrm{p}(\beta). \tag{5}$$

A. Observation (Noise) Model

The observation noise is independent and Gaussian with zero mean and variance equal to β^{-1}, that is, with (2)

$$\mathrm{p}(\mathbf{y}|\mathbf{w},\beta) = \mathcal{N}(\mathbf{y}|\Phi\mathbf{w},\beta^{-1}) \tag{6}$$

with a Gamma prior placed on β as follows:

$$\mathrm{p}(\beta|a^\beta,b^\beta) = \Gamma(\beta|a^\beta,b^\beta). \tag{7}$$

The Gamma distribution is defined as

$$\Gamma(\xi|a^\xi,b^\xi) = \frac{(b^\xi)^{a^\xi}}{\Gamma(a^\xi)}\xi^{a^\xi-1}\exp[-b^\xi\xi] \tag{8}$$

where $\xi > 0$ denotes a hyperparameter, $b^\xi > 0$ is the scale parameter, and $a^\xi > 0$ is the shape parameter. The mean and variance of ξ are given respectively by

$$\mathrm{Mean}[\xi] = \langle\xi\rangle = \frac{a^\xi}{b^\xi}, \quad \mathrm{Var}[\xi] = \frac{a^\xi}{(b^\xi)^2}. \tag{9}$$

The Gamma distribution is generally chosen as the prior for the inverse variance (precision) of a Gaussian distribution because it is its conjugate prior, which greatly simplifies the analysis and also includes the uniform distribution as a limiting case.

B. Signal Model

The l_1 regularization formulation in (4) is equivalent to using a Laplace prior on the coefficients $\mathbf{w}$, that is

$$\mathrm{p}(\mathbf{w}|\lambda) = \frac{\lambda}{2}\exp\left(-\frac{\lambda}{2}\|\mathbf{w}\|_1\right) \tag{10}$$

and using a *maximum a posteriori* (MAP) formulation with (6) and (10) for $\tau = \lambda/\beta$. However, this formulation of the Laplace prior does not allow for a tractable Bayesian analysis, since it is not conjugate to the conditional distribution in (6). To alleviate this, hierarchical priors are employed. In the following, we review the models utilized so far in the literature to model $\mathbf{w}$ and introduce the prior structure utilized in this work.

In [16], as the first stage of a hierarchical model, the following prior is employed on $\mathbf{w}$:

$$\mathrm{p}(\mathbf{w}|\boldsymbol{\gamma}) = \prod_{i=1}^{N} \mathcal{N}(w_i|0,\gamma_i) \tag{11}$$

where $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_N)$. In the second stage of the hierarchy, a Jeffrey's hyperprior is utilized independently on each γ_i, that is

$$\mathrm{p}(\gamma_i) \propto \frac{1}{\gamma_i}. \tag{12}$$

Observe that since

$$\mathrm{p}(\gamma_i) = \lim_{\zeta \to 0} \Gamma(\gamma_i|\zeta, 0) \tag{13}$$

we can obtain a sample from the prior distribution of each w_i independently by first obtaining a sample γ_i from a $\Gamma(\zeta, 0)$ distribution when $\zeta \to 0$ and then sampling a $\mathcal{N}(0, \gamma_i)$.

Alternatively, in [15], the prior model on $\mathbf{w}$ is formulated conditioned on the precision variables $\alpha_i = \gamma_i^{-1}$. A Gamma hyperprior is utilized on the precision variables, that is

$$\mathrm{p}\left(\alpha_i|a_i^{\alpha}, b_i^{\alpha}\right) = \Gamma\left(\alpha_i|a_i^{\alpha}, b_i^{\alpha}\right). \tag{14}$$

This formulation with the hierarchical prior in (11) and (14) is commonly referred to as the *relevance vector machine* (RVM), or *sparse Bayesian learning* (SBL) [12], [15]. Note, however, that both in the original work [15] and its adaptation to the compressive sensing problem [13], the shape and scale parameters are set respectively equal to $a_i^{\alpha} = 1$, $b_i^{\alpha} = 0$, thus obtaining *uniform* or *noninformative* distributions for these parameters.

When using noninformative priors on α_i, $\mathrm{p}(\alpha_i|a_i^{\alpha}, b_i^{\alpha})$ becomes

$$\mathrm{p}(\alpha_i|1,0) = \lim_{\zeta \to 1} \Gamma(\alpha_i|\zeta, 0). \tag{15}$$

It is important to mention that when changing variables from γ_i to α_i their corresponding maximum *a posteriori* estimations are not related by $(\alpha_i)_{\mathrm{MAP}} = 1/(\gamma_i)_{\mathrm{MAP}}$.

Values of a_i^{α} and b_i^{α} other than $a_i^{\alpha} = 1$, $b_i^{\alpha} = 0$ will result in Student's t distributions for the marginal distribution $\mathrm{p}(\mathbf{w})$. It is argued in [17] that Student's t priors will lead to less sparse solutions than RVM.

As explained in [14], compared to the separate Gaussian priors employed on the entries of $\mathbf{w}$ in the RVM framework, Laplace priors enforce the sparsity constraint more heavily by distributing the posterior mass more on the axes so that signal coefficients close to zero are preferred. Furthermore, the Laplace prior is also the prior that promotes sparsity to the largest extent while being log-concave [14]. The log-concavity provides the very useful advantage of eliminating local-minima since it leads to unimodal posterior distributions [14], [18], [19].

Based on the above, in this paper, we propose to use Laplace priors on the signal coefficients $\mathbf{w}$. In order to overcome the fact that the Laplace distribution is not conjugate to the observation model in (6), we model it in a hierarchical way by using the following hyperpriors on γ_i [16]

$$\mathrm{p}(\gamma_i|\lambda) = \Gamma(\gamma_i|1, \lambda/2) = \frac{\lambda}{2} \exp\left(-\frac{\lambda\gamma_i}{2}\right), \quad \gamma_i \geq 0,\ \lambda \geq 0 \tag{16}$$

and then using the Gaussian model in (11) to model $\mathrm{p}(\mathbf{w}|\boldsymbol{\gamma})$. In other words, we have

$$\begin{aligned} \mathrm{p}(\mathbf{w}|\lambda) &= \int \mathrm{p}(\mathbf{w}|\boldsymbol{\gamma})\mathrm{p}(\boldsymbol{\gamma}|\lambda) d\boldsymbol{\gamma} = \prod_i \int \mathrm{p}(w_i|\gamma_i)\mathrm{p}(\gamma_i|\lambda) d\gamma_i \\ &= \frac{\lambda^{N/2}}{2^N} \exp\left(-\sqrt{\lambda} \sum_i |w_i|\right). \end{aligned} \tag{17}$$

Finally, we model λ as the realization of the following Gamma hyperprior:

$$\mathrm{p}(\lambda|\nu) = \Gamma(\lambda|\nu/2, \nu/2). \tag{18}$$

The proposed modeling constitutes a three-stage hierarchical form. The first two stages of this hierarchical prior (11) and (16) result in a Laplace distribution $\mathrm{p}(\mathbf{w}|\lambda)$ [16], and the last stage (18) is embedded to calculate λ. This formulation can be shown to be very closely related to the convex variational formulation in [20], and the total-variation priors used in image restoration [21].

The prior distribution on λ is flexible enough so as to provide a range of restrictions on λ; from very vague information on λ

$$\mathrm{p}(\lambda) \propto \frac{1}{\lambda} \tag{19}$$

which would be obtained when $\nu \to 0$, to very precise information

$$\mathrm{p}(\lambda) = \begin{cases} 1, & \text{if } \lambda = 1 \\ 0, & \text{elsewhere} \end{cases} \tag{20}$$

which is obtained when $\nu \to \infty$. Note that by using a $\Gamma(a\nu/2, \nu/2)$, we have more flexibility regarding the hyperprior of λ but at the cost of having to estimate an additional parameter.

Observe that we can obtain a sample from the prior distribution of $\mathbf{w}$ by first sampling a $\Gamma(\nu/2, \nu/2)$ distribution to obtain λ, then sample a $\Gamma(1, \lambda/2)$ distribution N times to obtain γ_i, $i = 1, \ldots, N$ and finally sample $\mathcal{N}(0, \gamma_i)$ to obtain w_i.

We can now see a clear difference between how a realization from each of the prior distributions is obtained. While the $\{\gamma_i\}$ in (12) and $\{\alpha_i\}$ in (14) are obtained as realizations of independent distributions, the $\{\gamma_i\}$ values in (16) all come from a common distribution (see Fig. 1). The advantage of using the model in (12) and the model in (14) with $a_i^{\alpha} = 1, b_i^{\alpha} = 0$ is that there is no need to estimate the parameter λ [16]. However, as we will see in the next section, the inference based on the model in (12) is a particular case of the one based on the model in (16) which leads to the Laplace prior model. We will also show with experimental results that the performance of our model is superior to the alternative prior models of the signal coefficients.

By combining the stages of the hierarchical Bayesian model, the joint distribution can finally be defined as $\mathrm{p}(\mathbf{w}, \boldsymbol{\gamma}, \lambda, \beta, \mathbf{y}) = \mathrm{p}(\mathbf{y}|\mathbf{w}, \beta)\mathrm{p}(\beta)\mathrm{p}(\mathbf{w}|\boldsymbol{\gamma})\mathrm{p}(\boldsymbol{\gamma}|\lambda)\mathrm{p}(\lambda)$, where

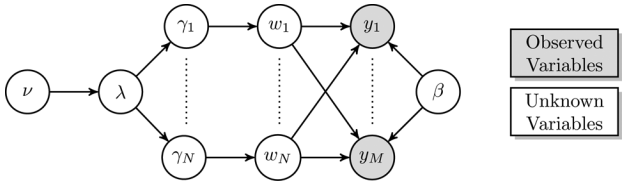

Fig. 1. Directed acyclic graph representing the Bayesian model.

$\mathrm{p}(\mathbf{y}|\mathbf{w},\beta)$, $\mathrm{p}(\beta)$, $\mathrm{p}(\mathbf{w}|\boldsymbol{\gamma})$, $\mathrm{p}(\boldsymbol{\gamma}|\lambda)$ and $\mathrm{p}(\lambda)$ are defined in (6), (7), (11), (16), and (18) respectively. The dependencies in this joint probability model are shown in graphical form in Fig. 1, where the arrows are used to denote the generative model. Note that the hierarchical structure can also be seen from the first four blocks from the left, which correspond to the variables ν, λ, $\gamma_1,\ldots\gamma_N$, and $w_1,\ldots,w_N$.

III. Bayesian Inference

As widely known, Bayesian inference is based on the posterior distribution

$$\mathrm{p}(\mathbf{w},\boldsymbol{\gamma},\lambda,\beta|\mathbf{y}) = \frac{\mathrm{p}(\mathbf{w},\boldsymbol{\gamma},\lambda,\beta,\mathbf{y})}{\mathrm{p}(\mathbf{y})}. \tag{21}$$

However, the posterior $\mathrm{p}(\mathbf{w},\boldsymbol{\gamma},\lambda,\beta|\mathbf{y})$ is intractable, since

$$\mathrm{p}(\mathbf{y}) = \int\int\int\int \mathrm{p}(\mathbf{w},\boldsymbol{\gamma},\lambda,\beta,\mathbf{y}) d\mathbf{w} d\boldsymbol{\gamma} d\lambda d\beta \tag{22}$$

cannot be calculated analytically. Therefore, approximation methods are utilized. In this paper, we utilize the evidence procedure (type-II maximum likelihood approach) to perform Bayesian inference.

A. Evidence Procedure

We will now derive the Bayesian inference using an evidence procedure with the conditional distribution in (6) and the priors in (11), (16), and (18). Our inference procedure is based on the following decomposition:

$$\mathrm{p}(\mathbf{w},\boldsymbol{\gamma},\lambda,\beta|\mathbf{y}) = \mathrm{p}(\mathbf{w}|\mathbf{y},\boldsymbol{\gamma},\beta,\lambda)\mathrm{p}(\boldsymbol{\gamma},\beta,\lambda|\mathbf{y}) \tag{23}$$

where the dependency on ν is dropped for clarity. Since $\mathrm{p}(\mathbf{w}|\mathbf{y},\boldsymbol{\gamma},\beta,\lambda) \propto \mathrm{p}(\mathbf{w},\mathbf{y},\boldsymbol{\gamma},\beta,\lambda)$, the distribution $\mathrm{p}(\mathbf{w}|\mathbf{y},\boldsymbol{\gamma},\beta,\lambda)$ is found to be a multivariate Gaussian distribution $\mathcal{N}(\mathbf{w}|\mu,\Sigma)$ with parameters

$$\mu = \Sigma\beta\Phi^T\mathbf{y} \tag{24}$$

$$\Sigma = [\beta\Phi^T\Phi + \Lambda]^{-1} \tag{25}$$

with

$$\Lambda = \mathrm{diag}(1/\gamma_i). \tag{26}$$

We now utilize $\mathrm{p}(\boldsymbol{\gamma},\beta,\lambda|\mathbf{y})$ in (23) to estimate the hyperparameters. In the type-II maximum likelihood procedure, we represent $\mathrm{p}(\boldsymbol{\gamma},\beta,\lambda|\mathbf{y})$ by a degenerate distribution where the distribution is replaced by a delta function at its mode, where we assume that this posterior distribution is sharply peaked around its mode [22]. Then, using $\mathrm{p}(\boldsymbol{\gamma},\beta,\lambda|\mathbf{y}) = (\mathrm{p}(\mathbf{y},\boldsymbol{\gamma},\beta,\lambda)/\mathrm{p}(\mathbf{y})) \propto \mathrm{p}(\mathbf{y},\boldsymbol{\gamma},\beta,\lambda)$, we estimate the hyperparameters by the maxima of the joint distribution $\mathrm{p}(\mathbf{y},\boldsymbol{\gamma},\beta,\lambda)$ which is obtained from $\mathrm{p}(\mathbf{y},\mathbf{w},\boldsymbol{\gamma},\beta,\lambda)$ by integrating out $\mathbf{w}$. Consequently, we have

$$\begin{aligned}\mathrm{p}(\mathbf{y},\boldsymbol{\gamma},\beta,\lambda) &= \int \mathrm{p}(\mathbf{y}|\mathbf{w},\beta)\mathrm{p}(\mathbf{w}|\boldsymbol{\gamma})\mathrm{p}(\boldsymbol{\gamma}|\lambda)\mathrm{p}(\lambda)\mathrm{p}(\beta)d\mathbf{w}\\ &= \left(\frac{1}{2\pi}\right)^{N/2}|\beta^{-1}\mathbf{I}+\Phi\Lambda^{-1}\Phi^T|^{-1/2}\\ &\quad\times\exp\left[-\frac{1}{2}\mathbf{y}^T(\beta^{-1}\mathbf{I}+\Phi\Lambda^{-1}\Phi^T)^{-1}\mathbf{y}\right]\\ &\quad\times\mathrm{p}(\boldsymbol{\gamma}|\lambda)\mathrm{p}(\lambda)\mathrm{p}(\beta)\\ &= \left(\frac{1}{2\pi}\right)^{N/2}|\mathbf{C}|^{-1/2}\exp\left[-\frac{1}{2}\mathbf{y}^T\mathbf{C}^{-1}\mathbf{y}\right]\\ &\quad\times\mathrm{p}(\boldsymbol{\gamma}|\lambda)\mathrm{p}(\lambda)\mathrm{p}(\beta)\end{aligned} \tag{27}$$

where $\mathbf{C} = (\beta^{-1}\mathbf{I} + \Phi\Lambda^{-1}\Phi^T)$ and $\mathbf{I}$ is the $M \times M$ identity matrix.

Instead of maximizing this distribution, we maximize equivalently its logarithm, which results in the following functional to be maximized:

$$\begin{aligned}\mathcal{L} = &-\frac{1}{2}\log|\mathbf{C}| - \frac{1}{2}\mathbf{y}^T\mathbf{C}^{-1}\mathbf{y} + N\log\frac{\lambda}{2} - \frac{\lambda}{2}\sum_i\gamma_i\\ &+\frac{\nu}{2}\log\frac{\nu}{2} - \log\Gamma(\nu/2) + \left(\frac{\nu}{2}-1\right)\log\lambda - \frac{\nu}{2}\lambda\\ &+(a^\beta-1)\log\beta - b^\beta\beta.\end{aligned} \tag{28}$$

Let us now state some equivalences that will be useful in solving this maximization problem. First, we obtain

$$|\mathbf{C}| = |\Lambda|^{-1}|\beta^{-1}\mathbf{I}||\Lambda+\beta\Phi^T\Phi| = |\Lambda|^{-1}|\beta^{-1}\mathbf{I}||\Sigma^{-1}| \tag{29}$$

using the determinant identity [23] and thus

$$\log|\mathbf{C}| = -\log|\Lambda| - N\log\beta - \log|\Sigma|. \tag{30}$$

Furthermore, using the Woodbury identity [24], we have

$$\begin{aligned}\mathbf{C}^{-1} &= (\beta^{-1}\mathbf{I}+\Phi\Lambda^{-1}\Phi^T)^{-1}\\ &= \beta\mathbf{I} - \beta\Phi(\Lambda+\beta\Phi^T\Phi)^{-1}\Phi^T\beta\\ &= \beta\mathbf{I} - \beta\Phi\Sigma\Phi^T\beta\end{aligned} \tag{31}$$

and, therefore

$$\begin{aligned}\mathbf{y}^T\mathbf{C}^{-1}\mathbf{y} &= \beta\mathbf{y}^T\mathbf{y} - \beta\mathbf{y}^T\Phi\Sigma\Phi^T\beta\mathbf{y}\\ &= \beta\mathbf{y}^T(\mathbf{y}-\Phi\mu)\\ &= \beta\|\mathbf{y}-\Phi\mu\|^2 + \beta\mu^T\Phi^T(\mathbf{y}-\Phi\mu)\\ &= \beta\|\mathbf{y}-\Phi\mu\|^2 + \mu^T\Lambda\mu.\end{aligned} \tag{32}$$

Using these identities, the derivative of $\mathcal{L}$ with respect to γ_i is given by

$$\frac{d\mathcal{L}}{d\gamma_i} = \frac{1}{2}\left[-\frac{1}{\gamma_i} + \frac{\langle w_i^2\rangle}{\gamma_i^2} - \lambda\right] \tag{33}$$

where $\langle w_i^2 \rangle = \mu_i^2 + \Sigma_{ii}$ with Σ_{ii} the i^{th} diagonal element of Σ. Setting this equal to zero results in

$$\gamma_i = -\frac{1}{2\lambda} + \sqrt{\frac{1}{4\lambda^2} + \frac{\langle \omega_i^2 \rangle}{\lambda}}. \tag{34}$$

The updates of the other hyperparameters are found similarly by taking the derivative of (28) with respect to each hyperparameter and setting it equal to zero. The updates found in this manner are given by

$$\lambda = \frac{N - 1 + \nu/2}{\sum_i \gamma_i/2 + \nu/2} \tag{35}$$

$$\beta = \frac{N/2 + a^\beta}{\langle ||\mathbf{y} - \Phi \mathbf{w}||^2 \rangle /2 + b^\beta} \tag{36}$$

where the expected value is calculated with respect to the conditional distribution of $\mathbf{w}$.

Finally, we can also estimate ν by maximizing (28) with respect to ν. This results in solving the following equation:

$$\log \frac{\nu}{2} + 1 - \psi\left(\frac{\nu}{2}\right) + \log \lambda - \lambda = 0. \tag{37}$$

This equation does not have a closed-form solution so it is solved numerically.

In summary, at each iteration of the algorithm, given estimates of $\boldsymbol{\gamma}$, β, and λ, the estimate of the distribution of $\mathbf{w}$ is calculated using (24) and (25), followed by the estimation of the variances γ_i from (34), the hyperparameter λ from (35), the noise inverse variance (precision) β from (36) and ν from (37), where the expected values needed in these equations are calculated using the current distribution of $\mathbf{w}$.

Note that the same update equations can be obtained by applying an expectation-maximization (EM) procedure instead of the direct maximization method employed in this section. Fixed point iterations [25] can also be applied to find $\boldsymbol{\gamma}$, β, and λ. Note also that a similar optimization procedure is used in [26] for a different modeling of the signal.

B. Fast Suboptimal Solutions

There is a major disadvantage of the method presented in the previous section; namely, it requires the solution of a linear system of N equations in (24), which requires $\mathbf{O}(N^3)$ computations. Moreover, since the system in (2) is underdetermined with $M \ll N$, numerical errors create major practical difficulties in solving this system. Although the matrix Σ can be written using the Woodbury matrix identity as follows:

$$\begin{aligned}\Sigma &= \Lambda^{-1} - \Lambda^{-1}\Phi^T(\beta^{-1}\mathbf{I} + \Phi\Lambda^{-1}\Phi^T)^{-1}\Phi\Lambda^{-1} \\ &= \Lambda^{-1} - \Lambda^{-1}\Phi^T\mathbf{C}^{-1}\Phi\Lambda^{-1}\end{aligned} \tag{38}$$

which requires the solution of only M linear equations, therefore, $\mathbf{O}(M^3)$ time, this is in practice more problematic due to numerical errors and it still does not scale up well for large-scale problems. Therefore, the algorithm presented in the previous section cannot be easily applied to practical problems, but it will serve us as the starting point in developing a practical algorithm as follows.

To promote sparsity and to decrease the computational requirements, only a single γ_i will be updated at each iteration of the algorithm instead of updating the whole vector $\boldsymbol{\gamma}$. As will be shown later, updating a single hyperparameter leads to very efficient updates of the matrix Σ and the mean μ. A fundamental observation is that if a single hyperparameter γ_i is set equal to zero, μ_i must be equal to 0, and so the corresponding entry is pruned out from the model. Since it is assumed that the vector $\mathbf{w}$ is sparse, many of its components are zero; therefore, most γ_i's are set equal to zero, and matrix Σ can be represented using fewer dimensions than $N \times N$. Exploiting these properties, one can obtain a much more efficient procedure than the algorithm presented in the previous section, by starting with an "empty" model ($\boldsymbol{\gamma} = 0$) and iteratively adding components to the model. In the following, we will present such a procedure.

A fundamental observation to obtain the fast suboptimal solution is that the matrix $\mathbf{C}$ in (28) can be written as follows:

$$\begin{aligned}\mathbf{C} &= \beta^{-1}\mathbf{I} + \sum_i \gamma_i \phi_i \phi_i^T \\ &= \beta^{-1}\mathbf{I} + \sum_{j \neq i} \gamma_j \phi_j \phi_j^T + \gamma_i \phi_i \phi_i^T \\ &= \mathbf{C}_{-i} + \gamma_i \phi_i \phi_i^T\end{aligned} \tag{39}$$

where $\mathbf{C}_{-i}$ denotes that the contribution of the i^{th} basis is not included. Using the Woodbury identity in (39), we obtain

$$\mathbf{C}^{-1} = \mathbf{C}_{-i}^{-1} - \frac{\mathbf{C}_{-i}^{-1}\phi_i\phi_i^T\mathbf{C}_{-i}^{-1}}{1/\gamma_i + \phi_i^T\mathbf{C}_{-i}^{-1}\phi_i} \tag{40}$$

and using the determinant identity, we obtain

$$|\mathbf{C}| = |\mathbf{C}_{-i}|\left|1 + \gamma_i \phi_i^T \mathbf{C}_{-i}^{-1}\phi_i\right|. \tag{41}$$

Substituting the last two equations in (28) and treating $\mathcal{L}$ as a function of $\boldsymbol{\gamma}$ only, we obtain

$$\begin{aligned}\mathcal{L}(\boldsymbol{\gamma}) =& -\frac{1}{2}\left[\log|\mathbf{C}_{-i}| + \mathbf{y}^T\mathbf{C}_{-i}^{-1}\mathbf{y} + \frac{\lambda}{2}\sum_{j\neq i}\gamma_j\right] \\ &+ \frac{1}{2}\left[\log\frac{1}{1+\gamma_i s_i} + \frac{q_i^2\gamma_i}{1+\gamma_i s_i} - \lambda\gamma_i\right] \qquad (42) \\ =& \mathcal{L}(\boldsymbol{\gamma}_{-i}) + l(\gamma_i) \qquad (43)\end{aligned}$$

where $l(\gamma_i) = (1/2)[\log(1/(1+\gamma_i s_i)) + (q_i^2\gamma_i/(1+\gamma_i s_i)) - \lambda\gamma_i]$ and q_i and s_i are defined as

$$s_i = \phi_i^T\mathbf{C}_{-i}^{-1}\phi_i \tag{44}$$

$$q_i = \phi_i^T\mathbf{C}_{-i}^{-1}\mathbf{y}. \tag{45}$$

Note that the quantities q_i and s_i do not depend on γ_i since $\mathbf{C}_{-i}^{-1}$ is independent of γ_i. Therefore, the terms related to a single hyperparameter γ_i are now separated from others. Let us now examine if the i^{th} basis should be included. A closed form solution of the maximum of $\mathcal{L}(\boldsymbol{\gamma})$, when only its i^{th} component is changed, can be found by holding other hyperparameters fixed, taking its derivative with respect to γ_i and setting it equal to zero. The derivative of $\mathcal{L}(\boldsymbol{\gamma})$ with respect to γ_i can be expressed as

$$\frac{d\mathcal{L}(\boldsymbol{\gamma})}{d\gamma_i} = \frac{dl(\gamma_i)}{d\gamma_i} = \frac{1}{2}\left[-\frac{s_i}{1+\gamma_i s_i} + \frac{q_i^2}{(1+\gamma_i s_i)^2} - \lambda\right]$$

$$= -\frac{1}{2}\left[\frac{\gamma_i^2\left(\lambda s_i^2\right)+\gamma_i\left(s_i^2+2\lambda s_i\right)+\left(\lambda+s_i-q_i^2\right)}{(1+\gamma_i s_i)^2}\right]. \tag{46}$$

Note that the numerator has a quadratic form while the denominator is always positive, and, therefore, $dl(\gamma_i)/d\gamma_i = 0$ is satisfied at

$$\gamma_i = \frac{-s_i(s_i+2\lambda) \pm s_i\sqrt{(s_i+2\lambda)^2 - 4\lambda\left(s_i - q_i^2+\lambda\right)}}{2\lambda s_i^2} \tag{47}$$

$$= \frac{-s_i(s_i+2\lambda) \pm s_i\sqrt{\Delta}}{2\lambda s_i^2} \tag{48}$$

where $\Delta = (s_i+2\lambda)^2 - 4\lambda(s_i - q_i^2+\lambda) > 0$. Observe that if $q_i^2 - s_i < \lambda$, then $\Delta^2 < s_i + 2\lambda$ and both solutions in (48) are negative, and since $dl(\gamma_i)/d\gamma_i|_{\gamma_i=0} < 0$, the maximum occurs at $\gamma_i = 0$. On the other hand if $q_i^2 - s_i > \lambda$, there are two real solutions, one negative and the variance estimate

$$\gamma_i = \frac{-s_i(s_i+2\lambda) + s_i\sqrt{(s_i+2\lambda)^2 - 4\lambda\left(s_i - q_i^2+\lambda\right)}}{2\lambda s_i^2}. \tag{49}$$

Since when $q_i^2 - s_i > \lambda$ we have $dl(\gamma_i)/d\gamma_i|_{\gamma_i=0} > 0$ and $dl(\gamma_i)/d\gamma_i|_{\gamma_i=\infty} < 0$, the obtained variance estimate in (49) maximizes $l(\gamma_i)$ and, therefore, $\mathcal{L}(\boldsymbol{\gamma})$.

In summary, the maximum of $\mathcal{L}(\boldsymbol{\gamma})$, when all components of $\boldsymbol{\gamma}$ except γ_i are kept fixed, is achieved at

$$\gamma_i = \begin{cases} \frac{-s_i(s_i+2\lambda)+s_i\sqrt{(s_i+2\lambda)^2-4\lambda\left(s_i-q_i^2+\lambda\right)}}{2\lambda s_i^2}, & \text{if } q_i^2 - s_i > \lambda \\ 0, & \text{otherwise.} \end{cases} \tag{50}$$

Note that in the case of $\gamma_i = 0$, the corresponding basis ϕ_i is pruned out from the model and μ_i is set equal to zero. Therefore, (50) provides a systematic method of deciding which basis vectors should be included in the model and which should be excluded. Note that as in the previous section, the estimate of λ is provided by (35).

It is crucial for computational efficiency that once a hyperparameter γ_i is updated using (50), the quantities s_i, q_i, μ and Σ are efficiently updated. Similarly to [27], the parameters q_i and s_i can be calculated for all basis vectors ϕ_i efficiently using the following identities:

$$S_i = \beta\phi_i^T\phi_i - \beta^2\phi_i^T\Phi\Sigma\Phi^T\phi_i \tag{51}$$

$$Q_i = \beta\phi_i^T\mathbf{y} - \beta^2\phi_i^T\Phi\Sigma\Phi^T\mathbf{y} \tag{52}$$

$$s_i = \frac{S_i}{1-\gamma_i S_i} \tag{53}$$

$$q_i = \frac{Q_i}{1-\gamma_i S_i} \tag{54}$$

where Σ and Φ include only the columns i that are included in the model ($\gamma_i \neq 0$). Moreover, Σ and μ can be updated very efficiently when only a single coefficient γ_i is considered, as in [27]. Utilizing these equations, we can obtain an iterative procedure by updating one hyperparameter γ_i at each iteration, and updating s_i, q_i, μ and Σ accordingly. The procedure is summarized below in Algorithm 1.

Algorithm 1 FAST LAPLACE

1: INPUTS: Φ, $\mathbf{y}$
2: OUTPUTS: $\mathbf{w}$, Σ, γ
3: Initialize all $\gamma_i = 0$, $\lambda = 0$
4: **while** convergence criterion not met **do**
5: Choose a γ_i (or equivalently choose a basis vector ϕ_i)
6: **if** $q_i^2 - s_i > \lambda$ AND $\gamma_i = 0$ **then**
7: Add γ_i to the model
8: **else if** $q_i^2 - s_i > \lambda$ AND $\gamma_i > 0$ **then**
9: Re-estimate γ_i
10: **else if** $q_i^2 - s_i < \lambda$ **then**
11: Prune i from the model (set $\gamma_i = 0$)
12: **end if**
13: Update Σ and μ
14: Update s_i, q_i
15: Update λ using (35)
16: Update ν using (37)
17: **end while**

At step 5 of the algorithm, a candidate γ_i must be selected for updating. This can be done by randomly choosing a basis vector ϕ_i, or by calculating each γ_i and choosing the one that results in the greatest increase in $\mathcal{L}(\boldsymbol{\gamma})$ in (42), which results in a faster convergence. The latter is the method implemented in this work. Finally, the updates of Σ, μ, s_i, and q_i in the add, delete, and re-estimate operations are the same as those in the RVM formulation (see [27, Appendix A] for details).

An important step in the algorithm is the estimation of the noise precision β, which is done in the previous section using (36). Unfortunately, this method cannot be used in practice in this fast algorithm since the proposed algorithm is constructive and the reconstruction and, therefore, the estimate in (36) are unreliable at early iterations. Due to the underdetermined nature of the compressive sensing problem, once the estimate of β is very far from its true value, the reconstruction quality is also significantly affected. Therefore, we fix the estimate of this parameter in the beginning of the algorithm using $\beta = 0.01\|\mathbf{y}\|_2^2$ inspired by [8] and [13]. Alternatively, this parameter can be integrated out from the model as in [28].

Note that unlike other constructive (or greedy) methods such as OMP [9], StOMP [29], and gradient pursuit methods [11], included basis vectors can also be deleted once they are determined to be irrelevant. This is a powerful feature of the algorithm, since errors in the beginning of the reconstruction process can be fixed in later stages by effectively pruning out irrelevant basis vectors which can drive the algorithm away from the optimal result.

Let us complete this section by comparing the variance estimates provided by the relevance vector machine (where $\lambda = 0$) with the ones provided by the proposed method in terms of sparsity. The estimate γ_i^{RVM} in the RVM framework is given by [27]

$$\gamma_i^{\mathrm{RVM}} = \arg\max_{\gamma_i} \frac{1}{2}\left[\log\frac{1}{1+\gamma_i s_i} + \frac{q_i^2\gamma_i}{1+\gamma_i s_i}\right] \tag{55}$$

while as we have seen the estimate provided by the modeling using the Laplace distribution is given by

$$\begin{aligned}\gamma_i^{\mathrm{L}} &= \arg\max_{\gamma_i} l(\gamma_i) \\ &= \arg\max_{\gamma_i} \frac{1}{2}\left[\log\frac{1}{1+\gamma_i s_i} + \frac{q_i^2\gamma_i}{1+\gamma_i s_i} - \lambda\gamma_i\right]. \end{aligned}\tag{56}$$

Clearly, the RVM model corresponds to the particular case of $\lambda = 0$ in our model. The solution of (55) is given by

$$\gamma_i^{\mathrm{RVM}} = \begin{cases} \frac{q_i^2 - s_i}{s_i^2}, & \text{if } q_i^2 - s_i > 0 \\ 0, & \text{otherwise.} \end{cases} \tag{57}$$

Let us now examine the difference $\gamma_i^{\mathrm{RVM}} - \gamma_i^{\mathrm{L}}$. When $q_i^2 - s_i < \lambda$, we have

$$\gamma_i^{\mathrm{RVM}} - \gamma_i^{\mathrm{L}} = \begin{cases} 0, & \text{if } q_i^2 - s_i < 0 \\ \gamma_i^{\mathrm{RVM}}, & \text{if } 0 \le q_i^2 - s_i < \lambda. \end{cases} \tag{58}$$

When $q_i^2 - s_i \ge \lambda$, the derivative of the function $l(\gamma_i)$ at $\gamma_i = \gamma_i^{\mathrm{RVM}}$ is $-\lambda < 0$. Since $dl(\gamma_i)/d\gamma_i|_{\gamma_i=0} > 0$, the maximum of $l(\gamma_i)$ occurs at a smaller value γ_i^{L} than γ_i^{RVM}. Consequently, we always have

$$\gamma_i^{\mathrm{RVM}} \ge \gamma_i^{\mathrm{L}}. \tag{59}$$

Therefore, the estimates γ_i^{L} using the Laplace prior are always smaller than the estimates γ_i^{RVM} of the relevance vector machine. Note also that compared to RVM more components will possibly be pruned out from the model when $\lambda > 0$, since the cardinality of the set $\{w_i\}$ for which $q_i^2 - s_i > \lambda$ is smaller than of that of the set $\{w_i\}$ for which $q_i^2 - s_i > 0$. These observations imply that the solution obtained by the proposed method is at least as sparse as the one provided by the RVM. This will also be shown empirically in Section IV.

IV. Experiments

In this section, we present experimental results with both 1-D synthetic signals and 2-D images to demonstrate the performance of the proposed method. We considered experimental setups used widely in the literature. In the experiments reported below, we concentrated on the fast algorithm presented in Section III-B due to its wider applicability in practical settings. Although it is suboptimal in theory, it provides better reconstruction results than the algorithm in Section III-A since the computational cost and increased numerical errors render the optimal algorithm impractical. This is especially evident when applying compressive sensing reconstruction algorithms to large-scale problems, such as images. Note that this is also observed when applying RVM to machine learning problems [15], [27] and to CS [13].

The source code developed to obtain the results shown in this section is available online at http://ivpl.eecs.northwestern.edu/.

A. One-Dimensional Synthetic Signals

We use the following default setup in the experimental results reported in this section. Four different types of signals of length N are generated, where T coefficients at random locations of the signals are drawn from four different probability distributions, and the rest $(N - T)$ of the coefficients are set equal to zero. The nonzero coefficients of the sparse signals are realizations of the following four distributions: 1) uniform ± 1 random spikes, 2) zero-mean unit variance Gaussian, 3) unit variance Laplace, and 4) Sstudent's t with 3 degrees of freedom.

As the measurement matrix Φ we chose a uniform spherical ensemble, where the columns ϕ_i are uniformly distributed on the sphere R^N. Other measurement matrices such as partial Fourier and uniform random projection (URP) ensembles gave similar results, and, therefore, they are not reported here.

In the experiments, we fix $N = 512$ and $T = 20$ and vary the number of measurements M from 40 to 120 in steps of 5. Moreover, we present results with noiseless and noisy acquisitions, where for the noisy observations we added zero mean white Gaussian noise with standard deviation 0.03. We repeated each experiment 100 times and report the average of all experiments.

In the first set of experiments, we compare the effect of different choices of the parameter λ on the reconstruction performance. We ran the algorithm presented in Section III-B with $\lambda = 0$, $\lambda = 1$, $\lambda = 10$, and λ estimated using (35). As mentioned in Section II, $\lambda = 0$ corresponds to the RVM formulation [27] which will be denoted by BCS following [13]. Moreover, we show results when the parameter ν is set equal to zero and when it is also estimated automatically using (37).

The reconstruction error is calculated as $\|\hat{\mathbf{w}} - \mathbf{w}\|_2^2/\|\mathbf{w}\|_2^2$, where $\hat{\mathbf{w}}$ and $\mathbf{w}$ are the estimated and true coefficient vectors, respectively. The criterion $\|\mathcal{L}(\boldsymbol{\gamma}^k) - \mathcal{L}(\boldsymbol{\gamma}^{k-1})\|^2 < 10^{-4}$ is used to terminate the iterative procedure.

Average reconstruction errors in 100 runs are shown for the noise-free case in Fig. 2 for all types of signals. It is clear that using nonzero values for λ results in lower reconstruction errors with all types of signals, and the $\lambda = 0$ case (BCS) gives the worst reconstruction error. Even arbitrarily selected nonzero values of λ (see cases with $\lambda = 1$ and $\lambda = 10$) result in better error rates. Automatically estimating λ using (35) results in the best reconstruction performance.

It is interesting to note that estimating the parameter ν automatically using (37) results in slightly worse performance than setting it equal to zero. This suggests that the elements of $\boldsymbol{\gamma}$ can be used to estimate λ in combination with the improper prior $\mathrm{p}(\lambda) \propto 1/\lambda$. In other words, not much more knowledge than the estimated $\boldsymbol{\gamma}$ is needed to estimate λ. Therefore, ν is fixed equal to zero in the remaining experiments.

The results of the same experimental setup with additive observation noise (zero mean white Gaussian noise with standard deviation 0.03) are shown in Fig. 3. Similar performance increases by using nonzero values of λ can be observed, with again estimating λ using (35) resulting in the best performance. Note that although the algorithms result in higher reconstruction errors than in the noise-free case and perfect reconstruction is not attained, good reconstruction performances are still obtained.

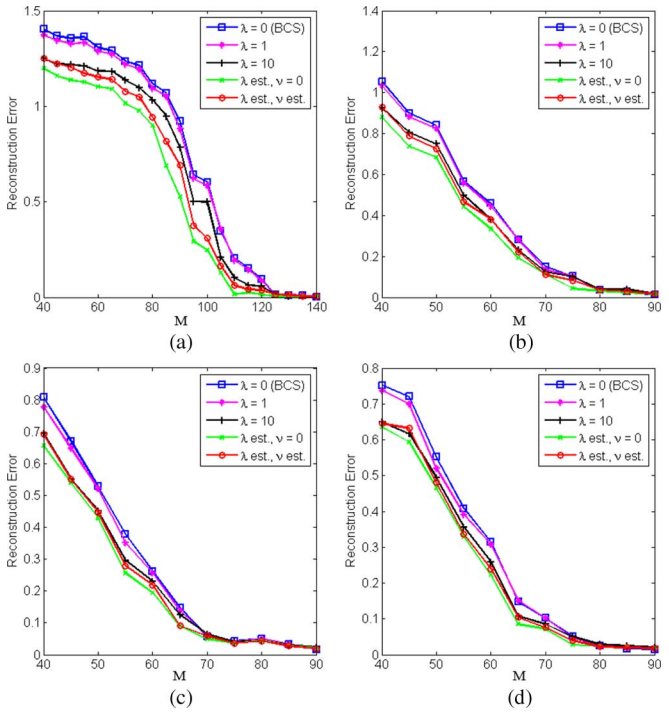

Fig. 2. Number of measurements M versus reconstruction error for the noise-free case resulting from different values of λ. (a) Uniform spikes ± 1; nonuniform spikes drawn from (b) zero mean unit variance Gaussian, (c) unit variance Laplace, (d) student's T with 3 degrees of freedom. In (b), (c), and (d), values corresponding to $M > 90$ are not shown as the error rates are negligible.

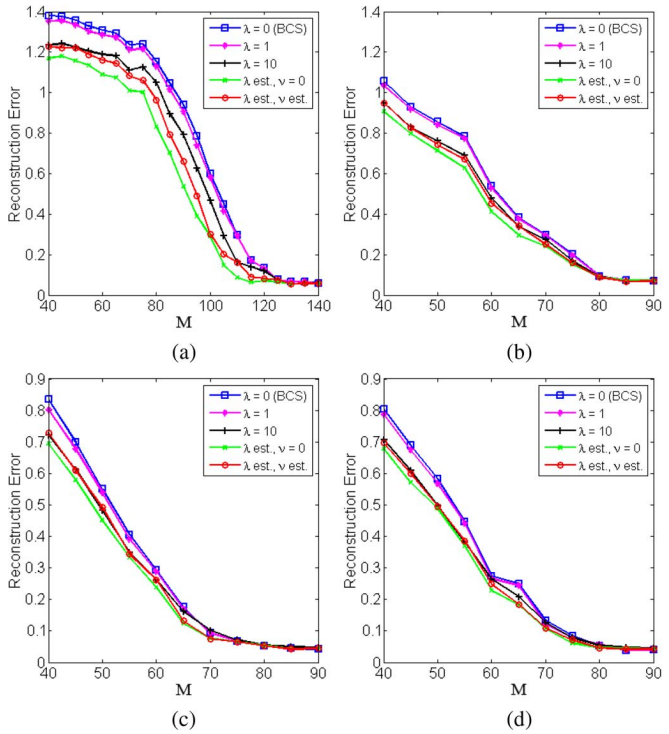

Fig. 3. Number of measurements M versus reconstruction error with noisy observations with different values of λ. (a) Uniform spikes ± 1; nonuniform spikes drawn from (b) zero mean unit variance Gaussian, (c) unit variance Laplace, (d) student's T with 3 degrees of freedom. In (b), (c), and (d), values corresponding to $M > 90$ are not shown as the error rates converged.

In summary, the experimental results suggest that the proposed framework clearly provides improved reconstruction performance over the RVM framework with only a slight difference in computations due to the calculation of (35).

In the second set of experiments, we repeat the same experiment and compare the proposed method with the algorithms BCS [13], BP [7], OMP [9], StOMP with CFAR thresholding (denoted by FAR) [29], and GPSR [8]. For all algorithms, their MATLAB implementations in the corresponding websites are used. The required algorithm parameters are set according to their default setups and in some cases adjusted for improved performance. The algorithms BCS, OMP, and FAR are greedy constructive algorithms like the proposed method, and the algorithms BP and GPSR are global optimization algorithms. We ran the GPSR method both with and without the "debiasing" option, and reported the best result. In the results reported below, Laplace denotes the proposed method where λ is estimated using (35) and the parameter ν is set equal to 0.

Average reconstruction errors of 100 runs are shown for the noise-free case in Fig. 4 for all types of signals. It is clear that the proposed algorithm outperforms all other methods in terms of reconstruction error except for the first signal, for which it provides the best performance after BP and GPSR. However, BP and GPSR result in worse performance than other methods for the rest of the signals. Note that with both algorithms we tuned the algorithm parameters by trial-and-error to achieve their best performance. On the other hand, both BCS and the proposed method do not require parameter tuning. Despite this fact, note that the proposed method provides the best overall performance among all methods.

Examples of reconstructions of the uniform spikes signal are shown in Fig. 5. An important property of the Bayesian methods BCS and Laplace is that they provide the posterior distribution of the unknown signal rather than point estimates. This distribution estimate can be used to provide uncertainty estimates of the coefficients using the covariance matrix Σ, which are shown as error-bars in Fig. 5. These error-bars are variance estimates of the coefficients corresponding to the diagonal elements Σ_{ii} of matrix Σ. Besides being a measure of the uncertainty of the reconstruction, the covariance matrix Σ can also be used to adaptively design the measurement matrix Φ [13], [14].

We repeat the same experiment this time with additive observation noise (zero mean white Gaussian noise with standard deviation 0.03). Average reconstruction errors of 100 runs are shown in Fig. 6. The reconstruction errors obtained by the algorithms are slightly higher than in the noise-free case, and even with a high number of measurements exact reconstructions are not obtained. However, the algorithms still provide accurate reconstructions with a low error rate. Note that the proposed method Laplace again provides the best overall performance for a reasonable number of observations.

For the 1-D experiments reported in this section, the average running times are around 0.1 s for BCS and Laplace, around 0.15 s for BP, and around 0.01 s for the other methods. Therefore, the proposed method and BCS are computationally slightly more demanding than other methods except BP, but such differences are small and they are considered justified considering the improvement in error rates obtained by the proposed method.

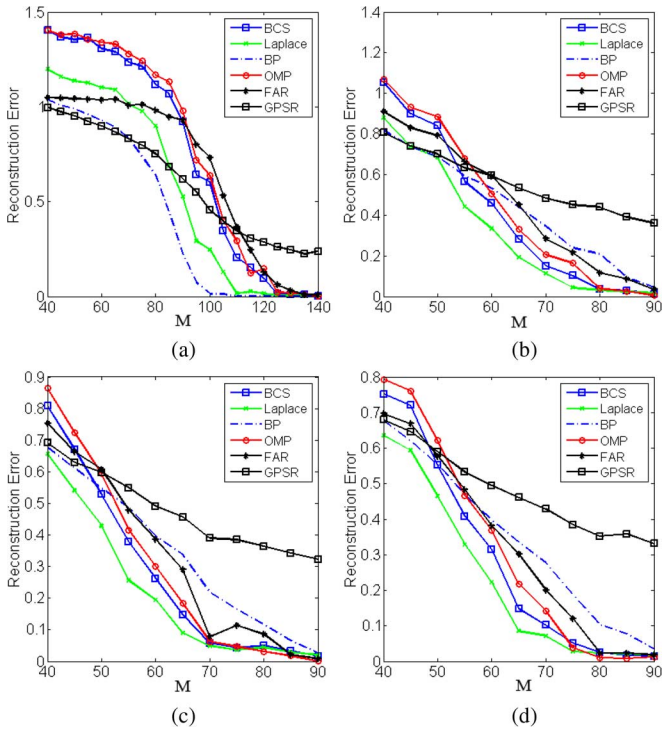

Fig. 4. Number of measurements M versus reconstruction error for the noise-free case for different algorithms. (a) Uniform spikes ± 1; nonuniform spikes drawn from (b) zero mean unit variance Gaussian, (c) unit variance Laplace, (d) student's T with 3 degrees of freedom. In (b), (c), and (d) values corresponding to $M > 90$ are not shown as the error rates converged.

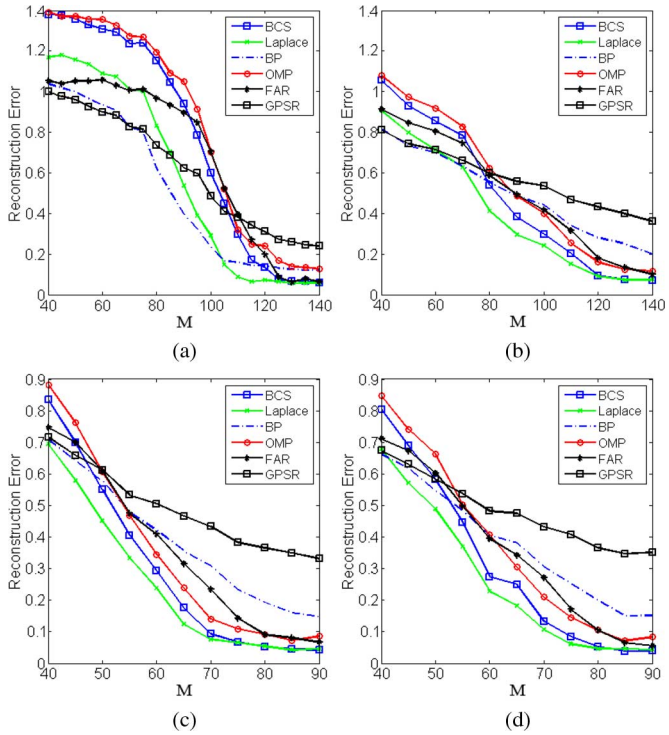

Fig. 6. Number of measurements M versus reconstruction error with noisy observations for different algorithms. (a) Uniform spikes ± 1; nonuniform spikes drawn from (b) zero mean unit variance Gaussian, (c) unit variance Laplace, (d) student's T with 3 degrees of freedom. In (b), (c), and (d) values corresponding to $M > 90$ are not shown for clarity as the error rates converged.

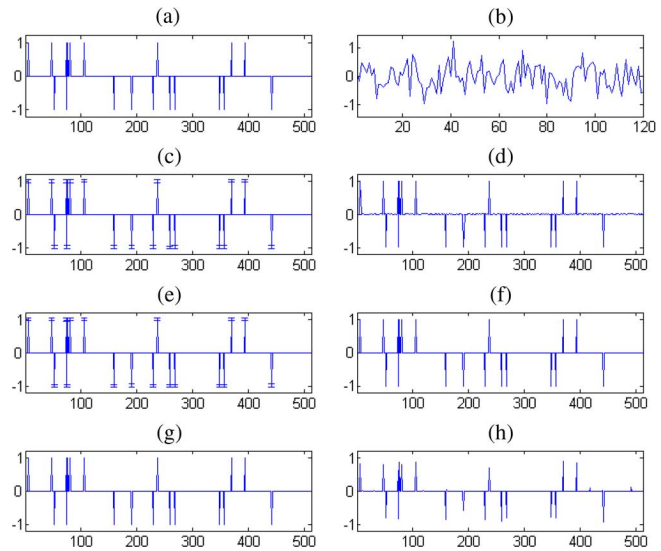

Fig. 5. Reconstruction of uniform spikes signal with $\mathrm{N} = 512, \mathrm{M} = 120$, and $\mathrm{T} = 20$. (a) Original Signal, (b) Observation, Reconstructions with (c) Laplace, (d) BP, (e) BCS, (f) OMP, (g) FAR, and (h) GPSR. All reconstructions have negligible errors except GPSR with reconstruction $\mathrm{error} = 0.2186$. The error bars in (c) and (e) correspond to the estimated variances of the coefficients.

As will be shown in the next section, the proposed method is computationally very competitive when applied to larger scale problems, such as images.

B. Images

In this section, we present a comparison between the proposed method and a number of existing methods on a widely used experimental setup, namely the multiscale CS reconstruction [30] of the 512×512 Mondrian image. We adapted the same test parameters as in the SparseLab package [31]: The multiscale CS scheme is applied on the wavelet transform of the image with a "symmlet8" wavelet with the coarsest scale 4 and finest scale 6. The number of wavelet samples is $N = 4096$, the number of measurements is $M = 2713$, and the measurement matrices are drawn from a uniform spherical distribution. We compared the performance of the algorithms BP, BCS, and StOMP with CFAR and CFDR thresholding with the proposed method. The parameters of the algorithms BP, CFAR, and CFDR are chosen as in the SparseLab package. As in the previous section, the parameters of BCS and the proposed method are solely estimated from the measurements.

The reconstruction error is calculated as $\|\hat{\mathbf{f}}-\mathbf{f}\|_2^2/\|\mathbf{f}\|_2^2$, where $\hat{\mathbf{f}}$ and $\mathbf{f}$ are the estimated and true images, respectively. Since the measurement matrices are random, the experiment is repeated 100 times and their average is reported. Average reconstruction errors, running times and the number of nonzero components in the reconstructed images are shown in Table I, where "Linear" denotes linear reconstruction with $N = 4096$ measurements and represents the best reconstruction performance that can be achieved. It is clear that although BCS and Laplace have nearly the same error rate, Laplace is faster and the reconstructed image is sparser. In fact, Laplace provides the sparsest

TABLE I
AVERAGE RECONSTRUCTION ERRORS, RUNNING TIMES AND NUMBER OF NONZERO COMPONENTS FOR MULTISCALE CS RECONSTRUCTION OF THE *Mondrian* IMAGE

	Mondrian		
	# Nonzeros	Time (s)	Error
Linear	4096	-	0.13325
BP	4096	78.254	0.13933
CFAR	1139.2	13.88	0.14971
CFDR	2177.3	7.86	0.20867
BCS	1174.2	18.343	0.1443
Laplace	1078.7	15.372	0.1451

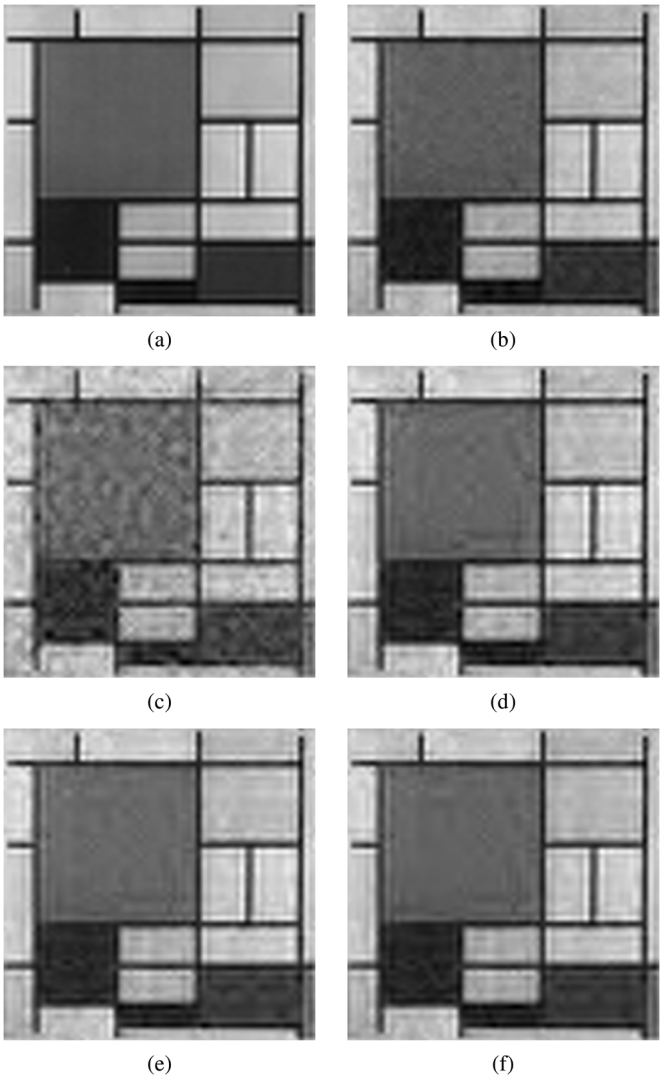

Fig. 7. Examples of reconstructed Mondrian images using a multiscale compressive sensing scheme by (a) linear reconstruction (error $= 0.13325$), (b) BP (error $= 0.13874$, time $= 76.555$ s, no. of nonzero components $= 4096$), (c) StOMP with FDR thresholding (error $= 0.1747$, time $= 6.48$ s, no. of nonzero components $= 2032$), (d) StOMP with FAR thresholding (error $= 0.14673$, time $= 19.759$ s, no. of nonzero components $= 1196$), (e) BCS (error $= 0.14233$, time $= 16.086$ s, no. of nonzero components $= 1145$) and (f) Laplace (error $= 0.14234$, time $= 15.982$, no. of nonzero components $= 1125$).

reconstructed image among all methods. The CFDR method, although it is the fastest, has the worst reconstruction error, and the BP method, although it has the best reconstruction error, has the largest computation time. Laplace and CFAR are clearly the methods that should be preferred, having near-best reconstruction errors and smallest computation times, where CFAR being slightly faster and Laplace having slightly lower reconstruction error. Examples of reconstructed images are shown in Fig. 7 where it can be observed that these methods provide fairly good reconstructions.

In summary, experimental results with both 1-D synthetic signals and 2-D images show that the proposed method provides improved performance in reconstruction quality with competitive performance in computational resources.

V. CONCLUSION

In this paper, we formulated the compressing sensing problem from a Bayesian perspective, and presented a framework to simultaneously model and estimate the sparse signal coefficients. Using this framework, we compared different sparsity priors, and proposed the use of a hierarchical form of Laplace priors on signal coefficients. We have shown that the relevance vector machine is a special case of our formulation, and that our hierarchical prior modeling provides solutions with a higher degree of sparsity and lower reconstruction errors. We proposed a constructive (greedy) algorithm resulting from this formulation, which updates the signal coefficients sequentially in order to achieve low computation times and efficiency in practical problems. The proposed algorithm estimates the unknown signal coefficients simultaneously along with the unknown model parameters. The model parameters are estimated solely from the observation, and, therefore, the proposed algorithm does not require user intervention unlike most existing methods. We demonstrated that overall, the proposed algorithm results in higher performance than most state-of-the-art algorithms. Moreover, the proposed method provides estimates to the distributions of the unknown signal which can be used to measure their uncertainty. The theoretical framework and the proposed algorithm are easy to implement and generalize to investigate further uses of the Bayesian framework in compressive sensing.

REFERENCES

[1] E. J. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.
[2] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.
[3] R. A. DeVore, "Deterministic constructions of compressed sensing matrices," *J. Complexity*, vol. 23, no. 4–6, pp. 918–925, 2007.
[4] S. Gurevich, R. Hadani, and N. Sochen, "On some deterministic dictionaries supporting sparsity," *J. Fourier Anal. Appl.*, vol. 14, pp. 859–876, Dec. 2008.
[5] M. Lustig, D. L. Donoho, and J. M. Pauly, "Sparse MRI: The application of compressed sensing for rapid MR imaging," *Magn. Res. Med.*, vol. 58, no. 6, pp. 1182–1195, Dec. 2007.
[6] J. Haupt, W. U. Bajwa, M. Rabbat, and R. Nowak, "Compressed sensing for networked data [A different approach to decentralized compression]," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 92–101, Mar. 2008.

[7] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61, 1999.
[8] M. Figueiredo, R. Nowak, and S. J. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," *IEEE Trans. Sel. Topics Signal Process.*, vol. 1, no. 4, pp. 586–597, Dec. 2007.
[9] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4655–4666, Dec. 2007.
[10] D. L. Donoho and J. Tanner, "Sparse nonnegative solution of underdetermined linear equations by linear programming," *Proc. Nat. Acad. Sci.*, vol. 102, no. 27, pp. 9446–9451, 2005.
[11] T. Blumensath and M. E. Davies, "Gradient pursuits," *IEEE Trans. Signal Process.*, vol. 56, no. 6, pp. 2370–2382, Jun. 2008.
[12] D. P. Wipf and B. D. Rao, "Sparse Bayesian learning for basis selection," *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2153–2164, Aug. 2004.
[13] S. Ji, Y. Xue, and L. Carin, "Bayesian compressive sensing," *IEEE Trans. Signal Process.*, vol. 56, no. 6, pp. 2346–2356, Jun. 2008.
[14] M. W. Seeger and H. Nickisch, "Compressed sensing and Bayesian experimental design," in *Proc. Int. Conf. Machine Learning (ICML)*, 2008, pp. 912–919.
[15] M. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, pp. 211–244, 2001.
[16] M. A. T. Figueiredo, "Adaptive sparseness for supervised learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 9, pp. 1150–1159, Sep. 2003.
[17] D. Wipf, J. Palmer, and B. D. Rao, "Perspectives on sparse Bayesian learning," *Advances in Neural Information Processing Systems*, no. 16, 2004.
[18] J. Palmer, D. Wipf, K. Kreutz-Delgado, and B. Rao, "Variational EM algorithms for non-Gaussian latent variable models," in *Advances in Neural Information Processing Systems 18*, Y. Weiss, B. Schölkopf, and J. Platt, Eds. Cambridge, MA: MIT Press, 2006, pp. 1059–1066.
[19] D. Wipf, J. Palmer, B. D. Rao, and K. Kreutz-Delgado, "Performance analysis of latent variable models with sparse priors," presented at the IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP), Honolulu, HI, May 2007.
[20] M. Girolami, "A variational method for learning sparse and overcomplete representations," *Neural Comput.*, vol. 13, no. 11, pp. 2517–2532, 2001.
[21] S. D. Babacan, R. Molina, and A. K. Katsaggelos, "Parameter estimation in TV image restoration using variational distribution approximation," *IEEE Trans. Image Process.*, no. 3, pp. 326–339, 2008.
[22] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer-Verlag, 2006.
[23] K. V. Mardia, J. T. Kent, and J. M. Bibby, *Multivariate Analysis*. New York: Academic, 1979.
[24] G. H. Golub and C. F. Van Loan, *Matrix Computations (Johns Hopkins Studies in Mathematical Sciences)*. Baltimore, MD: Johns Hopkins Univ. Press, Oct. 1996.
[25] D. J. C. MacKay, "Bayesian interpolation," *Neural Comput.*, vol. 4, no. 3, pp. 415–447, 1992.
[26] R. Molina, A. K. Katsaggelos, and J. Mateos, "Bayesian and regularization methods for hyperparameter estimation in image restoration," *IEEE Trans. Image Process.*, vol. 8, no. 2, pp. 231–246, Feb. 1999.
[27] M. Tipping and A. Faul, "Fast marginal likelihood maximisation for sparse Bayesian models," in *Proc. 9th Int. Workshop Artificial Intelligence and Statistics*, C. M. Bishop and B. J. Frey, Eds., 2003.
[28] S. Ji, D. Dunson, and L. Carin, "Multi-task compressive sensing," *IEEE Trans. Signal Process.*, vol. 57, no. 1, pp. 92–106, Jan. 2009.
[29] D. L. Donoho, I. Drori, Y. Tsaig, and J. L. Starck, Sparse Solution of Underdetermined Linear Equations by Stagewise Orthogonal Matching Pursuit 2006 [Online]. Available: http://www-stat.stanford.edu/~donoho/Reports/2006/StOMP-20060403.pdf, Tech. Rep.
[30] Y. Tsaig and D. L. Donoho, "Extensions of compressed sensing," *Signal Process.*, vol. 86, no. 3, pp. 549–571, 2006.
[31] Sparselab [Online]. Available: http://sparselab.stanford.edu/

S. Derin Babacan (S'02) was born in Istanbul, Turkey, in 1981. He received the B.Sc. degree from Bogazici University, Istanbul, in 2004, and the M.Sc. degree from Northwestern University, Evanston, IL, in 2006, where he is currently pursuing the Ph.D. degree in the Department of Electrical Engineering and Computer Science.

He is a Research Assistant with the Image and Video Processing Laboratory, Northwestern University. His primary research interests include image restoration, image and video compression, super resolution, and computer vision.

Mr. Babacan is the recipient of an IEEE International Conference on Image Processing Paper Award (2007).

Rafael Molina (M'87) was born in 1957. He received the degree in mathematics (statistics) in 1979 and the Ph.D. degree in optimal design in linear models in 1983.

He became a Professor of computer science and artificial intelligence at the University of Granada, Granada, Spain, in 2000. His areas of research interest are image restoration (applications to astronomy and medicine), parameter estimation in image restoration, super resolution of images and video, and blind deconvolution.

Aggelos K. Katsaggelos (S'87–M'87–SM'92–F'98) received the Diploma degree in electrical and mechanical engineering from the Aristotelian University of Thessaloniki, Greece, in 1979, and the M.S. and Ph.D. degrees in electrical engineering from the Georgia Institute of Technology, Atlanta, in 1981 and 1985, respectively.

In 1985, he joined the Department of Electrical and Computer Engineering, Northwestern University, Evanston, IL, where he is currently a Professor. He was the holder of the Ameritech Chair of Information Technology (1997–2003). He is also the Director of the Motorola Center for Seamless Communications and a member of the Academic Affiliate Staff, Department of Medicine, Evanston Hospital. He has published extensively in the areas of signal processing, multimedia transmission, and computer vision. He is the Editor of *Digital Image Restoration* (Springer-Verlag, 1991), coauthor of *Rate-Distortion Based Video Compression* (Kluwer, 1997), coeditor of *Recovery Techniques for Image and Video Compression and Transmission* (Kluwer. 1998), and coauthor of *Super-Resolution for Images and Video* (Claypool, 2007) and *Joint Source-Channel Video Transmission* (Claypool, 2007). He is also the coinventor of twelve international patents.

Dr. Katsaggelos has served the IEEE and other Professional Societies in many capacities. He is currently a member of the Publication Board of the IEEE Proceedings and has served as Editor-in-Chief of the IEEE SIGNAL PROCESSING MAGAZINE (1997–2002) and a member of the Board of Governors of the IEEE Signal Processing Society (1999–2001). He is the recipient of the IEEE Third Millennium Medal (2000), the IEEE Signal Processing Society Meritorious Service Award (2001), an IEEE Signal Processing Society Best Paper Award (2001), an IEEE International Conference on Multimedia and Expo Paper Award (2006), and an IEEE International Conference on Image Processing Paper Award (2007). He is a Distinguished Lecturer of the IEEE Signal Processing Society (2007–2008).