Bayesian Group-Sparse Modeling and Variational Inference

S. Derin Babacan, *Member, IEEE*, Shinichi Nakajima, Minh N. Do, *Senior Member, IEEE*

Abstract

In this paper, we present a general class of multivariate priors for group-sparse modeling within the Bayesian framework. We show that special cases of this class correspond to multivariate versions of several classical priors used for sparse modeling. Hence, this general prior formulation is helpful in analyzing the properties of different modeling approaches and their connections. We derive the estimation procedures with these priors using variational inference for fully Bayesian estimation. In addition, we discuss the differences between the proposed inference and deterministic inference approaches with these priors. Finally, we show the flexibility of this modeling by considering several extensions such as multiple measurements, within-group correlations and overlapping groups.

I. Introduction

We consider the general linear model given by

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

where $M \times 1$ observations $\mathbf{y}$ of the original unknown signal $\mathbf{w}$ are taken with an $M \times N$ measurement matrix (or dictionary) $\mathbf{\Phi} = [\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_N]$, and $\mathbf{n}$ represents the noise. This paper is concerned with the problem of finding an estimate of the unknown signal $\mathbf{w}$ from the observations $\mathbf{y}$. Generally, the case of interest is the $M \ll N$ regime, which makes the problem challenging and requires appropriate modeling of the unknown signal $\mathbf{w}$.

Problems of the general form (1) are very common in signal processing, statistics, neuroscience and machine learning. Typical applications include compressive sensing, sparse representation, super resolution, source localization, variable/model selection and prediction, among many others. A general design principle in these approaches is sparsity, which amounts to finding the most important components of $\mathbf{w}$ and suppressing the elements with relatively lower importance. In this design, the unknown vector $\mathbf{w}$ is assumed to contain a small number of nonzero elements, while the majority of the components are zero. This assumption is translated into the optimization problem for finding $\mathbf{w}$ using sparsity-promoting penalty functions, of which the most common example is the l_1-norm based formulation given by

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

This formulation is commonly referred to as basis pursuit [1] or lasso [2]. It implicitly models the noise $\mathbf{n}$ as zero-mean white Gaussian distributed with variance $\beta^{-1}/2$, and τ is the regularization parameter controlling the strength of the enforced sparsity. A large number of optimization methods have been developed for solving (2). In addition, different sparse signal models have been proposed extending the l_1-norm to the more general l_p-norm with $0 < p \leq 1$.

In the traditional sparse modeling, the sparsity constraint is imposed on individual components of $\mathbf{w}$. Recently, a different modeling approach has emerged where sparsity is enforced on groups instead of the individual components. This *group-sparse* (also called block-sparse) approach is a natural generalization of the traditional sparse modeling methods. It effectively models the *structural* properties of the signal by clustering relevant signal components together, such that dependencies between signal components are taken into account. It is also shown to lead to higher performance in pruning out irrelevant components compared to independent modeling of the coefficients [3].

S. Derin Babacan is with the Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, IL, USA. e-mail: dbabacan@illinois.edu

Shinichi Nakajima is with the Optical Research Laboratory, Nikon Corporation, Tokyo, 140-8601 Japan. e-mail: nakajima.s@nikon.co.jp

Minh N. Do is with the Department of Electrical and Computer Engineering and the Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, IL, USA. e-mail: minhdo@illinois.edu

Group-sparsity has recently been considered in compressive sensing [3]–[7] and machine learning [8]–[12], and is also closely related to signal modeling within union of subspaces [3], [13]–[15]. It has rapidly found applications in, e.g., imaging [16], [17] and network analysis [18], demonstrating promising performance.

A general optimization formulation for group-sparse regularization is

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \beta \left\| \mathbf{y} - \mathbf{\Phi}\mathbf{w} \right\|_2^2 + \tau \left\| \mathbf{w} \right\|_{1,2}, \tag{3}$$

where $\| \cdot \|_{1,2}$ denotes the combined l_1/l_2-norm with

$$\left\| \mathbf{w} \right\|_{1,2} = \sum_{i=1}^{G} \left\| \mathbf{w}_i \right\|_2 , \quad \left\| \mathbf{w}_i \right\|_2 = \sqrt{\mathbf{w}_i^T \mathbf{w}_i} , \tag{4}$$

where $\mathbf{w}_i$ denotes the i^{th} group, and G is the number of groups. Each group i contains d_i elements, such that $\sum_{i=1}^{G} d_i = N$ if the groups are not overlapping. It is clear that this formulation includes the traditional l_1-based formulation as a special case (when $d_i = 1, \ \forall i$).

The optimization problem (3) is similar to the l_1-based optimization, and thus some l_1-based approaches can be applied to this problem with some modifications. Deterministic methods directly addressing the problem (3) have been developed in [19]–[21], and in group-lasso methods [11], [12]. Several Bayesian approaches have been developed for group-sparse modeling: the Bayesian group-lasso [22] proposed to use multivariate Laplace priors on separate groups, and provided a sampling scheme for inference. A similar group-sparse prior is used in covariance estimation problem in [9]. In [8], [23], Laplacian scale mixtures have been used for the construction of the group-sparse prior, and the inference is performed using expectation-maximization (EM).

An important issue in all sparse reconstruction problems is choosing the regularization parameters β and τ. Clearly, optimizing (3) jointly with respect to them is not suitable since it results in the trivial solution $\beta = \tau = 0$. A similar problem is encountered when the problem is converted to weighted least squares problems, as in iteratively reweighted least squares (IRLS) with l_p-priors [24]–[26]. Deterministic heuristic methods are devised for parameter estimation, such as L-curves [27] or penalizing the trivial solution [24]. A more systematic approach can be obtained using Bayesian inference, as shown in this article.

In this paper, we present a Bayesian approach for group-sparse modeling and inference. Using a normal variance mixture formulation, we present the hierarchical construction of a general signal prior suitable for modeling group-sparse signals. This general signal prior contains a large class of distributions as special cases, obtained via different selections of distributions in the hierarchical construction. Using this general formulation, we explore different options for group-sparse modeling, analyze their connections, and their sparsity-enforcing properties. We show that some of the special cases of this generalized prior correspond to several standard models used in the sparse and group-sparse reconstruction literature. For estimation using this class of priors, we provide the hierarchical inference rules using the variational Bayesian (VB) approach for a fully-Bayesian estimation (i.e., including algorithmic parameters). We compare the proposed inference with deterministic inference approaches, and show the thresholding properties of different priors both in deterministic and Bayesian frameworks. Finally, we consider several extended modeling possibilities within Bayesian group-sparse modeling, such as within-group correlations and overlapping groups, and consider the multiple measurement vector case.

The rest of this paper is organized as follows. Section II provides the hierarchical construction of the generalized group-sparse prior using normal variance mixtures. We also derive its special cases and show their properties. In Section III, we develop fully-Bayesian inference methods using these priors via variational Bayesian approximation. Properties of the modeling and inference in comparison with deterministic approaches are discussed in Section IV. Several extensions to Bayesian group-sparse modeling are provided in Section V. Empirical evaluation of different aspects of the group-sparse modeling are presented in Section VI, and conclusions are drawn in Section VII.

II. Bayesian Group-Sparse Modeling

The Bayesian modeling of (1) requires the definition of a joint distribution of all unknown and observed quantities. This joint distribution typically includes the conditional distribution for the observations $\mathbf{y}$, and a prior that models the characteristics of the unknown signal $\mathbf{w}$. In the following, we first present a class of distributions suitable for group-sparse modeling of $\mathbf{w}$ using variance mixtures of Gaussian distributions. We then derive its special cases

and show the connections between them and models proposed in the literature. Finally, we complete the Bayesian model by specifying the observation model and hyperpriors assigned to the parameters of all distributions.

We use the following notation throughout this paper. Vectors are denoted by small-case bold letters $\mathbf{w}$, while matrices are in capital bold letters $\mathbf{W}$. $\mathrm{diag}(\mathbf{a})$ is a diagonal matrix with vector $\mathbf{a}$ as its diagonal, and $\langle\,\cdot\,\rangle$ denotes the expectation with respect to the corresponding distribution.

A. Signal Models

For modeling the unknown signal $\mathbf{w}$, we first define G groups of coefficients such that the vector $\mathbf{w}_i$ contains d_i signal coefficients assigned to group i. The case with $G = N$, $d_i = 1,\ \forall i$ corresponds to independent sparse modeling of the coefficients.

Assuming *a priori* independence between groups, we express the signal prior as

$$\mathrm{p}(\mathbf{w}|\mathbf{z}) = \prod_{i=1}^{G} \mathrm{p}(\mathbf{w}_i\,|z_i)\,, \tag{5}$$

where $\mathbf{z}$ is the vector containing all z_i. Sparsity is enforced on each group via the conditional priors $\mathrm{p}(\mathbf{w}_i\,|z_i)$. For their representation, we use the normal variance mixture model [28] (also called scale mixtures of Gaussians [29], [30]). Specifically, we represent each group $\mathbf{w}_i$ as

$$\mathbf{w}_i = \sqrt{z_i}\,\mathbf{x}\,, \tag{6}$$

where $z_i > 0$ and $\mathbf{x}$ is a standard multivariate Gaussian variable, i.e., $\mathbf{x} \sim \mathcal{N}(\mathbf{0}_{d_i}, \mathbf{I}_{d_i})$ with $\mathbf{0}_{d_i}$ a zero vector of length d_i and $\mathbf{I}_{d_i}$ the $d_i \times d_i$ identity matrix. It is clear that given z_i, $\mathbf{w}_i$ is a multivariate Gaussian variable with zero mean and variance $z_i\,\mathbf{I}_{d_i}$, that is

$$\mathrm{p}(\mathbf{w}_i\,|z_i) = \mathcal{N}(\mathbf{0}_{d_i}, z_i\,\mathbf{I}_{d_i})\,. \tag{7}$$

Notice that the coefficients within each group are not independent. The marginal probability distribution of $\mathbf{w}_i$ can be found by integrating out the latent variables as

$$\mathrm{p}(\mathbf{w}_i) = \int_0^\infty \mathrm{p}(\mathbf{w}_i\,|\,z_i)\,\mathrm{p}(z_i)\,dz_i\,. \tag{8}$$

Here, $\mathrm{p}(z_i)$ is called the mixing distribution and determines the form of the marginal distribution $\mathrm{p}(\mathbf{w}_i)$.

Normal variance mixtures have been extensively used in the literature for representing a large number of distributions, and for deriving efficient inference procedures for parameter estimation. A variety of distributions can be represented in this fashion by different selection of the mixing distribution $\mathrm{p}(z_i)$.

In this paper, for the mixing distribution $\mathrm{p}(z_i)$ we consider the generalized inverse Gaussian (GIG) distribution

$$\mathrm{p}(z_i|a_i, b_i, \lambda_i) = \frac{(a_i/b_i)^{\lambda_i/2}}{2K_{\lambda_i}(\sqrt{a_i b_i})}\, z_i^{\lambda_i - 1} \exp\left(-\frac{1}{2}\left(a_i\, z_i + b_i\, z_i^{-1}\right)\right)\,, \tag{9}$$

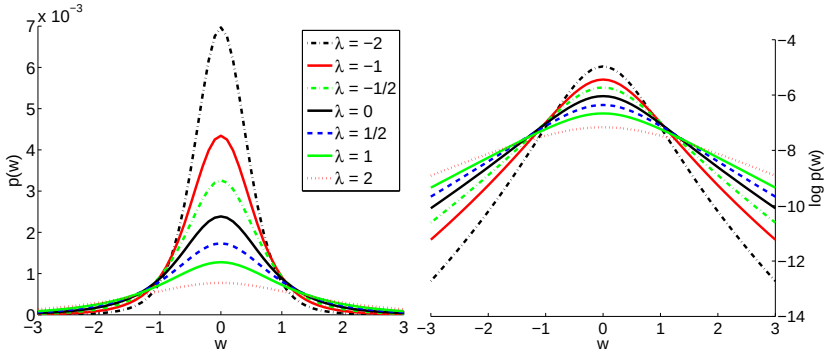
where K_{λ_i} is the modified Bessel function of the second kind. The moments of this distribution are given by [31]

$$\langle z_i^p \rangle = \frac{K_{\lambda_i + p}(\sqrt{a_i b_i})}{K_{\lambda_i}(\sqrt{a_i b_i})} \left(\frac{b_i}{a_i}\right)^{p/2}\,. \tag{10}$$

With this mixing density, the marginal distribution of $\mathbf{w}_i$ is found from (8) as the generalized hyperbolic (GH) distribution [28]

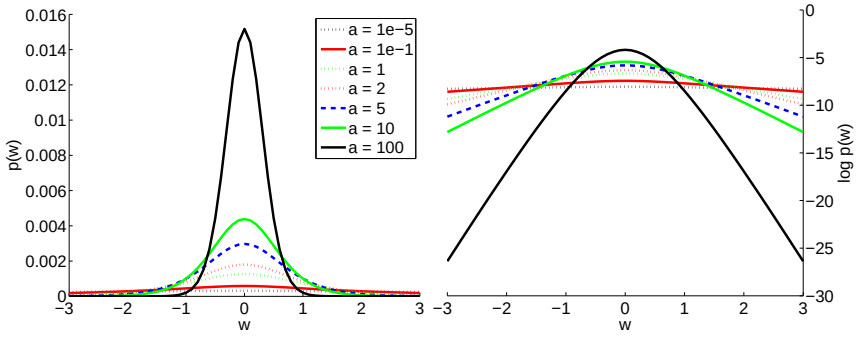
$$\mathrm{p}(\mathbf{w}_i|a_i, b_i, \lambda_i) = \frac{a_i^{d_i/4}}{(2\pi)^{d_i/2}} \frac{b_i^{-\lambda_i/2}}{K_{\lambda_i}(\sqrt{a_i b_i})} \frac{K_{\lambda_i - d_i/2}(\sqrt{a_i}\,\sqrt{b_i + \|\mathbf{w}_i\|_2^2})}{\left(b_i + \|\mathbf{w}_i\|_2^2\right)^{d_i/4 - \lambda_i/2}}\,. \tag{11}$$

In this paper, we chose the GIG distribution as the mixing distribution as it includes a fairly broad class of distributions commonly used as hyperpriors, and the resulting marginal distribution, the GH distribution, again covers a large number of distributions as special cases. Due to this generalization, we are able to analyze the connections between different modeling strategies. As we shall see in the following, several special cases correspond to standard priors commonly used in sparse modeling.



(a) (b)

Fig. 1. Generalized hyperbolic distributions (a) and log-distributions (b) with varying λ_i, when $a_i = 1$, $b_i = 1$ ($d = 2$, the cross-section is shown).



(a) (b)

Fig. 2. Generalized hyperbolic distributions (a) and log-distributions (b) with varying a, when $\lambda_i = 1$, $b = 1$ ($d = 2$, the cross-section is shown).

To see the rich family of distributions that can be obtained from the GH distribution, distributions obtained with varying values of a_i, b_i and λ_i are depicted in Figs. 1-3. It can be seen that both the central and tail behavior can be varied using different parameter values, and as will be shown later, the resulting distributions have different estimation characteristics. In the following, we consider the special cases of the GH distribution at the limit parameter values, along with the mixing distributions. Let us first give some expressions on asymptotic approximations of the modified Bessel function that will be useful:

$$K_\lambda(x) \approx \frac{1}{2}\Gamma(\lambda)\left(\frac{x}{2}\right)^{-\lambda}, \quad \text{for } \lambda > 0,\ x \to 0 \tag{12}$$

$$K_\lambda(x) \approx \frac{1}{2}\Gamma(-\lambda)\left(\frac{x}{2}\right)^{\lambda}, \quad \text{for } \lambda < 0,\ x \to 0 \tag{13}$$

$$K_0(x) \approx -\ln(x) \tag{14}$$

$$K_\lambda(x) \approx \sqrt{\frac{\pi}{2x}}\exp(-x), \quad \text{for } x \to \infty, \tag{15}$$

and for integer λ,

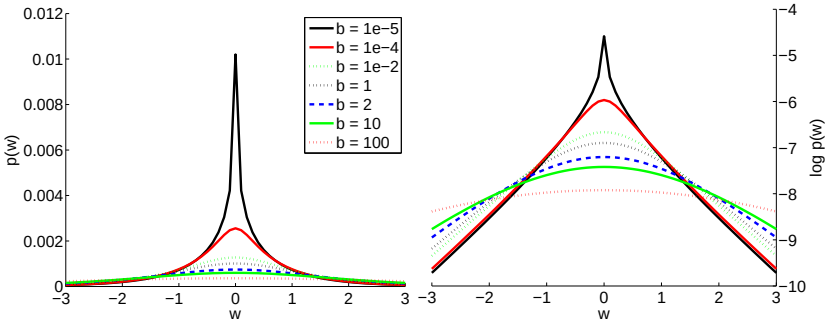
$$K_{\lambda+1/2}(x) = \sqrt{\frac{\pi}{2x}}\exp(-x)\left[1 + \sum_{i=1}^{\lambda}\frac{(\lambda+i)!}{(\lambda-i)!i!}(2x)^{-i}\right]. \tag{16}$$

1) McKay's Bessel function distribution: When $b_i \to 0$ with $\lambda_i > 0$, the mixing GIG distribution reduces to the gamma distribution, given by

$$\mathrm{p}(z_i|a_i,\lambda_i) = \frac{a_i^{\lambda_i/2}}{2^{\lambda_i}\Gamma(\lambda_i)} z_i^{\lambda_i-1}\exp\left(-\frac{1}{2}a_i z_i\right). \tag{17}$$

The corresponding marginal distribution is

$$\mathrm{p}(\mathbf{w}_i|a_i,\lambda_i) = \frac{a_i^{d_i/4+\lambda_i/2}}{\pi^{d_i/2}\,2^{d_i/2+\lambda_i-1}} \frac{(\|\mathbf{w}_i\|_2)^{\lambda_i-d_i/2}}{\Gamma(\lambda_i)} K_{\lambda_i-d_i/2}(\sqrt{a_i}\,\|\mathbf{w}_i\|_2), \tag{18}$$



(a) (b)

Fig. 3. Generalized hyperbolic distributions (a) and log-distributions (b) with varying b, when $\lambda_i = 1$, $a = 1$ ($d = 2$, the cross-section is shown).

which is McKay's Bessel function distribution [32]–[34] (also called multivariate variance-gamma [35], multivariate generalized Laplace [33], or multivariate K distribution [36], [37]).

We now consider two special cases of (18) that are related to the Laplace distribution. In the case $\lambda_i = 1$, the mixing distribution becomes the exponential distribution

$$\mathrm{p}(z_i|a_i) = \frac{a_i}{2} \exp\left(-\frac{1}{2} a_i\, z_i\right) , \tag{19}$$

such that the marginal becomes

$$\mathrm{p}(\mathbf{w}_i|a_i) = \frac{a_i^{d_i/4+1/2}}{(2\pi)^{d_i/2}} \; (\|\mathbf{w}_i\|_2)^{1-d_i/2} \; K_{1-d_i/2}(\sqrt{a_i}\,\|\mathbf{w}_i\|_2)\,. \tag{20}$$

To see the relation with the univariate Laplace distribution, we can use (16) and rewrite (20) for odd d_i as

$$\mathrm{p}(\mathbf{w}_i|a_i) \propto \frac{\exp(-\sqrt{a_i}\|\mathbf{w}_i\|_2)}{(\|\mathbf{w}_i\|_2)^{d_i/2-1/2}} \left[\sum_{i=1}^{(d_i-3)/2} \frac{((d_i-3)/2+i)!}{((d_i-3)/2-i)!\,i!} \,(2\sqrt{a_i}\|\mathbf{w}_i\|_2)^{-i} + 1\right] . \tag{21}$$

The similarity to the univariate case can be seen from the exponential term, and noticing that all other terms vanish with $d_i = 1$. Note, however, that there are additional terms that are power functions of $\|\mathbf{w}_i\|_2$. A more directly related case can be found by the selection $\lambda_i = (d_i+1)/2$, which simplifies (18) using (16) as

$$\mathrm{p}(\mathbf{w}_i|a_i) \propto a_i^{d_i/2} \exp(-\sqrt{a_i}\|\mathbf{w}_i\|_2)\,, \tag{22}$$

in which case the mixing distribution is a gamma distribution given by

$$\mathrm{p}(z_i|a_i) = \frac{a_i^{(d_i+1)/4}}{2^{(d_i+1)/2}\Gamma((d_i+1)/2)}\, z_i^{d_i/2-1/2} \exp\left(-\frac{1}{2} a_i\, z_i\right) . \tag{23}$$

Both distributions (20) and (22) were termed as multivariate Laplace distributions in the literature: the form in (20) is used in [37], [38] due to the similarity of the hierarchical structure to the univariate case, and (22) is used in the Bayesian group-lasso method [22] due to the similarity of the marginal distributions. Here we will use the term multivariate Laplace for the distribution in (22) since it has an estimation behavior similar to the univariate case (see Section IV-A). The distribution in (20) will be referred to as McKay($\lambda = 1$). Notice that both distributions reduce to the univariate Laplace distribution when $d_i = 1$.

It is also possible to integrate out a_i from $\mathrm{p}(\mathbf{w}_i|a_i)$ by assigning a gamma hyperprior on $\sqrt{a_i}$. When $\lambda_i = (d_i+1)/2$, the corresponding marginal has a closed form and is given by

$$\mathrm{p}(\mathbf{w}_i|k_a,\theta_a) = \Gamma(d_i+k_a-1)\,[\theta_a + \|\mathbf{w}_i\|_2]^{-(d_i+k_a)} , \tag{24}$$

which is the multivariate version of the generalized double Pareto distribution [39], [40].

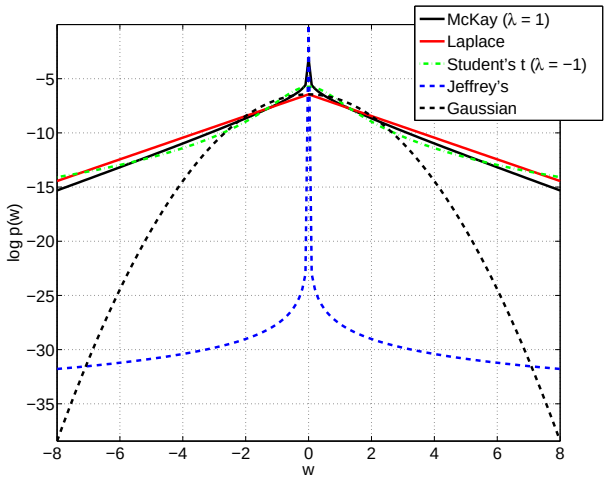


Fig. 4. McKay, Laplace, Student's t, Jeffrey's and Gaussian log-distributions ($d = 2$, the cross-section is shown).

2) Multivariate Student's t: When $a_i \to 0$ with $\lambda_i < 0$, we have the inverse gamma distribution as the mixing density

$$\mathrm{p}(z_i|b_i,\lambda_i) = \frac{(b_i/2)^{-\lambda_i}}{\Gamma(-\lambda_i)}\, z_i^{\lambda_i - 1} \exp\left(-\frac{1}{2} b_i\, z_i^{-1}\right). \tag{25}$$

The corresponding marginal is given by

$$\mathrm{p}(\mathbf{w}_i|b_i,\lambda_i) = \left(\frac{1}{\pi}\right)^{d_i/2} \frac{\Gamma(-\lambda_i + d_i/2)}{b_i^{\lambda_i}\,\Gamma(-\lambda)} \left(b_i + \|\mathbf{w}_i\|_2^2\right)^{\lambda_i - d_i/2}, \tag{26}$$

which is a multivariate Student's t distribution with $-2\lambda_i$ degrees of freedom.

Finally, when $a_i \to 0$, $b_i \to 0$ and $\lambda_i \to 0$, we have the Jeffrey's non-informative prior $\mathrm{p}(z_i) \propto z_i^{-1}$. In this case, the marginal distribution becomes

$$\mathrm{p}(\mathbf{w}_i) \propto \left(\frac{1}{\|\mathbf{w}_i\|_2}\right)^{d_i}. \tag{27}$$

In summary, the variance mixture model with the GIG mixture distribution includes a number of classical distributions as special cases at the limiting values of its parameters. In the following, we mainly limit our discussion to the four distributions described above, i.e, multivariate McKay($\lambda = 1$), Laplace, Student's t distributions and Jeffrey's prior. These distributions along with the corresponding parameter selections are summarized in Table I.

The log-distributions for all cases are shown in Fig. 4, along with the Gaussian distribution. It is evident that all distributions have heavy-tails, which is generally considered to be a desirable property for enforcing sparsity and variable selection.

B. Complete Model

After the signal model is defined, we complete the Bayesian model characterization by modeling the observations $\mathbf{y}$ in (1). Assuming independent Gaussian noise with zero mean and variance equal to β^{-1}, the conditional distribution is expressed as

$$\mathrm{p}(\mathbf{y}|\mathbf{w},\beta) = \mathcal{N}(\mathbf{y}|\mathbf{\Phi}\mathbf{w},\beta^{-1}), \tag{28}$$

with a conjugate gamma prior placed on β as

$$\mathrm{p}(\beta|k_\beta,\theta_\beta) = \Gamma(\beta|k_\beta,\theta_\beta). \tag{29}$$

A prior is called conjugate if it leads to a posterior distribution that has the same functional form as the prior [41]. The use of conjugate priors significantly simplify the form of posterior distributions. Combining (28), (29) and the hierarchical signal prior (7) and (9), we define the joint probability distribution as

$$\mathrm{p}(\mathbf{y},\mathbf{w},\mathbf{z},\beta) = \mathrm{p}(\mathbf{y}|\mathbf{w},\beta)\,\mathrm{p}(\mathbf{w}|\mathbf{z})\,\mathrm{p}(\mathbf{z}|\mathbf{a},\mathbf{b},\boldsymbol{\lambda})\,\mathrm{p}(\beta)\,\mathrm{p}(\mathbf{a},\mathbf{b}). \tag{30}$$

TABLE I
SUMMARY OF DISTRIBUTIONS AND PARAMETER ESTIMATES

Parameter Values	Distribution	Mixing Distribution	Variance Parameter Update $\langle z_i^{-1} \rangle$	Hyperparameter update
-	Generalized Hyperbolic	Generalized Inverse Gaussian	$\frac{\sqrt{a}}{\sqrt{\langle \|\mathbf{w}_i\|_2^2 \rangle + b}} \frac{K_{\lambda_i - d_i/2-1}(\sqrt{a}\sqrt{\langle \|\mathbf{w}_i\|_2^2 \rangle + b_i})}{K_{\lambda_i - d_i/2}(\sqrt{a_i}\sqrt{\langle \|\mathbf{w}_i\|_2^2 \rangle + b_i})}$	see below for a_i, b_i; λ_i not provided
$b_i \to 0$, $\lambda_i > 0$	McKay's Bessel function	Gamma	$\frac{\sqrt{a_i}}{\sqrt{\langle \|\mathbf{w}_i\|_2^2 \rangle}} \frac{K_{\lambda_i - d_i/2-1}(\sqrt{a_i}\sqrt{\langle \|\mathbf{w}_i\|_2^2 \rangle})}{K_{\lambda - d_i/2}(\sqrt{a_i}\sqrt{\langle \|\mathbf{w}_i\|_2^2 \rangle})}$	$\langle a_i \rangle = (k_a + \lambda_i)\left(\theta_a + \frac{\langle z_i \rangle}{2}\right)^{-1}$
$b_i \to 0$, $\lambda_i = (d_i+1)/2$	Multivariate Laplace	Gamma	$\frac{\sqrt{a_i}}{\sqrt{\langle \|\mathbf{w}_i\|_2^2 \rangle}}$	$\langle a_i \rangle = \left(k_a + \frac{d_i+1}{2}\right)\left(\theta_a + \frac{\langle z_i \rangle}{2}\right)^{-1}$
$a_i \to 0$, $\lambda_i < 0$	Multivariate Student's t	Inverse Gamma	$\frac{d_i/2-\lambda}{\frac{1}{2}\left(\langle \|\mathbf{w}_i\|_2^2 \rangle + b_i\right)}$	$\langle b_i \rangle = (k_b - \lambda_i)\left(\theta_b + \frac{\langle z_i^{-1} \rangle}{2}\right)^{-1}$
$a_i \to 0$, $b_i \to 0$, $\lambda_i \to 0$	Jeffrey's	Jeffrey's	$\frac{d_i}{\langle \|\mathbf{w}_i\|_2^2 \rangle}$	None

where $\mathbf{z}$, $\mathbf{a}$, $\mathbf{b}$, $\boldsymbol{\lambda}$ are vectors containing z_i, a_i, b_i, and λ_i, $\forall i$ respectively. The hyperprior $\mathrm{p}(\mathbf{a}, \mathbf{b})$ is used to model the parameters $\mathbf{a}$ and $\mathbf{b}$ for their estimation, and will be discussed in Section III-C.

III. VARIATIONAL INFERENCE

Bayesian inference is based on the posterior distribution $\mathrm{p}(\boldsymbol{\xi}|\mathbf{y}) = \mathrm{p}(\boldsymbol{\xi}, \mathbf{y})/\mathrm{p}(\mathbf{y})$, where $\boldsymbol{\xi}$ denotes the set of all unknowns such that $\boldsymbol{\xi} = \{\mathbf{w}, \mathbf{z}, \beta, \mathbf{a}, \mathbf{b}\}$. However, as in many multidimensional problems, the Bayesian model defined with the joint distribution in (30) does not allow for exact inference as the marginal distribution $\mathrm{p}(\mathbf{y})$ is intractable. Therefore, approximation methods must be used for the inference. In the following, we use the variational Bayesian (VB) approximation [42], [43], which has attractive computational properties along with high estimation performance. With the definition of the joint distribution in (30), the variational Bayes method provides a distribution $\mathrm{q}(\boldsymbol{\xi})$ that approximates the posterior $\mathrm{p}(\boldsymbol{\xi}|\mathbf{y})$. Specifically, $\mathrm{q}(\boldsymbol{\xi})$ is found by minimizing the Kullback-Leibler (KL) divergence between the approximation and the unknown posterior as [42], [43]

$$\mathrm{q}^*(\boldsymbol{\xi}) = \underset{\mathrm{q}(\boldsymbol{\xi})}{\arg\min}\ \mathrm{KL}(\mathrm{q}(\boldsymbol{\xi}) \,||\, \mathrm{p}(\boldsymbol{\xi}|\mathbf{y})) = \underset{\mathrm{q}(\boldsymbol{\xi})}{\arg\min} \int \mathrm{q}(\boldsymbol{\xi}) \log \frac{\mathrm{q}(\boldsymbol{\xi})}{\mathrm{p}(\boldsymbol{\xi}|\mathbf{y})} d\xi \tag{31}$$

$$= \underset{\mathrm{q}(\boldsymbol{\xi})}{\arg\min} \int \mathrm{q}(\boldsymbol{\xi}) \log \frac{\mathrm{q}(\boldsymbol{\xi})}{\mathrm{p}(\boldsymbol{\xi}, \mathbf{y})} d\xi + \mathrm{const}\,, \tag{32}$$

where $\mathrm{p}(\mathbf{y}, \boldsymbol{\xi})$ is the joint probability distribution given in (30). To solve this optimization, the only assumption needed is an appropriate factorization of $\mathrm{q}(\boldsymbol{\xi})$. Here we use the mean-field approximation [42] with

$$\mathrm{q}(\mathbf{w}, \mathbf{z}, \beta, \mathbf{a}, \mathbf{b}) = \mathrm{q}(\mathbf{w})\, \mathrm{q}(\mathbf{z})\, \mathrm{q}(\beta)\, \mathrm{q}(\mathbf{a}, \mathbf{b})\,. \tag{33}$$

Using this factorization in (32), the distributions $\mathrm{q}(\boldsymbol{\xi}_k)$ of each variable $\boldsymbol{\xi}_k \in \boldsymbol{\xi}$ is found as [42], [43]

$$\log \mathrm{q}^*(\boldsymbol{\xi}_k) = \left\langle \log \mathrm{p}(\mathbf{y}, \boldsymbol{\xi}) \right\rangle_{\mathrm{q}(\boldsymbol{\xi} \backslash \boldsymbol{\xi}_k)} + \mathrm{const}, \tag{34}$$

$$= \left\langle \log \mathrm{p}(\mathbf{y}|\mathbf{w}, \beta)\, \mathrm{p}(\mathbf{w}|\mathbf{z})\, \mathrm{p}(\mathbf{z}|\mathbf{a}, \mathbf{b}, \boldsymbol{\lambda})\, \mathrm{p}(\beta)\, \mathrm{p}(\mathbf{a}, \mathbf{b}) \right\rangle_{\mathrm{q}(\boldsymbol{\xi} \backslash \boldsymbol{\xi}_k)} + \mathrm{const}, \tag{35}$$

where $\boldsymbol{\xi} \backslash \boldsymbol{\xi}_k$ denotes the set $\boldsymbol{\xi}$ with $\boldsymbol{\xi}_k$ removed. Individual distributions $\mathrm{q}(\boldsymbol{\xi}_k)$ are updated by (35) at each iteration by fixing the remaining distributions $\mathrm{q}(\boldsymbol{\xi} \backslash \boldsymbol{\xi}_k)$, which corresponds to an alternating minimization of the KL divergence in (32). This iterative procedure is repeated until the KL distance converges.

The VB method is a generalization of the maximum *a posteriori* (MAP) and expectation-maximization (EM) methods. The EM estimates can be found by restricting some distributions $\mathrm{q}(\boldsymbol{\xi}_k)$ to be degenerate, i.e., delta distributions at a particular value. On the other hand, MAP solutions can be found by restricting all of the distributions to be degenerate. When a distribution is degenerate, it can be shown from (32) that its corresponding estimation amounts to minimizing the negative expected log joint distribution $-\left\langle \log \mathrm{p}(\boldsymbol{\xi}, \mathbf{y}) \right\rangle_{\mathrm{q}(\boldsymbol{\xi} \backslash \boldsymbol{\xi}_k)}$, which reduces to the log joint distribution in the case of MAP. We will discuss the MAP estimation in more detail in Section IV.

In the following subsections, we provide the explicit forms of the update rules for all unknowns. For notational simplicity, the optimal distributions are denoted by q instead of q^*.

A. Signal Estimate

From (35), the posterior approximation of $\mathbf{w}$ is found as a multivariate Gaussian

$$\mathrm{q}(\mathbf{w}) = \mathcal{N}(\langle \mathbf{w} \rangle, \boldsymbol{\Sigma}_{\mathbf{w}}) \,, \tag{36}$$

with parameters

$$\langle \mathbf{w} \rangle = \boldsymbol{\Sigma}_{\mathbf{w}} \, \beta \, \boldsymbol{\Phi}^T \mathbf{y} \,, \tag{37}$$

$$\boldsymbol{\Sigma}_{\mathbf{w}} = \left(\beta \, \boldsymbol{\Phi}^T \boldsymbol{\Phi} + \boldsymbol{\Lambda} \right)^{-1} \tag{38}$$

$$= \boldsymbol{\Lambda}^{-1} - \boldsymbol{\Lambda}^{-1} \boldsymbol{\Phi}^T \left(\beta^{-1} \mathbf{I} + \boldsymbol{\Phi} \boldsymbol{\Lambda}^{-1} \boldsymbol{\Phi}^T \right)^{-1} \boldsymbol{\Phi} \boldsymbol{\Lambda}^{-1} \,, \tag{39}$$

with $\boldsymbol{\Lambda} = \mathrm{diag}\left(\langle z_i^{-1} \rangle \right)$, with each $\langle z_i^{-1} \rangle$ repeated d_i times[1]. It can be seen from (38) that except when $\boldsymbol{\Phi}^T \boldsymbol{\Phi} = \mathbf{I}$, the groups are *a posteriori* dependent, despite the *a priori* independence assumption in (5). Sparsity in the groups occur when particular variables $\langle z_i^{-1} \rangle \to \infty$, in which case the i^{th} group is pruned out from the signal estimate[2]. Notice also the estimation of $\boldsymbol{\Sigma}_{\mathbf{w}}$ requires the inversion of an $N \times N$ matrix using (38), and an $M \times M$ matrix using (39).

B. Estimation of the Variance Parameters

The crucial part of (37) is the estimates of z_i^{-1}, which control the sparsity and hence the structure of the signal estimate. Here we derive the estimation rules for the general case with the GIG hyperprior, from which the special cases can easily be obtained.

First, with some algebra, it can be derived from (35) in combination with (33) that the distribution $\mathrm{q}(\mathbf{z})$ factorizes over $\mathrm{q}(z_i)$, such that

$$\mathrm{q}(\mathbf{z}) = \prod_{i=1}^{G} \mathrm{q}(z_i) \,. \tag{40}$$

Therefore, in the following we provide the update rules for each distribution $\mathrm{q}(z_i)$. Using (35), we find the approximate posterior $\mathrm{q}(z_i)$ from (7) and (9) as a GIG distribution

$$\mathrm{q}(z_i) \propto \; z_i^{\lambda_i - 1 - d_i/2} \exp\left(-\frac{1}{2} \left(a_i \, z_i + z_i^{-1} \left(\langle \|\mathbf{w}_i\|_2^2 \rangle + b_i \right) \right) \right) \,, \tag{41}$$

with the expectation $\langle \|\mathbf{w}_i\|_2^2 \rangle$ computed as

$$\langle \|\mathbf{w}_i\|_2^2 \rangle = \|\langle \mathbf{w}_i \rangle\|_2^2 + \mathrm{trace}\left(\boldsymbol{\Sigma}_{\mathbf{w}_i} \right) \,, \tag{42}$$

where $\boldsymbol{\Sigma}_{\mathbf{w}_i}$ denotes the submatrix of $\boldsymbol{\Sigma}_{\mathbf{w}}$ corresponding to the i^{th} group. The posterior estimate of $\langle z_i^{-1} \rangle$ can be calculated by the moments of this distribution in (10) as

$$\langle z_i^{-1} \rangle = \frac{\sqrt{a_i}}{\sqrt{\langle \|\mathbf{w}_i\|_2^2 \rangle + b}} \, \frac{K_{\lambda_i - d_i/2 - 1}(\sqrt{a_i} \, \sqrt{\langle \|\mathbf{w}_i\|_2^2 \rangle + b_i})}{K_{\lambda_i - d_i/2}(\sqrt{a_i} \, \sqrt{\langle \|\mathbf{w}_i\|_2^2 \rangle + b_i})} \,. \tag{43}$$

The update rules for the limiting cases can be found from this general form, and are shown in the third column of Table I.

[1]Notice that this assumes non-overlapping groups; overlapping groups will be discussed later.

[2]The modeling used in this paper does not allow for exact sparsity. However, sparsity occurs in practice when estimates $\langle z_i^{-1} \rangle$ become very large such that the coefficients in the i^{th} group are numerically indistinguishable from zero.

C. Estimation of the Hyperparameters a_i and b_i

Notice that in the general case (43), the posterior estimate of z_i^{-1} contains the hyperparameters a_i, b_i, and λ_i, which determine the shape of the enforced distribution on $\mathbf{w}_i$. With the variational approximation, their posterior distributions can be estimated using (35) as well, with the appropriate selection of the hyperpriors $\mathrm{p}(a_i)$, $\mathrm{p}(b_i)$ and $\mathrm{p}(\lambda_i)$ (or with a joint hyperprior $\mathrm{p}(a_i, b_i, \lambda_i)$). However, in the general case with GIG mixing distribution, the joint estimation of all a_i, b_i and λ_i is challenging: the estimation of λ_i requires numerical solutions (instead of analytical closed form updates), and when all parameters are jointly estimated, the accuracy greatly depends on the initial estimates.

Therefore, we instead provide hyperparameter estimates of a_i and b_i in the special cases, and leave λ_i as a free parameter.

1) McKay's Bessel function distribution: Recall that with $b_i \to 0$ and $\lambda_i > 0$, we have the gamma distribution (17) as the mixing density. As the corresponding hyperprior for a_i, we choose the conjugate gamma distribution

$$\mathrm{p}(a_i) = \Gamma(a_i; k_a, \theta_a)\,, \tag{44}$$

with the shape parameter k_a and the inverse scale parameter θ_a. The posterior becomes

$$\mathrm{q}(a_i) \propto \Gamma(a_i; k_a + \lambda_i, \theta_a + \frac{\langle z_i \rangle}{2})\,, \tag{45}$$

with the corresponding update

$$\langle a_i \rangle = (k_a + \lambda_i) \left(\theta_a + \frac{\langle z_i \rangle}{2} \right)^{-1} . \tag{46}$$

The moment $\langle z_i \rangle$ can be found from (41) using (10).

2) Multivariate Student's t: When $a_i \to 0$ with $\lambda_i < 0$, the mixing distribution (25) is an inverse gamma distribution in terms of z_i, but it is a gamma distribution with respect to the parameter b_i. Hence we choose the gamma distribution that is conjugate for b_i

$$\mathrm{p}(b_i) = \Gamma(b_i; k_b, \theta_b)\,. \tag{47}$$

The posterior distribution is found as a gamma distribution

$$\mathrm{q}(b_i) \propto \Gamma(b_i; k_b - \lambda_i, \theta_a + \frac{\langle z_i^{-1} \rangle}{2})\,, \tag{48}$$

with mean

$$\langle b_i \rangle = (k_b - \lambda_i) \left(\theta_b + \frac{\langle z_i^{-1} \rangle}{2} \right)^{-1} . \tag{49}$$

D. Estimation of the noise variance

The Bayesian methodology allows for the estimation of the noise variance as well. Using the prior in (29), the posterior of β becomes a gamma distribution, and β can be estimated using its mean as

$$\langle \beta \rangle = \frac{2k_\beta + M}{2\theta_\beta + \langle \|\mathbf{y} - \mathbf{\Phi}\mathbf{w}\|_2^2 \rangle}\,, \tag{50}$$

with the expectation given by

$$\langle \|\mathbf{y} - \mathbf{\Phi}\mathbf{w}\|_2^2 \rangle = \|\mathbf{y} - \mathbf{\Phi}\langle \mathbf{w} \rangle\|_2^2 + \mathrm{trace}\left(\mathbf{\Phi}^T \mathbf{\Phi}\, \mathbf{\Sigma}_\mathbf{w} \right)\,. \tag{51}$$

E. Summary

The signal priors presented in the previous sections, along with the corresponding mixing distributions and variational estimation rules are summarized in Table I. The algorithm alternates between estimating the signal $\mathbf{w}$ using (37), and the variances $\mathbf{z}$ and hyperparameters $\mathbf{a}$, $\mathbf{b}$ using the equations shown in Table I, according to the selected signal distribution.

The normal variance mixture with the GIG mixing distribution is extremely flexible, and encompasses a large family of distributions some of which can be used for modeling group-sparse signals. Other, non-standard, distributions can also be obtained by further extending the hierarchical construction and marginalization. The advantages of using the variance mixture formulation are the tractable properties of the Gaussian distribution obtained for the signal estimate in (36) and the conjugate prior mechanism that allows for closed-form estimation of the parameters.

In this work, we used a three-level hierarchical estimation procedure, involving the estimation of $\mathbf{w}_i$, z_i, a_i and b_i in alternating fashion. Instead, two-level hierarchical estimation procedures can be devised using the marginal distributions $\mathrm{p}(\mathbf{w}_i|a_i, b_i, \lambda_i)$ and appropriate hyperpriors on a_i and b_i (therefore bypassing the estimation of z_i). This approach is a generalization of Laplacian scale mixtures [8]. However, this approach brings some difficulties: First, the marginal distributions have complicated forms and the corresponding conjugate hyperpriors on a_i and b_i are hard to find. Second, the marginal distributions generally do not allow for closed form updates of the posterior mean $\mathbf{w}$. Finally, the posterior mean updates of a_i and b_i in general require expectations that do not have closed forms. Hence, fully-Bayesian inference with this two-level hierarchy is generally hard. Note, however, that if parameter estimation is not desired, deterministic approaches can be used (see Section IV) with relative ease with some forms of the marginal distributions, e.g., the Laplace distribution. This approach is closely related to reweighted l_1-minimization schemes [25], [26] and the EM approach presented in [8].

IV. Comparison with Deterministic Estimation

The signal priors considered in Section II-A can also be used in a deterministic maximum *a posteriori* (MAP) framework, which is commonly encountered in the literature. Using a deterministic framework allows us to show some interesting connections between different signal priors and also compare and demonstrate some properties of the variational Bayesian estimation described before.

When considering MAP optimization with the Bayesian model in this paper, two approaches can be considered.

A. MAP estimation using marginal distributions

By forming the joint probability distribution $\mathrm{p}(\mathbf{y}|\mathbf{w}, \beta)\mathrm{p}(\mathbf{w}|\mathbf{a}, \mathbf{b}, \lambda)$ using the observation model in (28) and the generalized hyperbolic distribution in (11) as the signal prior, and applying a $\log$-transform, we obtain the MAP estimate as

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} \log \mathrm{p}(\mathbf{y}|\mathbf{w}, \beta) + \sum_{i=1}^{G} \log \mathrm{p}(\mathbf{w}_i|a_i, b_i, \lambda_i) \tag{52}$$

$$= \arg\min_{\mathbf{w}} \beta\|\mathbf{y} - \mathbf{\Phi}\mathbf{w}\|_2^2 - \sum_{i=1}^{G} 2\log \frac{K_{\lambda_i - d_i/2}(\sqrt{a_i}\sqrt{b_i + \|\mathbf{w}_i\|_2^2})}{\left(b_i + \|\mathbf{w}_i\|_2^2\right)^{d_i/4 - \lambda_i/2}}\,. \tag{53}$$

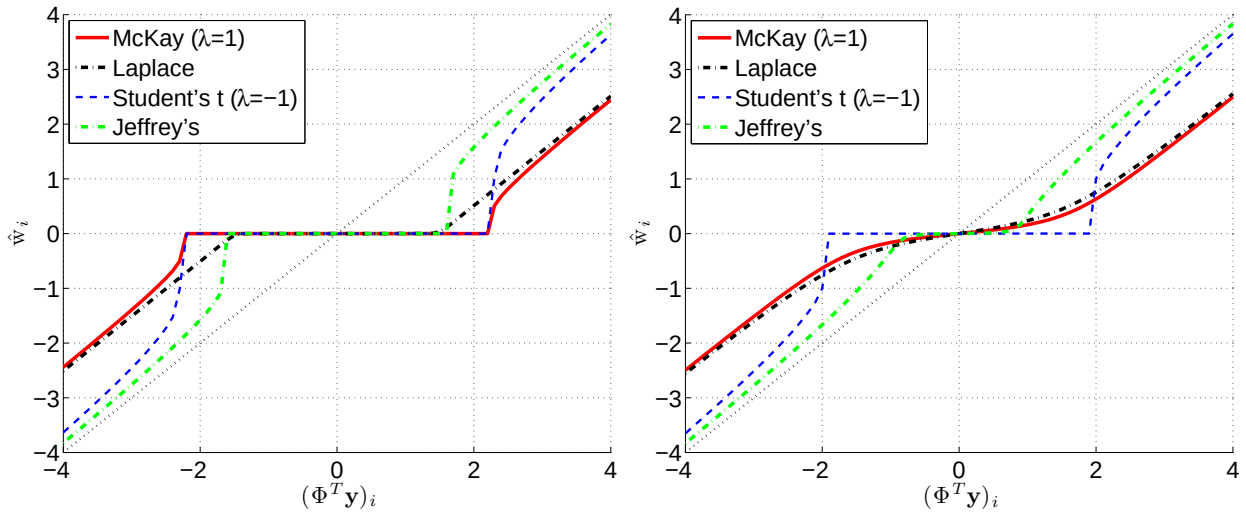
Note that the mode of the posterior distribution is sought within this formulation. In the general case with nonzero a_i, b_i, and λ_i, closed form updates for $\mathbf{w}_i$ cannot be found and numerical solutions are required. However, closed-form updates can easily be found in the case of multivariate Laplace (22) and t-distributions (26), and Jeffrey's prior (27).

In the case of multivariate Laplace priors, the optimization problem becomes

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \beta\|\mathbf{y} - \mathbf{\Phi}\mathbf{w}\|_2^2 + \sum_{i=1}^{G} \sqrt{a_i}\,\|\mathbf{w}_i\|_2\,, \tag{54}$$

which is equivalent to the l_1/l_2-norm formulation in (3). With the multivariate t-distributions, we have

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \beta\|\mathbf{y} - \mathbf{\Phi}\mathbf{w}\|_2^2 + \sum_{i=1}^{G} (d_i/2 - \lambda_i)\log\left(b_i + \|\mathbf{w}_i\|_2^2\right)\,. \tag{55}$$



(a) MAP (b) VB

Fig. 5. Thresholding functions for the McKay ($\lambda = 1$), multivariate Laplace, multivariate Student's t ($\lambda = -1$), and Jeffrey's. The dotted line is $\hat{w}_i = (\mathbf{\Phi}^T \mathbf{y})_i$.

Although the connection between this problem and the l_1/l_2-norm formulation in (3) is not immediately clear, they are in fact related. Consider the following l_p-norm based group-sparse estimation problem

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \; \beta \|\mathbf{y} - \mathbf{\Phi}\mathbf{w}\|_2^2 + \sum_{i=1}^{G} \tau \left(b_i + \|\mathbf{w}_i\|_2\right)^p \;, \tag{56}$$

with $0 < p \leq 2$. Notice that $p = 1$ recovers the l_1/l_2-norm minimization in (3). Using the formula

$$\lim_{p \to 0} \; \frac{1}{p} \left(b_i + \|\mathbf{w}_i\|_2 - 1\right)^p = \log\left(b_i + \|\mathbf{w}_i\|_2^2\right) \;, \tag{57}$$

it can be seen that the multivariate t prior is a limiting case of the l_p-norm based group-sparse estimation procedure. In addition, in the case of Jeffrey's priors, the penalty function is the limiting case of $\|\mathbf{w}_i\|_2^p$ as $p \to 0$. In this regard, the Laplace and t-distributions can be thought to be at the opposite ends of the l_p-norm penalties; while Laplace prior leads to an l_1-based method, t-distributions enforce sparsity similar to l_0-norms. The generalized l_p-norm based formulation with $0 < p < 1$ can be constructed using Gaussian variance mixtures as well, but the mixing distribution is an alpha-stable distribution without a closed-form, which makes the inference very hard.

Using the MAP formulation in (53), we can also analyze the thresholding properties of different distributions when $\mathbf{\Phi}$ is orthonormal, i.e., $\mathbf{\Phi}^T\mathbf{\Phi} = \mathbf{I}$. In this case, the problem decouples into G optimization problems (the groups become independent), and can be solved for each group separately as

$$\hat{\mathbf{w}}_i = \arg\min_{\mathbf{w}_i} \; -2\beta \, \mathbf{w}_i^T \left(\mathbf{\Phi}^T \, \mathbf{y}\right)_i + \mathbf{w}_i^T \mathbf{w}_i + \sum_{i=1}^{G} \log \mathrm{p}(\mathbf{w}_i | a_i, b_i, \lambda_i) \,. \tag{58}$$

The thresholding functions for different distributions for fixed $\mathbf{a}$, $\mathbf{b}$ and β are shown in Fig. 5(a). The multivariate Laplace distribution has a soft-thresholding behavior (similar to the univariate case), while the behavior of all other distributions is similar to hard-thresholding, including the McKay ($\lambda = 1$) distribution. In addition, the multivariate Laplace and McKay ($\lambda = 1$) priors have a constant bias independent of the signal value. Student's t and Jeffrey's priors do not have this disadvantage; the bias converges to zero as the signal magnitude increases. On the other hand, the Laplace prior is continuous at the thresholding value, whereas the others have discontinuities, which is generally considered as a disadvantage since small changes in the data might lead to large changes in the estimation [44].

In comparison, the thresholding functions obtained by the variational Bayesian inference described in the previous sections is shown in Fig. 5(b). It can be observed that all thresholding curves become smoother, and in fact, none of the priors lead to a thresholding rule: the estimates are only "almost" sparse, i.e., they have very small values in

an interval but are never exactly zero. Interestingly, the thresholding function of the Jeffrey's prior now exhibits a soft-thresholding behavior while the bias is again converging to zero as the signal magnitude increases. On the other hand, the thresholding property of the Laplace and McKay ($\lambda = 1$) is decreased. However, it should be emphasized that when $\mathbf{a}$, $\mathbf{b}$ and β are not constant but also estimated, all priors lead to exact thresholding rules.

Finally, an important remark is that simultaneous estimation of the parameters $\mathbf{a}$ and $\mathbf{b}$ cannot in general be performed using the MAP formulation if the hyperpriors $\mathrm{p}(\mathbf{a})$ and $\mathrm{p}(\mathbf{b})$ are not suitably chosen. The objective (53) becomes unbounded from below for some values of parameters k_a, θ_a, k_b, θ_b, in which case the global minimum is obtained at the trivial solution $\mathbf{w} = \mathbf{0}$, $\mathbf{a} \to \mathbf{0}$ and $\mathbf{b} \to \mathbf{0}$. Therefore, other methods should be employed, such as cross-validation or L-curves [27].

B. Hierarchical estimation

A second method is to use the hierarchical representations of the distributions, and consider the joint minimization problem as

$$\hat{\mathbf{w}}, \hat{\mathbf{z}} = \arg\max_{\mathbf{w},\mathbf{z}} \log \mathrm{p}(\mathbf{y}|\mathbf{\Phi}, \mathbf{w}) + \sum_{i=1}^{G} \log \mathrm{p}(\mathbf{w}_i|z_i) + \log \mathrm{p}(z_i|a_i, b_i, \lambda_i) \tag{59}$$

$$= \arg\min_{\mathbf{w},\mathbf{z}} \beta\|\mathbf{y} - \mathbf{\Phi}\mathbf{w}\|_2^2 - \log|\mathbf{\Lambda}| + \mathbf{w}^T \mathbf{\Lambda}\, \mathbf{w} - \sum_{i=1}^{G} 2 \log \mathrm{p}(z_i|a_i, b_i, \lambda_i)\,, \tag{60}$$

with $\mathbf{\Lambda} = \mathrm{diag}\left(z_i^{-1}\right)$. A common method for optimization is to consider an alternating iteratively reweighted minimization problem, where the estimation of $\mathbf{w}$ is done by holding $\mathbf{z}$ fixed and vice versa [24]. However, the distribution $\mathrm{p}(z_i|a_i, b_i, \lambda_i)$ and parameters k_a, θ_a, k_b, θ_b should be chosen carefully as some selections (e.g., $\lambda_i, a_i, b_i \to 0$) cause the objective to be unbounded at $\mathbf{w} = \mathbf{0}$ and $\mathbf{z} \to \mathbf{0}$, leading to the trivial solution. This problem is also observed in the Gaussian probabilistic matrix factorization with flat hyperpriors [45]: while the variational Bayesian inference allows the estimation of the hyperparameters, MAP estimation fails and gives the trivial solution.

One possible solution is to bound the objective by replacing $\mathbf{w}^T\mathbf{\Lambda}\mathbf{w}$ with $\mathbf{w}^T\mathbf{\Lambda}\mathbf{w} + \epsilon\,\mathbf{\Lambda}$ where ϵ is generally chosen in a heuristic manner [24], [26]. In this case, the signal estimate $\hat{\mathbf{w}}$ is the same as (37), while z_i^{-1} estimated using

$$\hat{z}_i^{-1} = \arg\min_{z_i^{-1}} z_i^{-1}(\|\mathbf{w}_i\|_2^2 + \epsilon) - d_i \log z_i^{-1} - 2\log \mathrm{p}(z_i|a_i, b_i, \lambda_i)\,. \tag{61}$$

One important difference between the MAP and Bayesian inference with this hierarchy can be observed by comparing (61) with the update rules in Section III-B: While MAP uses $\|\mathbf{w}_i\|_2^2 + \epsilon$ for updating z_i, the Bayesian method uses $\langle\|\mathbf{w}_i\|_2^2\rangle = \|\mathbf{w}_i\|_2^2 + \mathrm{trace}\,(\mathbf{\Sigma}_{\mathbf{w}_i})$. The last term makes the z_i parameters *a posteriori* dependent, while they are independent in the MAP approach. The Bayesian methodology provides a statistical interpretation of the term ϵ: it is the estimate of the posterior variance of the group i. In deterministic approaches, a decreasing sequence of ϵ is shown to provide better empirical performance. From the Bayesian perspective, this is also expected; the variance estimate generally decreases at each iteration with more accurate estimates of the signal. This connection is also observed in [46].

Note, however, calculation of $\mathrm{trace}\,(\mathbf{\Sigma}_{\mathbf{w}_i})$ significantly increases the computational complexity of Bayesian inference, since the inversion of either an $N \times N$ or $M \times M$ matrix is required using (38) and (39), respectively. The signal estimate in (37) does not require this inversion and has the same complexity as the MAP approach. The explicit calculation of $\Sigma_{\mathbf{w}}$ is prohibitive in high dimensional problems. A very simple and crude approximation, which surprisingly gives good results in some cases, is to only invert the diagonal elements of $\mathbf{\Sigma}_{\mathbf{w}}$ and calculate the trace terms. With this approximation, the computations reduce from $\mathrm{O}(N^3)$ (or $\mathrm{O}(M^3)$) to $\mathrm{O}(N)$. We evaluate this approximation in Section VI-D.

V. Extensions

In this section we discuss some extensions of the group-sparse modeling within the Bayesian framework, along with the resulting estimation schemes using variational inference.

A. Group-sparsity in multiple-measurements

Group-sparsity can also be used in the multiple measurement vector (MMV) problem. Here, the observations are expressed as

$$\mathbf{Y} = \mathbf{\Phi}\,\mathbf{W} + \mathbf{N}\,, \tag{62}$$

where each row of $\mathbf{W} \in \mathbb{R}^{N\times K}$ corresponds to K related variables with similar sparsity profile, and the groups are again defined over the columns. Matrix $\mathbf{N}$ represents the noise with independent zero-mean Gaussian variables as entries. To accommodate this generative model, we modify the mixture model as

$$\mathbf{W}_i = \sqrt{z_i}\,\mathbf{X}\,, \tag{63}$$

with matrix $\mathbf{W}_i$ is extracted from $\mathbf{W}$ using the rows contained in group i, and each column of $\mathbf{X}$ is a standard multivariate Gaussian variable.

The inference procedures presented so far can accommodate this modeling as well, with small changes in the updates. With some algebra, it is not hard to see that the posterior distribution of $\mathbf{W}$ becomes factorized with respect to its columns, and all columns have the same covariance matrices $\mathbf{\Sigma_W}$, such that

$$\mathrm{q}(\mathbf{W}) = \mathcal{N}(\langle\mathbf{W}\rangle, \mathbf{\Sigma_W})\,, \tag{64}$$

with parameters

$$\langle\mathbf{W}\rangle = \mathbf{\Sigma_W}\,\beta\,\mathbf{\Phi}^T\,\mathbf{Y}\,, \tag{65}$$

$$\mathbf{\Sigma}_\mathbf{W}^{-1} = \beta\,\mathbf{\Phi}^T\mathbf{\Phi} + \mathbf{\Lambda}\,, \tag{66}$$

with $\mathbf{\Lambda}$ defined as before in (38). The change affecting the posterior updates of the variances z_i is the use of the Frobenius norm $\|\mathbf{W}_i\|_F^2$ instead of $\|\mathbf{w}_i\|_2^2$, such that instead of (42) we have

$$\langle\|\mathbf{W}_i\|_F^2\rangle = \|\langle\mathbf{W}_i\rangle\|_F^2 + K\,\mathrm{trace}\,(\mathbf{\Sigma}_{\mathbf{W}_i})\,. \tag{67}$$

In addition, d_i is replaced with Kd_i in all updates of the parameter z_i.

B. Within-group correlations

The framework considered until now correlated the coefficients within each group through the use of a single parameter only, as can be seen from (7). We can, however, embed additional correlation structure into the formulation, by the modification

$$\mathrm{p}(\mathbf{w}_i\,|z_i) = \mathcal{N}(\mathbf{0}_{d_i}, z_i\,\mathbf{C}_i^T\mathbf{C}_i)\,, \tag{68}$$

where $\mathbf{C}_i^T\mathbf{C}_i$ is the within-group covariance matrix, and $\mathbf{z}_i$ again is used to control the sparsity. The variance-mixtures are defined in this case as

$$\mathbf{w}_i = \sqrt{z_i}\,\mathbf{C}_i\mathbf{x}\,. \tag{69}$$

Hence, the signal $\mathbf{w}$ is a linear transformation of a multivariate Gaussian variable. Note that the matrix $\mathbf{C}_i$ represents an integral-type operator, which generates data $\mathbf{w}_i$ from white noise [47], [48]. This type of modeling, generally referred to as *analysis-based modeling* [49], is useful in modeling signals that are not sparse themselves but can be represented sparsely in some transform domain, with images as typical examples. In this case, the inverse covariance $\left(\mathbf{C}_i^T\mathbf{C}_i\right)^{-1} = \mathbf{D}_i^T\mathbf{D}_i$ is chosen as a high-pass operator. For instance, when a forward wavelet transform is applied to the image, the resulting wavelet coefficients contain a small number of significant groups, and the remaining majority of the coefficients have negligible magnitudes.

Only small changes are needed in the inference procedures to accommodate this change in the modeling. Specifically, the signal update becomes

$$\langle\mathbf{w}\rangle = \mathbf{\Sigma_w}\,\beta\,\mathbf{\Phi}^T\,\mathbf{y}\,, \tag{70}$$

$$\mathbf{\Sigma}_\mathbf{w}^{-1} = \beta\,\mathbf{\Phi}^T\mathbf{\Phi} + \mathbf{D}^T\,\mathbf{\Lambda}\,\mathbf{D}\,, \tag{71}$$

with $\mathbf{D}$ is a block-diagonal matrix with the $\mathbf{D}_i$-matrices on the diagonal, i.e., $\mathbf{D} = \mathrm{diag}(\mathbf{D}_i)$. All variance parameter update equations have the same form with $\|\mathbf{w}_i\|_2^2$ replaced with $\|\mathbf{D}_i\,\mathbf{w}_i\|_2^2$.

C. Overlapping Groups

In some problems, the group structure is designed to be overlapping [19], [50]. This property is desired in certain applications (especially in bioinformatics [50]), or in cases where the group structure is not known *a priori*, in which case overlapping groups might alleviate problems associated with wrong groupings. If only a few groups are overlapping, then a simple way to incorporate this in the modeling and inference is to explicitly duplicate the signal coefficients and columns of the dictionary $\mathbf{\Phi}$ that correspond to elements belonging to multiple groups [50]. However, this approach leads to increased computational requirements when many groups overlap.

The approach with expanding the signal dimensions by duplication can readily be handled with modeling and inference schemes presented so far. Here we consider the approach without duplication. We do not change the signal modeling and use the signal prior in (5). Notice, however, that the coefficients in multiple groups will have multiple variances z_i associated with them. Specifically, the factorized signal prior is given by

$$\mathrm{p}(\mathbf{w}|\mathbf{z}) = \prod_{i=1}^{G} \mathrm{p}(\mathbf{w}_i\,|z_i) = \prod_{k=1}^{N} \prod_{i\in\Omega_k} \mathrm{p}(w_k|z_i) \tag{72}$$

where Ω_k is the index set of groups the coefficient w_k belongs to. For coefficients that belong to multiple groups, the prior inverse variances will be added, e.g., if w_k belongs to groups i and j, the corresponding prior inverse variance is given by $z_i^{-1} + z_j^{-1}$. With this modeling, the only modification in the inference is in the construction of matrix $\mathbf{\Lambda}$ when estimating the signal $\mathbf{w}$. Specifically, we have

$$Z_{kk} = \sum_{i\in\Omega_k} \langle z_i^{-1} \rangle\,, \quad k = 1, \ldots, N. \tag{73}$$

It should be noted that in this formulation, overlapping groups will have an effect on each other during inference (due to the added inverse variances). Therefore, higher sparsity might be enforced on coefficients that belong to many groups. This effect does not exist with the duplication approach discussed above. Nevertheless, this scheme proved to be useful for estimation when the group structure is unknown (see the empirical results in Section VI).

VI. Empirical Evaluation

In this section, we present experimental results demonstrating the performance of different signal priors in group-sparse signal estimation problems. We focus on the multivariate signal priors McKay($\lambda = 1$), Student's t, Laplace and Jeffrey's. We examine the effect of group size and the selection of the groups, and demonstrate the utility of modeling with overlapping groups in problems where the group structure is unknown. Finally, we compare the performance of variational Bayesian inference with full covariance estimation and the approximation described in Section IV.

As a baseline comparison, we use the state-of-the-art l_1-norm basis pursuit method SPG [21], which is a deterministic optimization approach based on spectral projection. This method is very fast and provides high estimation performance, and hence is suitable for comparing the modeling and inference procedures described in this paper in terms of estimation accuracy and computational requirements.

The source code developed to obtain the results shown in this section is available online in `https://netfiles.uiuc.edu/dbabacan/www/software.html`.

A. Comparison of Signal Priors

To compare the estimation performance of different signal priors, we generated a collection of signals of length $N = 300$ including (i) a sparse signal with 60 coefficients Gaussian-distributed with variance 1, and the remaining coefficients zero, (ii) a non-sparse signal with 60 coefficients Gaussian-distributed with variance 1, and the remaining coefficients Gaussian distributed with variance 10^{-3}, (iii) Student's t and (iv) generalized double Pareto (GDP) distributed signals, which are considered to be compressible with appropriate parameter selections [39] (we use $\lambda = -1/2$ for the Student's t in (26) and $k_a = 3/4$ for the GDP in (24)). Example realizations of these signals are shown in Fig. 6. A variety of signal characteristics are captured with this collection. Only signal (i) is sparse, whereas signals (iii) and (iv) are compressible, and signal (ii) is neither sparse nor compressible. Although only signal (i) is exactly sparse, more coefficients in signals (iii) and (iv) are closer to zero, and therefore the compressibility

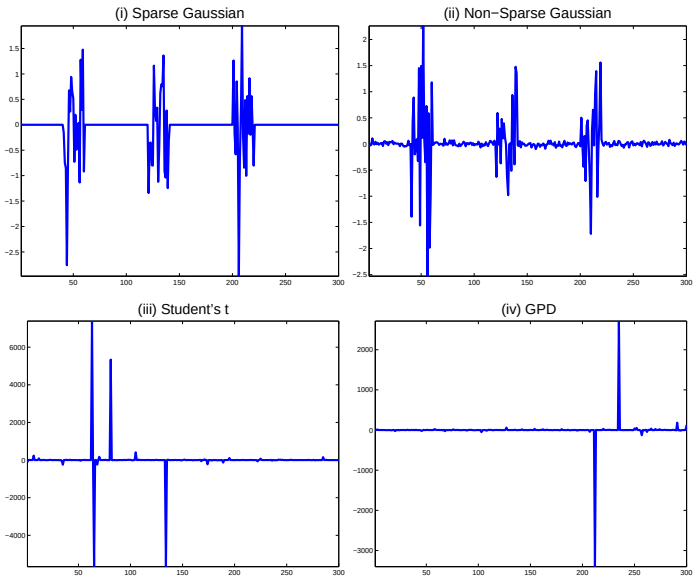


Fig. 6. Example realizations of different signals used in the experiments.

of these signals can be ordered as (iii) $\approx$ (iv) $>$ (i) $>$ (ii). In addition, the energy distribution within the signal coefficients is very different. In signals (i) and (ii), the magnitude difference between the important and spurious coefficients is small (generally less than an order of magnitude), whereas in signals (iii) and (iv) this difference can be very large (e.g., several orders of magnitude).

We fix the group size to 20 ($d_i = 20\ \forall i$, number of groups $G = 15$) and consider two strategies for grouping the signal coefficients: (1) random grouping and (2) ordered grouping where the coefficients of the original vector $\mathbf{w}$ are sorted according to their magnitudes, and groups are created by dividing the N sorted indices into G clusters. Note that this corresponds to an "oracle" grouping with respect to the coefficient magnitudes. With signal (i), in both strategies we generate the groups such that they either contain all non-zero coefficients or all zero coefficients (but without magnitude ordering). A completely random grouping results in significant loss in performance and will be discussed later.

The $M \times N$ matrix $\mathbf{\Phi}$ is generated by drawing its entries from a standard Gaussian distribution and normalizing the columns to have unit l_2-norm. White Gaussian noise with variance 10^{-6} is added to obtain measurements $\mathbf{y}$. The unknown vector $\mathbf{w}$ is estimated using the variational Bayesian methods with multivariate McKay($\lambda = 1$), Laplace, Student's t ($\lambda = -1$) and Jeffrey's priors. The hyperparameters k_a, k_b, θ_a and θ_b are set equal to 10^{-5} to obtain broad hyperpriors on the parameters a_i and b_i. The noise variance β^{-1} is estimated along with the unknown vector with all methods. The SPG method does not provide means to estimate this parameter, so the true noise variance is given to this method.

To measure the reconstruction performance, we use the relative reconstruction error $\|\hat{\mathbf{w}} - \mathbf{w}\|_2 / \|\mathbf{w}\|_2$ where $\hat{\mathbf{w}}$ is the estimated signal and $\mathbf{w}$ is the true signal, respectively. The convergence criterion is $\|\hat{\mathbf{w}}^n - \hat{\mathbf{w}}^{n-1}\|_2 / \|\hat{\mathbf{w}}^{n-1}\|_2 < 10^{-10}$ where n is the iteration number. The experiments are repeated 100 times with different realizations of matrices $\mathbf{\Phi}$, noise and signals $\mathbf{w}$. Average estimation results comparing the signal priors with different signals and varying M/N ratios are depicted in Fig. 7. The results with random grouping are denoted by (R) and the ones with ordered grouping are denoted by (O).

Several observations can be made from Fig. 7: First, the proposed method outperforms the deterministic approach SPG in all test cases with all priors, while the performance difference varies depending on the underlying signal. The performance difference is especially prominent with the sparse signal (i), where the proposed methods achieve reconstruction errors close to 10^{-3} with as low as $0.3M$ measurements, while SPGL1 requires more than $0.7M$ measurements to obtain this error level. With the other signals the performance difference is also clear, in some cases getting close to an order of magnitude.

Second, all priors provide good signal estimates with the highly compressible signals (i), (iii) and (iv) at even low measurement levels. McKay($\lambda = 1$), Student's t and Jeffrey's priors result in more accurate estimation compared to Laplace at all measurement levels with the sparse signal (i). However, the performance of the priors is close with

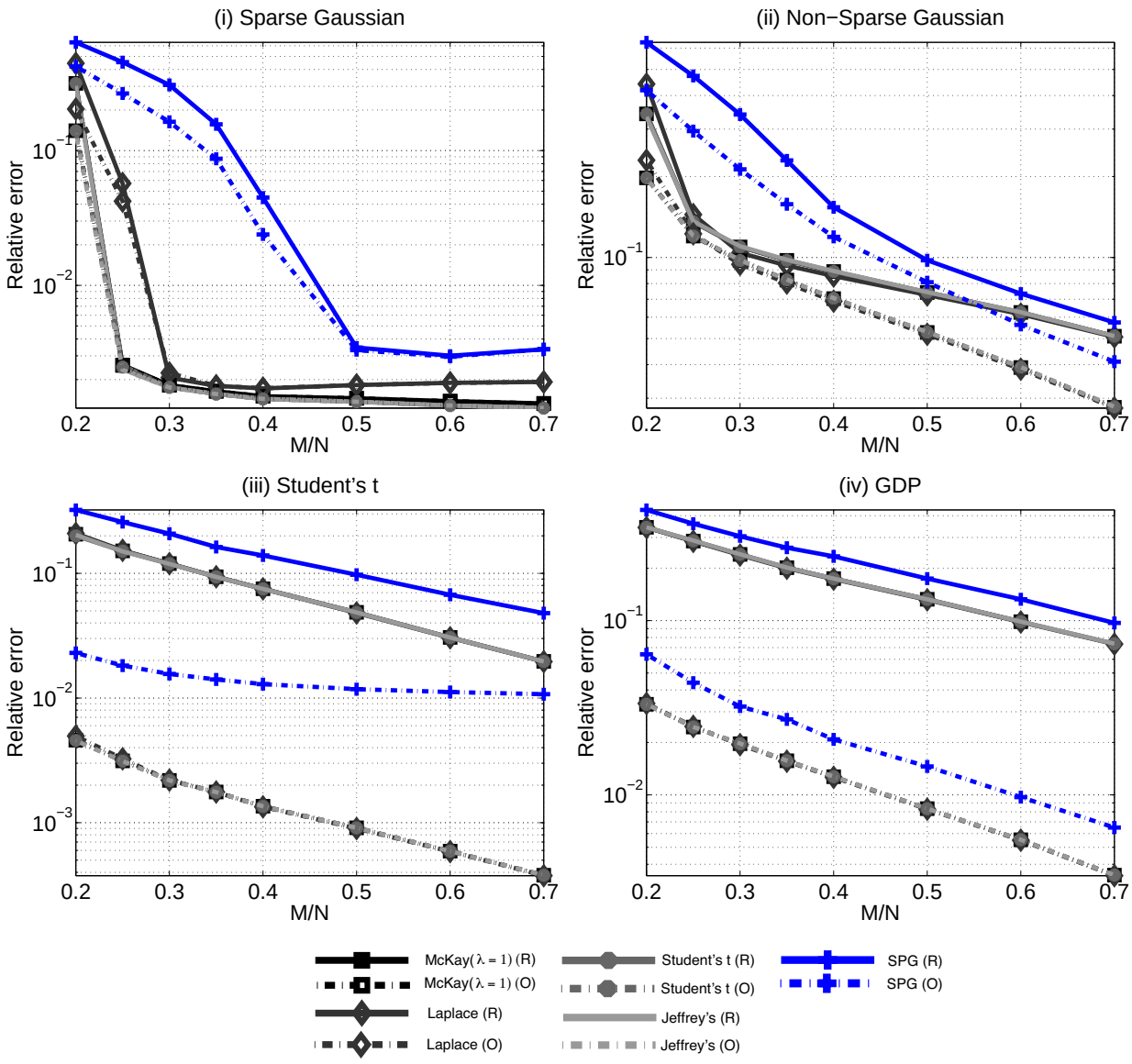


Fig. 7. Comparison of estimation performances of signal priors with different signals. Signal types are denoted at the top of each figure, and the legend is common to all figures. (R) denotes random grouping (solid lines), and (O) denotes ordered grouping (dashed lines). Note that the curves corresponding to Mackay ($\lambda = 1$), Student's t and Jeffrey's priors almost coincide.

the other signals. Especially Student's t and Jeffrey's priors give very similar results, and we empirically observed that many b_i parameters of Student's t are driven to very small values during iterations making the distribution similar to Jeffrey's prior.

The performance of all priors is much lower with the non-sparse signal (ii). Interestingly, the Student's t and Jeffrey's priors again provide very good results even though they enforce sparsity to an higher extent. It can be argued that these priors are very effective in selecting the most important coefficients even with non-sparse signals where the difference between the important and unimportant coefficients is not high.

Finally, it is clear that the grouping strategy makes a significant difference in estimation performance. Grouping coefficients with high magnitude differences results in severe degradations in estimation performance. The degradation in performance is not as severe with signal (i), where in both cases we classified the groups as zero/nonzero. The result of completely random grouping is shown in Fig. 10. An interesting observation is that grouping via magnitude ordering is not as important as identifying the nonzero coefficients when the magnitude differences within the signal is not large (such as signal (i)).

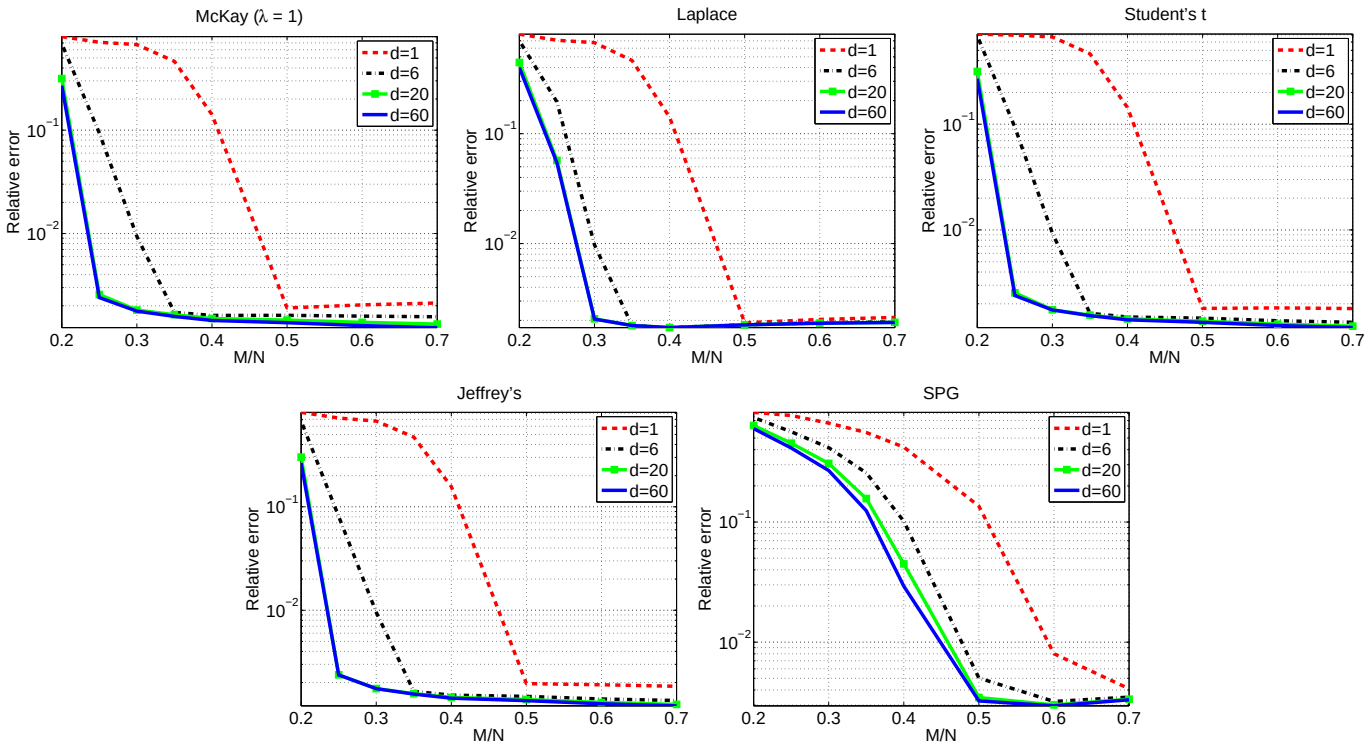


Fig. 8. Effect of group size on the estimation performance with different signal priors.

B. Effect of group size

It is evident that the strategy of selecting the groups has a significant effect on the estimation performance. Here we additionally investigate the effect of the group size on the estimation performance. We report experiments with the sparse signal (i) in the previous section to precisely control the sparsity level of the signal, since these signals contain exactly 60 nonzero coefficients. We vary the group sizes as $d_i = 1, 2, 6, 20, 60$. The groups are selected randomly but groups contain either all non-zero or all zero coefficients. Note that the case with $d_i = 1$ corresponds to standard sparse reconstruction without groups.

Simulation results with different M/N ratios are shown in Fig. 8. It is clear that with all priors, grouping the coefficients result in significant gains in estimation even when the grouping is done randomly (without the information on the ordering of their magnitudes). While all priors have similar and high performance, the Laplace prior is generally slightly inferior compared to others. Finally, as in the previous experiments, the proposed method outperforms the SPG method independent of the selected signal prior: the proposed method typically requires at least $0.2M$ less measurements to obtain the reconstruction errors provided by SPGL1, independent of the group size.

Overall, based on the experimental results, it can be observed that all signal priors approximately provide the same estimation performance. Due to additional complexity in the estimation rules with the McKay($\lambda = 1$), Student's t and Laplace priors, the Jeffrey's prior is favorable as the corresponding estimation procedure does not involve complex special functions and thus is much simpler. However, note that in this work we only consider signal reconstruction; other priors might prove useful in applications where the goal is data interpretation instead of reconstruction.

C. Overlapping groups

The group size and selection is critical in estimation performance, as demonstrated in the previous sections. However, neither of them are known *a priori* in general practical settings without additional structural information of the unknown signal $\mathbf{w}$. In this section, we demonstrate the utility of the modeling with overlapping groups (Section V-C) in cases where no information is available about the signal structure.

We again experiment with the sparse signals (i) with 60 nonzero coefficients and 240 zero coefficients. The nonzero coefficients are chosen uniformly at random and drawn from a standard Gaussian distribution. We consider

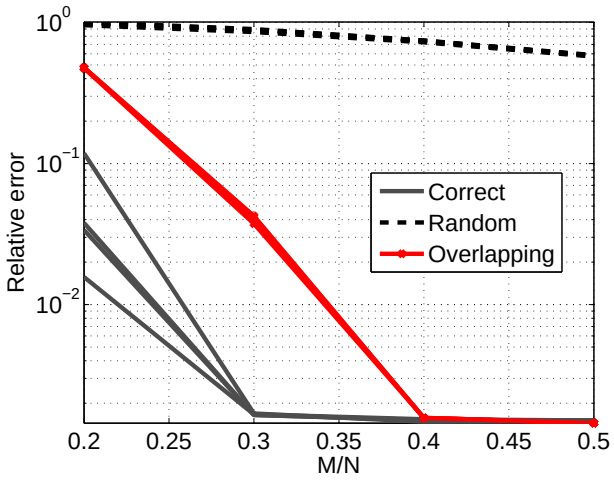


Fig. 9. Comparison of correct, random and overlapping groups. Different distributions with the same grouping are plotted with the same shape curves.

three scenarios with a fixed group size of $d_i = 20$: 1) "Correct" grouping where the nonzero coefficient locations are known and groups either contain all nonzero or all zero coefficients, 2) groups are constructed randomly, and 3) overlapping groups, where the signal is divided sequentially into $d_i = 20$ coefficient groups with 50% overlap (i.e., group 1 contains coefficients 1 to 20, group 2 contains coefficients 10 to 30, and so on). Estimation results with these strategies (average of 100) are shown in Fig. 9. It is clear that modeling with overlapping groups results in significant improvement compared to random grouping. Random construction of the groups does not provide good estimates, whereas the difference between the overlapping and correct grouping is not too large. Overlapping grouping therefore can be used for instance to first estimate the group structure which can then be used in an additional inference step for improved performance.

D. Effect of the Covariance Approximation

As mentioned earlier, one disadvantage of the Bayesian methods presented in this paper is the need to compute the covariance matrix $\mathbf{\Sigma_w}$, which is computationally intensive and makes the inference not scalable to problems with high dimensional data. On the other hand, the approximation to the covariance matrix described in Section IV significantly reduces the computational load and provides a more efficient inference procedure. In this section, we evaluate the effect of this approximation in terms of estimation accuracy and speed.

Similar to the previous sections, we generate sparse signals with nonzero coefficients drawn from a standard Gaussian distribution. The signal size is chosen as 500, the number of nonzero coefficients are set to 100, and the group size is fixed to 20 where the non-zero group locations are assumed to be known. We use (39) to compute the full covariance matrices as $M < N$. Fig. 10 compares the estimation performance and the corresponding running times with and without the covariance approximation. While the methods with full covariance matrices have significantly lower estimation error (especially at low M/N ratios), the running times are drastically increasing with increasing M (approximately in the order of M^3). On the other hand, while the estimation performance is significantly decreased, the running times of the methods with covariance approximation are approximately constant for all M levels, indicating that they are scalable to high-dimensional problems. While not investigated in this paper, a possible method to achieve both high estimation performance and computational efficiency is to divide the problem into inner- and outer-loops to reduce the number of updates of the covariance matrix [51]. Finally, it is evident that even with the covariance approximation the proposed methods still provide comparable or better performance than SPG with approximately same running times.

VII. Conclusions

In this paper, we presented a general multivariate signal prior construction suitable for group-sparse modeling. Using the normal-variance mixture hierarchy, we have shown that this signal model includes multivariate versions of a number of signal models commonly used in the literature for sparse signal modeling. Therefore, this construction

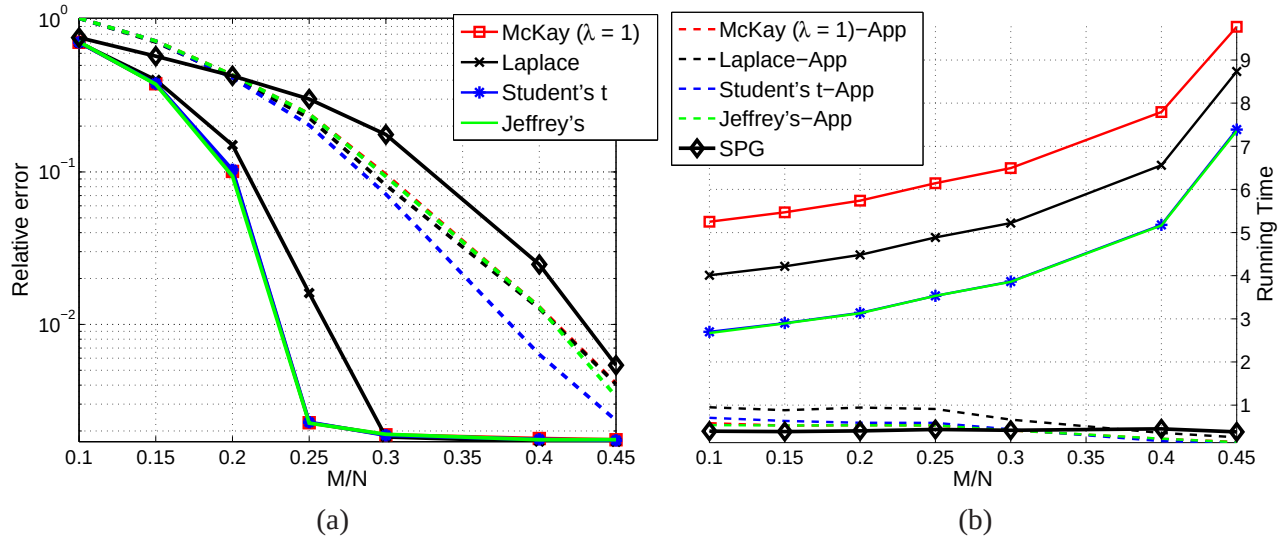


Fig. 10. Comparison of variational Bayesian approaches with different distributions with the full and approximate covariance matrices (denoted by "App"). (a) Reconstruction errors, (b) running times. The legends are common to both figures.

is useful to compare their characteristics and analyze their connections. We provided estimation rules with these priors using variational Bayesian inference and empirically demonstrated their estimation performance. Experimental results suggest that the proposed formulation is very powerful and provides better estimation performance than state-of-the-art deterministic approaches. In addition, we showed that while all priors provide very similar performances, Jeffrey's prior is an attractive choice due to its high estimation performance and simple update rules. We also provided and evaluated a simple approximation for scalable inference in large-scale problems. Finally, we have discussed some extensions of group-sparse modeling within the Bayesian methodology and have shown that the proposed method is very flexible and can easily be used for a wide range of problems involving group-sparse modeling.

REFERENCES

[1] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1999.

[2] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[3] Y. Eldar and M. Mishali, "Robust recovery of signals from a structured union of subspaces," *IEEE Transactions on Information Theory*, vol. 55, no. 11, pp. 5302–5316, 2009.

[4] M. Stojnic, F. Parvaresh, and B. Hassibi, "On the reconstruction of block-sparse signals with an optimal number of measurements," *IEEE Transactions on Signal Processing*, vol. 57, no. 8, pp. 3075–3085, 2009.

[5] M. Stojnic, "l2/l1-optimization in block-sparse compressed sensing and its strong thresholds," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 2, pp. 350–357, 2010.

[6] J. Huang and T. Zhang, "The benefit of group sparsity," *Arxiv preprint arXiv:0901.2962*, 2009.

[7] Y. Eldar, P. Kuppinger, and H. Bölcskei, "Block-sparse signals: Uncertainty relations and efficient recovery," *IEEE Transactions on Signal Processing*, vol. 58, no. 6, pp. 3042–3054, June 2010.

[8] P. J. Garrigues and B. A. Olshausen, "Group sparse coding with a Laplacian scale mixture prior," in *Advances in Neural Information Processing Systems 23*. MIT Press, 2010.

[9] B. M. Marlin, M. Schmidt, and K. P. Murphy, "Group sparse priors for covariance estimation," in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, ser. UAI '09. Arlington, Virginia, United States: AUAI Press, 2009, pp. 383–392.

[10] L. Meier, S. Van De Geer, and P. Bhlmann, "The group lasso for logistic regression," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 70, no. 1, pp. 53–71, 2008.

[11] J. Friedman, T. Hastie, and R. Tibshirani, "A note on the group lasso and a sparse group lasso," *Arxiv preprint arXiv:1001.0736*, 2010.

[12] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[13] Y. Lu and M. Do, "A theory for sampling signals from a union of subspaces," *IEEE Transactions on Signal Processing*, vol. 56, no. 6, pp. 2334–2345, 2008.

[14] ——, "Sampling signals from a union of subspaces," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 41–47, 2008.

[15] T. Blumensath and M. Davies, "Sampling theorems for signals from the union of finite-dimensional linear subspaces," *IEEE Transactions on Information Theory*, vol. 55, no. 4, pp. 1872–1882, 2009.

[16] J. Huang, X. Huang, and D. N. Metaxas, "Learning with dynamic group sparsity," in *ICCV*, 2009, pp. 64–71.

[17] J. Baritaux, K. Hassler, M. Bucher, S. Sanyal, and M. Unser, "Sparsity-driven reconstruction for FDOT with anatomical priors," *IEEE Transactions on Medical Imaging*, vol. 30, no. 5, pp. 1143–1153, 2011.

[18] A. K. Bolstad, B. D. V. Veen, and R. Nowak, "Causal network inference via group sparse regularization," *IEEE Transactions on Signal Processing*, vol. 59, no. 6, pp. 2628–2641, 2011.

[19] S. Mosci, S. Villa, A. Verri, and L. Rosasco, "A primal-dual algorithm for group sparse regularization with overlapping groups," in *Advances in Neural Information Processing Systems 23*. MIT Press, 2010.

[20] X. Chen, Q. Lin, S. Kim, J. Peña, J. G. Carbonell, and E. P. Xing, "An efficient proximal-gradient method for single and multi-task regression with structured sparsity," *CoRR*, vol. abs/1005.4717, 2010.

[21] E. van den Berg and Friedlander, "Sparse optimization with least-squares constraints," *SIAM J. on Optimization*, vol. 21, no. 4, pp. 1201–1229, 2010.

[22] S. Raman, T. J. Fuchs, P. J. Wild, E. Dahl, and V. Roth, "The Bayesian group-lasso for analyzing contingency tables," in *ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009, pp. 881–888.

[23] A. Lee, F. Caron, A. Doucet, and C. Holmes, "A hierarchical Bayesian framework for constructing sparsity-inducing priors," *ArXiv:1009.1914*, Sept. 2010.

[24] I. Daubechies, R. DeVore, M. Fornasier, and C. S. Güntürk, "Iteratively reweighted least squares minimization for sparse recovery," *Comm. Pure Appl. Math.*, vol. 63, no. 1, pp. 1–38, 2010.

[25] E. J. Candes, M. Wakin, and S. Boyd, "Enhancing sparsity by reweighted l1 minimization," *J. Fourier Anal. Appl.*, no. 14, pp. 877–905, 2007.

[26] D. Wipf and S. Nagarajan, "Iterative reweighted l1 and l2 methods for finding sparse solutions," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 2, pp. 317 –329, april 2010.

[27] P. C. Hansen and D. P. O'Leary, "The use of the l-curve in the regularization of discrete ill-posed problems," *SIAM J. Sci. Comput.*, vol. 14, pp. 1487–1503, November 1993.

[28] O. Barndorff-Nielsen, J. Kent, and M. Sørensen, "Normal variance-mean mixtures and z distributions," *International Statistical Review/Revue Internationale de Statistique*, pp. 145–159, 1982.

[29] D. F. Andrews and C. L. Mallows, "Scale mixtures of normal distributions," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 36, no. 1, pp. 99–102, 1974.

[30] M. West, "On scale mixtures of normal distributions," *Biometrika*, vol. 74, no. 1, pp. 646–648, 1987.

[31] B. Jørgensen, *Statistical Properties of the Generalized Inverse Gaussian Distribution*, ser. Lecture Notes in Statistics. 9. Springer-Verlag., 1982.

[32] A. T. McKay, "A Bessel function distribution," *Biometrika*, vol. 24, no. 1/2, pp. 39–44, May 1932.

[33] S. Kotz, T. Kozubowski, and K. Podgorski, *The Laplace Distribution and Generalizations*. Birkhauser, 2001.

[34] J. Palmer, K. Kreutz-Delgado, B. Rao, and S. Makeig, "Modeling and estimation of dependent subspaces with non-radially symmetric and skewed densities," *Independent Component Analysis and Signal Separation*, pp. 97–104, 2007.

[35] D. B. Madan and E. Seneta, "The variance gamma (v.g.) model for share market returns," *The Journal of Business*, vol. 63, no. 4, pp. 511–24, October 1990.

[36] T. Eltoft, T. Kim, and T.-W. Lee, "Multivariate scale mixture of Gaussians modeling," in *Independent Component Analysis and Blind Signal Separation*, ser. Lecture Notes in Computer Science, J. Rosca, D. Erdogmus, J. Prncipe, and S. Haykin, Eds. Springer Berlin / Heidelberg, 2006, vol. 3889, pp. 799–806.

[37] A. P. Doulgeris and T. Eltoft, "Scale mixture of Gaussian modelling of polarimetric SAR data," *EURASIP J. Adv. Signal Process*, vol. 2010, pp. 2:1–2:12, January 2010.

[38] T. Eltoft, T. Kim, and T. Lee, "On the multivariate Laplace distribution," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 300–303, 2006.

[39] V. Cevher, "Learning with compressible priors," in *Advances in Neural Information Processing Systems*. MIT Press, 2009, vol. 22, pp. 261–269.

[40] A. Armagan, D. B. Dunson, and J. Lee, "Generalized double Pareto shrinkage," *Arxiv preprint arXiv:1104:0861v2*, 2011.

[41] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*. New York, Springer Verlag, 1985.

[42] C. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[43] D. G. Tzikas, A. C. Likas, and N. P. Galatsanos, "The variational approximation for bayesian inference," *IEEE Signal Processing Magazine*, vol. 25, pp. 131–146, Nov 2008.

[44] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

[45] S. Nakajima and M. Sugiyama, "Theoretical analysis of Bayesian matrix factorization," *Journal of Machine Learning Research*, vol. 12, pp. 2583–2648, 2011.

[46] D. P. Wipf and S. Nagarajan, "Sparse estimation using general likelihoods and non-factorial priors," in *Advances in Neural Information Processing Systems 22*. MIT Press, 2009, vol. 22.

[47] M. Unser, P. Tafti, and Q. Sun, "A unified formulation of Gaussian vs. sparse stochastic processes - part I: Continuous-domain theory," *arXiv:1108.6150v1*, 2011.

[48] M. Unser, P. Tafti, A. Amini, and Q. Sun, "A unified formulation of Gaussian vs. sparse stochastic processes - part II: Discrete-domain theory," *arXiv:1108.6150v1*, 2011.

[49] M. Elad, P. Milanfar, and R. Rubinstein, "Analysis versus synthesis in signal priors," *Inverse Problems*, vol. 23, no. 3, pp. 947–968, June 2007.

[50] L. Jacob, G. Obozinski, and J.-P. Vert, "Group lasso with overlap and graph lasso," in *Proceedings of the 26th Annual International Conference on Machine Learning*, ser. ICML '09. New York, NY, USA: ACM, 2009, pp. 433–440.

[51] M. W. Seeger and H. Nickisch, "Large scale variational inference and experimental design for sparse generalized linear models," *SIAM Journal on Imaging Sciences*, vol. 4, no. 1, pp. 166–199, 2011.