Hierarchical Infinite Divisibility for Multiscale Shrinkage

Xin Yuan, Vinayak Rao, Shaobo Han and Lawrence Carin
Department of Electrical and Computer Engineering
Duke University
Durham, NC 27708-0291 USA

***Abstract*—A new shrinkage-based construction is developed for a compressible vector $\boldsymbol{x} \in \mathbb{R}^n$, for cases in which the components of $\boldsymbol{x}$ are naturally associated with a tree structure. Important examples are when $\boldsymbol{x}$ corresponds to the coefficients of a wavelet or block-DCT representation of data. The method we consider in detail, and for which numerical results are presented, is based on the gamma distribution. The gamma distribution is a heavy-tailed distribution that is infinitely divisible, and these characteristics are leveraged within the model. We further demonstrate that the general framework is appropriate for many other types of infinitely-divisible heavy-tailed distributions. Bayesian inference is carried out by approximating the posterior with samples from an MCMC algorithm, as well as by constructing a variational approximation to the posterior. We also consider expectation-maximization (EM) for a MAP (point) solution. State-of-the-art results are manifested for compressive sensing and denoising applications, the latter with spiky (non-Gaussian) noise.**

I. Introduction

In signal processing, machine learning and statistics, a key aspect of modeling concerns the compact representation of data. Consider an n-dimensional signal $\boldsymbol{T}\boldsymbol{x}$, where $\boldsymbol{T} \in \mathbb{R}^{n\times n}$ and $\boldsymbol{x} \in \mathbb{R}^n$; the columns of $\boldsymbol{T}$ define a basis for representation of the signal, and $\boldsymbol{x}$ are the basis coefficients. Two important means of constituting the columns of $\boldsymbol{T}$ are via orthornormal wavelets or the block discrete cosine transform (DCT) [1]. The former is used in the JPEG2000 compression standard, and the latter in the JPEG standard; compression is manifested because many of the components of $\boldsymbol{x}$ may be discarded without significant impact on the accuracy of the reconstructed signal [2], [3]. Wavelet and block-DCT-based compression are applicable to a wide class of natural data. For such data, the original $\boldsymbol{x}$ typically has no values that are exactly zero, but many that are relatively small, and hence in this case $\boldsymbol{x}$ is termed "compressible."

In compression, one typically acquires the uncompressed data, sends it through filters associated with $\boldsymbol{T}$, and then performs compression on the resulting $\boldsymbol{x}$. In inverse problems, one is not given $\boldsymbol{x}$, and the goal is to recover it from given data. For example, in compressive sensing (CS) [4]–[6] one is typically given m linear projections of the data, $\boldsymbol{y} = \boldsymbol{H}\boldsymbol{T}\boldsymbol{x} + \boldsymbol{n}$, where $\boldsymbol{H} \in \mathbb{R}^{m\times n}$ and $\boldsymbol{n} \in \mathbb{R}^m$ is measurement noise (ideally $m \ll n$). The goal is to recover $\boldsymbol{x}$ from $\boldsymbol{y}$. In denoising applications [7] $\boldsymbol{H}$ may be the $n \times n$ identity matrix, and the goal is to recover $\boldsymbol{x}$ in the presence of noise $\boldsymbol{n}$, where the noise may be non-Gaussian. If one is performing recovery of missing pixels [8] (the "inpainting" problem), then $\boldsymbol{H}$ is defined by m rows of the $n \times n$ identity matrix. Other related problems include deblurring, where $\boldsymbol{H}$ may be a blur kernel [9].

To solve these types of problems, it is important to impart prior knowledge about $\boldsymbol{x}$, where compressibility is widely employed. As a surrogate for compressibility, there is much work on assuming that $\boldsymbol{x}$ is exactly sparse [4]–[6]; the residuals associated with discarding the small components of $\boldsymbol{x}$ are absorbed in $\boldsymbol{n}$. This problem has been considered from many perspectives, using methods from optimization [10]–[12] as well as Bayesian approaches [13]–[15]. In this paper we take a Bayesian perspective. However, it is demonstrated that many of the shrinkage priors we develop from that perspective have a corresponding regularizer from an optimization standpoint, particularly when considering a maximum *a posterior* (MAP) solution.

In Bayesian analysis researchers have developed methods like the relevance vector machine (RVM) [16], which has been applied successfully in CS [13]. Spike-slab priors have also been considered, in which exact sparsity is imposed [14]. Methods have also been developed to leverage covariates ($e.g.$, the spatial locations of pixels in an image [8]), using methods based on kernel extensions. Inference has been performed efficiently using techniques like Markov chain Monte Carlo (MCMC) [8], variational Bayes [17], belief propagation [15], and expectation propagation [18]. Expectation-maximization (EM) has been considered [19] from the MAP perspective.

Recently there has been significant interest on placing "global-local shrinkage" priors [20], [21] on the components of $\boldsymbol{x}$, the ith of which is denoted x_i. Each of these priors imposes the general construction $x_i \sim \mathcal{N}(0, \tau^{-1}\alpha_i^{-1})$, with a prior on α_i^{-1} highly concentrated at the origin, with heavy tails. The set of "local" parameters $\{\alpha_i\}$, via the heavy tailed prior, dictate which components of $\boldsymbol{x}$ have significant amplitude. Parameter τ is a "global" parameter, that scales $\{\alpha_i\}$ such that they are of the appropriate amplitude to represent the observed data. Recent examples of such priors include the horseshoe [22] and three-parameter-beta priors [23].

In this line of work the key to imposition of compressibility is the heavy-tailed prior on the $\{x_i\}$. An interesting connection has been made recently between such a prior and increments of general Lévy processes [21]. It has been demonstrated that nearly all of the above shrinkage priors may be viewed in

terms of increments of a Lévy process, with appropriate Lévy measure. Further, this same Lévy perspective may be used to manifest the widely employed spike-slab prior. Hence, the Lévy perspective is a valuable framework from which to view almost all priors placed on $\boldsymbol{x}$. Moreover, from the perspective of optimization, many regularizers on $\boldsymbol{x}$ have close linkages to the Lévy process [24].

All of the shrinkage priors discussed above assume that the elements of $\boldsymbol{x}$ are exchangeable; $i.e.$, the elements of $\{\alpha_i\}$ are drawn i.i.d. from a heavy-tailed distribution. However, the components of $\boldsymbol{x}$ are often characterized by a tree structure, particularly in the context of a wavelet [2], [14] or block-DCT [3], [17] representation of $\boldsymbol{T}$. It has been recognized that when solving an inverse problem for $\boldsymbol{x}$, leveraging this known tree structure may often manifest significantly improved results [25].

The principal contribution of this paper concerns the extension of the aforementioned shrinkage priors into a new framework, that exploits the known tree structure. We primarily focus on one class of such shrinkage priors, which has close connections from an optimization perspective to adaptive Lasso [26]. However, we also demonstrate that this specific framework is readily generalized to all of the shrinkage priors that may be represented from the Lévy perspective. This paper therefore extends the very broad class of Lévy-based shrinkage priors, leveraging the known tree structure associated with $\boldsymbol{T}$. There is a close connection between Lévy processes and infinitely-divisible random variables, which we discuss. The perspective of infinitely-divisible random variables is emphasized for simplicity, as the continuous-time Lévy process is not needed for the actual implementation.

We principally take a Bayesian perspective, and perform computations using MCMC and variational Bayesian analysis. However, we also perform a MAP-based point estimation, which demonstrates that this general tree-based structure may also be extended to an optimization perspective. Specific applications considered are estimation of $\boldsymbol{x}$ in the context of CS and denoising. For the latter we consider non-Gaussian noise, specifically noise characterized by the sum of Gaussian and spiky components. This is related to the recent interest in robust principal component analysis [27]. State-of-the-art results are realized for these problems.

The remainder of the paper is organized as follows. The basic setup of the inverse problems considered is detailed in Section II. In that section we also introduce the specific class of global-local priors for which example results are computed. In Section III we discuss the method by which we extend such a shrinkage prior to respect the tree structure of $\boldsymbol{x}$, when associated with expansions $\boldsymbol{T}$ tied to wavelet, block-DCT and other related bases. It is demonstrated in Section IV how the above model is a special case of a general class of tree-based models, based on infinitely-divisible random variables. In Section V we provide a detailed discussion of how the current work is related to the literature, and in Section VI we discuss how spiky noise is modeled. Three different means of performing inference are discussed in Section VII, with further details in the Appendix. An extensive set of numerical results and comparisons to other methods are presented in Section VIII, with conclusions discussed in Section IX.

II. Tree-Based Representation

A. Measurement model

Consider the wavelet transform of a two-dimensional (2D) image $\boldsymbol{F} \in \mathbb{R}^{n_x \times n_y}$; the image pixels are vectorized to $\boldsymbol{f} \in \mathbb{R}^n$, where $n = n_x n_y$. Under a wavelet basis $\boldsymbol{T} \in \mathbb{R}^{n \times n}$, $\boldsymbol{f}$ can be expressed as $\boldsymbol{f} = \boldsymbol{T}\boldsymbol{x}$, where the columns of $\boldsymbol{T}$ represent the orthonormal wavelet basis vectors, and $\boldsymbol{x}$ represents the wavelet coefficients of $\boldsymbol{f}$. The same type of tree-based model may be applied to a *block* discrete cosine transform (DCT) representation [3], [17]. In Section VIII experimental results are shown based on both a wavelet and block-DCT tree-based decomposition. However, for clarity of exposition, when presenting the model, we provide details in the context of the wavelet transform.

The measurement model for $\boldsymbol{f}$ is assumed manifested in terms of m linear projections, defined by the rows of $\boldsymbol{H} \in \mathbb{R}^{m \times n}$. Assuming additive measurement noise $\boldsymbol{n}$, we have:

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{f} + \boldsymbol{n} = \boldsymbol{\Psi}\boldsymbol{x} + \boldsymbol{n} \tag{1}$$

where $\boldsymbol{\Psi} = \boldsymbol{H}\boldsymbol{T}$. In compressive sensing (CS) [4] $m \ll n$, and it is desirable that the rows of $\boldsymbol{H}$ are incoherent with the columns of $\boldsymbol{T}$, such that $\boldsymbol{\Psi}$ is a dense matrix [4], [5], [28]. If one directly observes the image pixels, which we will consider in the context of image-denoising applications, $\boldsymbol{H} = \boldsymbol{I}$, with $\boldsymbol{I}$ symbolizing the identity matrix.

Assuming i.i.d. Gaussian noise with precision α_0, we have

$$\boldsymbol{y}|\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\Psi}\boldsymbol{x}, \alpha_0^{-1}\boldsymbol{I}). \tag{2}$$

The assumption of i.i.d. Gaussian noise may be inappropriate in some settings, and in Section VI we revisit the noise model.

B. Modeling the wavelet scaling coefficients

In most wavelet decompositions of a 2D image, there are three sets of trees[1]: one associated with high-high (HH) wavelet coefficients, one for high-low (HL) wavelet coefficients, and one for low-high (LH) coefficients [1]. In addition to the HH, HL and LH wavelet coefficients, there is a low-low (LL) scaling coefficient subimage, which does *not* have an associated tree; the LL scaling coefficients constitute a low-frequency representation of the original image.

The scaling coefficients are in general *not* sparse (since $\boldsymbol{f}$ is not sparse in general), and hence the scaling coefficients are modeled as

$$x_{0,i}|\tau_0, \alpha_0 \sim \mathcal{N}(0, \tau_0^{-1}\alpha_0^{-1}), \tag{3}$$

where τ_0 controls the "global" scale of the scaling coefficients, and α_0 is the noise precision as above (τ_0 represents the *relative* precision of the scaling coefficients, with respect to the noise precision α_0). Non-informative gamma priors are placed on τ_0 and α_0, $i.e.$, Ga$(10^{-6}, 10^{-6})$. This is a typical default

[1]There are classes of wavelets, for example polyharmonic wavelets [29], that do not use three wavelet trees, but rather only a single tree. This special class of wavelets is not considered here

setting for such hyperparameters [16], and in the context of all experiments, no model tuning has been performed.

C. Compressiblility of wavelet coefficients

Let $x_{\ell,i}$ represent coefficient i at level ℓ in the wavelet tree (for either a HH, HL, or LH tree). To make connections to existing shrinkage priors and regularizers (*e.g.*, adaptive Lasso [26]), we consider a generalization of the double-exponential (Laplace) prior [30], [31]. In Section IV we discuss that this model corresponds to one example of a broad class of shrinkage priors based on infinitely-divisible random variables, to which the model may be generalized.

Coefficient $x_{\ell,i}$ is assumed to be drawn from the distribution

$$\begin{aligned} & \frac{1}{2}\sqrt{\frac{\tau_\ell}{\gamma_{\ell,i}}}\exp(-|x_{\ell,i}|\sqrt{\frac{\tau_\ell}{\gamma_{\ell,i}}}) \\ = & \int \mathcal{N}(x_{\ell,i};0,\tau_\ell^{-1}\alpha^{-1})\mathrm{InvGa}(\alpha;1,(2\gamma_{\ell,i})^{-1})d\alpha \quad (4) \end{aligned}$$

where $\mathrm{InvGa}(\alpha;1,(2\gamma_{\ell,i})^{-1})$ is the inverse-gamma distribution for α, with parameters 1 and $(2\gamma_{\ell,i})^{-1}$. Parameter $\tau_\ell > 0$ is a "global" scaling for all wavelet coefficients at layer ℓ, and $\gamma_{\ell,i}$ is a "local" weight for wavelet coefficient i at that level. We place a gamma prior $\mathrm{Ga}(a,b)$ on the $\gamma_{\ell,i}$, and as discussed below one may set the hyperparameters (a,b) to impose that most $\gamma_{\ell,i}$ are small, which encourages that most $x_{\ell,i}$ are small. As far as inference for a point estimate for the model parameters by maximizing the log posterior is concerned, observe that the log of the prior in (4) corresponds to adaptive Lasso regularization [26]. Adaptive Lasso is distinguished from the original Lasso by employing coefficient-dependent Laplace parameters $\gamma_{\ell,i}$; although not originally viewed from a Bayesian standpoint, in the original Lasso [32] $\gamma_{\ell,i}$ is the same for all i.

The model (4) for wavelet coefficient $x_{\ell,i}$ may be represented in the hierarchical form

$$\begin{aligned} x_{\ell,i}|\tau_\ell,\alpha_{\ell,i} &\sim \mathcal{N}(0,\tau_\ell^{-1}\alpha_{\ell,i}^{-1}) \\ \alpha_{\ell,i}|\gamma_{\ell,i} &\sim \mathrm{InvGa}(1,(2\gamma_{\ell,i})^{-1}) \quad (5) \\ \gamma_{\ell,i} &\sim \mathrm{Ga}(a,b), \end{aligned}$$

introducing latent variables $\{\alpha_{\ell,i}\}$, rather than marginalizing them out as in (4); a vague/diffuse gamma prior is again placed on the scaling parameters τ_ℓ. While we have introduced many new latent variables $\{\alpha_{\ell,i}\}$, the form in (5) is convenient for computation, and with appropriate choice of (a,b) most $\{\alpha_{\ell,i}\}$ are encouraged to be large. The large $\alpha_{\ell,i}$ correspond to small $x_{\ell,i}$, and therefore this model imposes that most $x_{\ell,i}$ are small, *i.e.*, it imposes compressibility. In Section III we concentrate on the model for $\{\gamma_{\ell,i}\}$.

D. Continuous shrinkage priors vs. spike-slab priors

There are models of coefficients like $\boldsymbol{x}$ that impose exact sparsity [14], [33], [34], while the model in (5) imposes compressibility (many coefficients that are small, but not exactly zero). Note that if a component of $\boldsymbol{x}$ is small, but is set exactly to zero via a sparsity prior/regularizer, then the residual will be attributed to the additive noise $\boldsymbol{n}$. This implies that the sparsity-promoting priors will tend to over-shrink, under-estimating components of $\boldsymbol{x}$ and over-estimating the noise variance. This partially explains why we have consistently found that compressibility-promoting priors like (5) often are more effective in practice on real (natural) data than exact-sparsity-promoting models, like the spike-slab setup in [14]. These comparisons are performed in detail in Section VIII. In Section V-A we make further connections shrinkage priors developed in the literature.

III. Hierarchical Tree-Based Shrinkage

A. Infinite divisibility of gamma random variables

The gamma prior on $\gamma_{\ell,i}$ plays a key role in imposing the "local" shrinkage on coefficients $x_{\ell,i}$. We here briefly review properties of the gamma distribution, which are exploited in the hierarchical, tree-based shrinkage introduced below. Further, in Section IV we discuss that the gamma distribution is a special case of general heavy-tailed infinitely divisible random variables, which provide alternatives to the gamma distribution for shrinkage-based priors.

For any integer $n \geq 1$, assume $X_i \sim \mathrm{Ga}(a/n,b)$, for $i = 1,\ldots,n$. The sum of random variables $\sum_{i=1}^n X_i$ is equal in distribution to $X \sim \mathrm{Ga}(a,b)$. Random variables X drawn from $\mathrm{Ga}(a,b)$ are termed "infinitely divisible," because they may be constituted as $X = \sum_{i=1}^n X_i$, where X_i are i.i.d. random variables (here drawn from $\mathrm{Ga}(a/n,b)$), and $n > 1$ is an integer. The gamma distribution is one example from which infinitely divisible random variables are generated, the general properties of which are discussed in Section IV-A. In the context of the aforementioned gamma distribution, by taking the limit $n \to \infty$, it also follows that if we assume $X_i \sim \mathrm{Ga}(T_i a,b)$, with each $T_i \in (0,1)$ and with $\sum_{i=1}^n T_i = 1$, then $\sum_{i=1}^n X_i$ is equal in distribution to $X \sim \mathrm{Ga}(a,b)$.

We consider $a = b = 1$, and again $T_i \in (0,1)$ with $\sum_{i=1}^n T_i = 1$, for $n \geq 1$. For $X_i \sim \mathrm{Ga}(T_i,1)$, since $\sum_{i=1}^n X_i$ is equal in distribution to $X \sim \mathrm{Ga}(1,1)$, we have $\sum_{i=1}^n \mathbb{E}(X_i) = 1$, where the expectation $\mathbb{E}(X_i) = T_i$. Therefore, the size of T_i dictates the expected contribution of X_i to the expected sum, which equals one.

In what follows, we will assume $\gamma_{\ell,i} \sim \mathrm{Ga}(T_i^{(\ell-1)},1)$, where $T_i^{(\ell-1)} > 0$ and $\sum_i T_i^{(\ell-1)} = 1$; $T_i^{(\ell-1)}$ will be linked to *parent* wavelet coefficients situated at layer $\ell-1$, and the characteristics of these parent coefficients impacts the properties of the children coefficients at layer ℓ. From (4), large/small $\gamma_{\ell,i}$ encourage large/small $x_{\ell,i}$. Since $\mathbb{E}(\gamma_{\ell,i}) = T_i^{(\ell-1)}$, the size of $T_i^{(\ell-1)}$ controls the degree of shrinkage in $x_{\ell,i}$. In the model developed below, if a wavelet coefficient at layer $\ell-1$ is large/small, then the $T_i^{(\ell-1)}$ of its children coefficients are encouraged to be large/small. Via the "budget" $\sum_i T_i^{(\ell-1)} = 1$, only a small number of $T_i^{(\ell-1)}$ can be relatively large, thereby encouraging compressibility in the wavelet coefficients. The key observation is that the size of $T_i^{(\ell-1)}$ controls the expected value of $\gamma_{\ell,i}$, this motivating encoding the wavelet tree structure in the multiscale construction of $\{T_i^{(\ell-1)}\}$.

B. Hierarchical gamma shrinkage

If we simply impose the prior $\text{Ga}(a, b)$ on $\{\gamma_{\ell,i}\}$, we do not exploit the known tree structure of wavelets and what that tree imposes on the characteristics of the multi-layer coefficients [1], [35]. A wavelet decomposition of a 2D image yields a quadtree hierarchy ($n_c = 4$ children coefficients under each parent coefficient), as shown in Figure 1. A key contribution of this paper is the development of a general shrinkage-based decomposition that respects the tree structure.

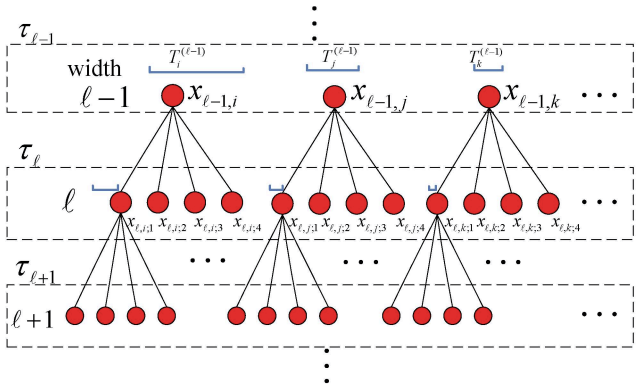

Fig. 1: Depiction of the layers of the trees, with $n_c = 4$ children considered. Note that each parent coefficient at layer $\ell - 1$ has a relative width $T_i^{(\ell-1)}$, and all widths at a given layer sum to one. The width of the parent is partitioned into windows of width $T_i^{(\ell-1)}/n_c$ for the children.

Let $\boldsymbol{x}$ represent all the coefficients in a wavelet decomposition of a given image, where $\boldsymbol{x}_\ell$ represents all coefficients across all trees at level ℓ in the wavelet hierarchy. We here consider all the wavelet trees for one particular class of wavelet coefficients (*i.e.*, HH, HL or LH), and the model below is applied independently as a prior for each. Level $\ell = 0$ is associated with the scaling coefficients, and $\ell = 1, 2, \ldots$ correspond to levels of wavelet coefficients, with $\ell = 1$ the root-node level. We develop a prior that imposes shrinkage *within* each layer ℓ, while also imposing structure *between* consecutive layers. Specifically, we impose the *persistence* property [35] of a wavelet decomposition of natural imagery: if wavelet coefficient i at level $\ell - 1$, denoted $x_{\ell-1,i}$, is large/small, then its children coefficients $\{x_{\ell,i;j}\}_{j=1,n_c}$ are also likely to be large/small.

A parameter $\gamma_{\ell,i}$ is associated with each node i at each level ℓ in the tree (see Figure 1). Let $\{\gamma_{\ell-1,i}\}$ represent the *set* of parameters for all coefficients at layer $\ell - 1$. The parameters $\{\gamma_{\ell,i;j}\}_{j=1,n_c}$ represent the n_c *children* parameters under *parent* parameter $\gamma_{\ell-1,i}$.

If there are n_1 wavelet trees associated with a given 2D image, there are n_1 root nodes at the top layer $\ell = 1$. For each of the coefficients associated with the root nodes, corresponding shrinkage parameters are drawn

$$\gamma_{1,i} \sim \text{Ga}(1/n_1, b) \tag{6}$$

which for large n_1 encourages that most $\gamma_{1,i}$ will be small, with a few outliers; this is a "heavy-tailed" distribution for appropriate selection of b, where here we set $b = 1$ (the promotion of compressibility is made more explicit in Section IV). From the discussion above, $\sum_i \mathbb{E}(\gamma_{1,i}) = 1$.

Assuming $n_{\ell-1}$ nodes at layer $\ell-1$, we define an "increment length" $T_i^{(\ell-1)}$ for node i at level $\ell - 1$:

$$T_i^{(\ell-1)} = \gamma_{\ell-1,i} / \sum_{i'=1}^{n_{\ell-1}} \gamma_{\ell-1,i'} \tag{7}$$

To constitute each child parameter $\gamma_{\ell,i;j}$, for $j = 1, \ldots, n_c$, we independently draw

$$\gamma_{\ell,i;j} \sim \text{Ga}(T_i^{(\ell-1)}/n_c, 1) \tag{8}$$

Because of the properties of the gamma distribution, if $T_i^{(\ell-1)}$ is *relatively* large (*i.e.*, $\gamma_{\ell-1,i}$ is relatively large), the children coefficients are also encouraged to be relatively large, since the increment $T_i^{(\ell-1)}/n_c$ is relatively large; the converse is true if $T_i^{(\ell-1)}$ is small. For each ℓ, we have $\sum_j \sum_i \mathbb{E}(\gamma_{\ell,i;j}) = 1$.

C. Dirichlet simplification

In (5), the parameter τ_ℓ provides a scaling of all precisions at layer ℓ of the wavelet tree. Of most interest to define which wavelet coefficients have large values are the *relative* weights of $\alpha_{\ell,i}$, with the relatively small elements corresponding to wavelet coefficients with large values. Therefore, for computational convenience, we consider scaled versions of $\gamma_{\ell,i}$, with the relatively large $\gamma_{\ell,i}$ associated with the relatively small-valued $\alpha_{\ell,i}$. Specifically, consider the normalized vector

$$\tilde{\boldsymbol{\gamma}}_{\ell,i} = \boldsymbol{\gamma}_{\ell,i} / \sum_{i'} \gamma_{\ell,i'}, \tag{9}$$

which corresponds to a draw from a Dirichlet distribution:

$$\tilde{\boldsymbol{\gamma}}_\ell \sim \text{Dir}(\beta_1^{(\ell)}, \ldots, \beta_{n_\ell}^{(\ell)}) \tag{10}$$

where $\beta_i^{(\ell)} = T_{pa(\ell,i)}^{(\ell-1)}/n_c$, and $pa(\ell, i)$ is the parent of the ith coefficient at layer ℓ. The last equation in (5) is replaced by (10), and all other aspects of the model remain unchanged; the $\tilde{\boldsymbol{\gamma}}_\ell$ are *normalized* parameters, with the scaling of the parameters handled by τ_ℓ at layer ℓ. One advantage of this normalization computationally is that all elements of $\tilde{\boldsymbol{\gamma}}_\ell$ are updated as a block, rather than one at a time, as implied by the last equation in (5).

The model for parameters $\tilde{\boldsymbol{\gamma}}_\ell$ at each level of the tree may be summarized as follows. At the root $\ell = 1$ layer

$$\tilde{\boldsymbol{\gamma}}_1 \sim \text{Dir}(1/n_1, \ldots, 1/n_1) \tag{11}$$

If $\tilde{\boldsymbol{\gamma}}_{\ell-1}$ are parameters at level $\ell-1$, then the child parameters follow (10), with $\beta_i^{(\ell)} = T_{pa(\ell,i)}^{(\ell-1)}/n_c = \tilde{\gamma}_{\ell-1,pa(\ell,i)}$.

D. Discussion

We define $\gamma_{\ell,i} = \sum_{j=1}^{n_c} \gamma_{\ell,i;j}$. Based upon the properties of the gamma distribution, discussed in Section III-A, $\gamma_{\ell,i} \sim \text{Ga}(T_i^{(\ell-1)}, 1)$, while for $\ell = 1$ we have $T_i^{(0)} = 1/n_1$ for all $i = 1, \ldots, n_1$. Therefore, at each level ℓ, $\sum_{i=1}^{n_i} \mathbb{E}(\gamma_{\ell,i}) = 1$, and $\mathbb{E}(\gamma_{\ell,i}) = T_i^{(\ell-1)}$. The *relatively* large components of $\{T_i^{(\ell-1)}\}$ are defined by which parents $\gamma_{\ell-1,i}$ are *relatively*

large. Via the tree-based construction discussed above, persistence between scales is imposed: if a parent coefficient is relatively large/small at a given layer, then its children coefficients are encouraged to be as well.

IV. Infinite Divisibility and Model Generalization

A. Lévy-Khintchine theorem

The Lévy-Khintchine theorem [36] specifies the form of the characteristic equation for any random variable G that is infinitely divisible. Specifically, a random variable G is infinitely divisible if and only if its characteristic equation satisfies

$$\mathbb{E}(e^{iuG}) = \exp[\{i\delta u - \tfrac{1}{2}u^2\Sigma + \textstyle\int_{\mathbb{R}-\{0\}}[e^{iuy} - 1 - iuy1(|y| \leq 1)]\nu(dy)\}], \quad (12)$$

where ν is a Lévy measure satisfying $\int_{\mathbb{R}-\{0\}}(|y|^2 \wedge 1)\nu(dy) < \infty$. The terms $\delta \in \mathbb{R}$ and $\Sigma \in \mathbb{R}_+$ correspond to, respectively, the mean and variance of a Gaussian random variable. The term associated with Lévy measure ν corresponds to a compound Poisson distribution with rate $\nu(dy)$. Any infinitely divisible random variable may be expressed as $G = G^{(N)} + G^{(CP)}$, where $G^{(N)} \sim \mathcal{N}(\delta, \Sigma)$ and $G^{(CP)}$ is associated with the compound Poisson distribution.

Now consider $T_i \in (0, 1)$ for $i = 1, \ldots, n$, with $\sum_{i=1}^n T_i = 1$. We may define an infinitely divisible random variable G_i associated with each T_i, the characteristic equation for which satisfies

$$\mathbb{E}(e^{iuG_i}) = \exp[T_i\{i\delta u - \tfrac{1}{2}u^2\Sigma + \textstyle\int_{\mathbb{R}-\{0\}}[e^{iuy} - 1 - iuy1(|y| \leq 1)]\nu(dy)\}], \quad (13)$$

From (12), $G = \sum_{i=1}^n G_i$ is also infinitely divisible. Each G_i may be expressed as $G_i = G_i^{(B)} + G_i^{(CP)}$, where $G_i^{(N)} \sim \mathcal{N}(\delta T_i, \Sigma T_i)$, and $G_i^{(CP)}$ corresponds to a compound Poisson distribution with rate $T_i\nu$. From above, each $\{G_i\}$ and G are based on the same underlying (δ, Σ, ν); G corresponds to an increment length of $T = 1$, while G_i is based on length T_i, with $\sum_{i=1}^n T_i = 1$. As first discussed in [20] by choosing particular settings of δ, Σ and ν, the vector $\{G_i\}$ may be used to impose sparsity or compressibility.

The expression (13) corresponds to the mass under a continuous-time (stochastic) Lévy process of time length (increment) T_i, and [20] emphasizes the Lévy-process perspective. Since we are not interested in the continuous-time Lévy-process itself, and only in the associated integrated mass, the infinitesimally-divisible random variable perspective emphasized above is sufficient for our purposes. However, this relationship underscores the close connections between infinitesimally divisible random variables and Lévy processes [37]–[39].

Assume that we wish to impose that the components of $\boldsymbol{x} \in \mathbb{R}^n$ are compressible. Then the ith component x_i may be assumed drawn $x_i \sim \mathcal{N}(0, \tau G_i)$, where τ is a "global" scaling, independent of i. Since only a small subset of the $\{G_i\}_{i=1,n}$ are *relatively* large for an appropriate infinitely divisible random variable (with $\delta = \Sigma = 0$, and heavy-tailed ν [20]), most components x_i are encouraged to be small. This global scaling τ and local G_i looks very much like the class of models in (5), which we now make more precise.

B. Tree-based model with general infinitely divisible random variables

At the root layer $\ell = 1$, with n_1 wavelet coefficients, we define $T_i^{(1)} = 1/n_1$ for all $i \in \{1, \ldots, n_1\}$. From (13) we define random variables $G_{1,i}$, $i = 1, \ldots, n_1$. With a heavy-tailed infinitely divisible random variable, and large n_1, most $G_{1,i}$ will be relatively small, imposing compressibility on the wavelet trees.

Assume access to $G_{\ell-1,i}$, $i = 1, \ldots, n_{\ell-1}$ at layer $\ell - 1$. The lengths of the increments at layer ℓ are defined analogous to (7), specifically

$$T_i^{(\ell-1)} = |G_{\ell,i}| / \sum_{i'}^{n_\ell - 1} |G_{\ell,i'}| \quad (14)$$

We use the absolute value in $|G_{\ell,i}|$ because, depending on the choice of (δ, Σ, ν), $G_{\ell,i}$ may be negative. The n_c children of coefficient i at layer $\ell - 1$ have increment widths $T_i^{(\ell-1)}/n_c$. Using each of these increment lengths, which sum to one, we draw associated $G_{\ell,i;j}$, with characteristic function in (13). Therefore, if $G_{\ell,i}$ is relatively large/small compared to the other coefficients at layer ℓ, then the increments of its children coefficients, $T_i^{(\ell-1)}/n_c$, are also relatively large/small. Since the expected relative strength of the random variable is tied to the width of the associated increment, this construction has the property of imposing a persistence in the strength of the wavelet coefficients across levels ℓ (relatively large/small parent coefficients enourage relative large/small children coefficients).

C. Shrinkage and sparsity imposed by different Lévy families

Each choice of parameters (δ, Σ, ν) yields a particular hierarchy of random variables $\{G_{\ell,i}\}$. In the context of imposing compressibility or sparsity, we typically are interested in $\delta = \Sigma = 0$, and different choices of ν yield particular forms of sparsity or shrinkage. For example, consider

$$\nu(dy) = y^{-1}\exp(-yb)dy \quad (15)$$

defined for $y > 0$. The associated random variable is $G_{\ell,i} \sim \text{Ga}(T_i^{(\ell-1)}, b)$. Therefore, the model considered in Section III is a special case of the Lévy measure in (15), which corresponds to a gamma process. The random variables $G_{\ell,i}$ correspond to the $\gamma_{\ell,i}$ in (5).

Note the singularity in (15) at $y = 0$, and also the heavy tails for large y. Hence, the gamma process imposes that most $G_{\ell,i}$ will be very small, since the gamma-process is concentrated around $y = 0$, but the heavy tails will yield some large $G_{\ell,i}$; this is why the increments of appropriate Lévy processes have close connections to compressibility-inducing priors [21].

There are many choices of Lévy measures one may consider; the reader is referred to [20], [21], [40] for a discussion

of these different Lévy measures and how they impose different forms of sparsity/compressibility. We have found that the gamma-process construction for $\{\gamma_{\ell,i}\}$, which corresponds to Section III, works well in practice for a wide range of problems, and therefore it is the principal focus of this paper.

We briefly note another class of Lévy measures that is simple, and has close connections to related models, discussed in Section V. Specifically, consider $\nu(dy) = \nu^+ H(dy)$, where ν^+ is a positive constant and $H(dy)$ is a probability density function, *i.e.*, $\int H(dy) = 1$. This corresponds to a compound Poisson distribution in (12), assuming $\delta = \Sigma = 0$. In this case the random variables can be positive and negative, if H admits random variables on the real line (*e.g.*, if H corresponds to a Gaussian distribution). Rather than using this to model $\gamma_{\ell,i}$, which must be positive, the $G_{\ell,i}$ are now used to directly model the wavelet coefficients $x_{\ell,i}$. We demonstrate in Section V-B that this class of constructions has close connections to existing models in the literature for modeling wavelet coefficients.

V. Connections to Previous Work

A. Other forms of shrinkage

Continuous shrinkage priors, like (5) and its hierarchical extensions developed here, enjoy advantages over conventional discrete mixture [41], [42] (*e.g.*, spike-slab) priors. First, shrinkage may be more appropriate than selection in many scenarios (selection is manifested when exact sparsity is imposed). For example, wavelet coefficients of natural images are in general p-compressible [43], with many small but non-zero coefficients. Second, continuous shrinkage avoids several computational issues that may arise in the discrete mixture priors (which impose explicit sparsity) [44], such as a combinatorial search over a large model space, and high-dimensional marginal likelihood calculation. Shrinkage priors often can be characterized as Gaussian scale mixtures (GSM), which naturally lead to simple block-wise Gibbs sampling updates with better mixing and convergence rates in MCMC [20], [23]. Third, continuous shrinkage reveals close connections between Bayesian methodologies and frequentist regularization procedures. For example, in Section II-C we noted the connection of the proposed continuous-shrinkage model to adaptive Lasso. Additionally, the iteratively reweighted ℓ_2 [45] and ℓ_1 [46] minimization schemes could also be derived from the EM updating rules in GSM and Laplace scale mixture (LSM) models [19], [47], respectively.

Recently, aiming to better mimic the marginal behavior of discrete mixture priors, [20], [21] present a new global-local (GL) family of GSM priors,

$$x_i|\tau,\lambda_i \sim \mathcal{N}(0,\tau\lambda_i), \quad \lambda_i \sim f, \quad \tau \sim g, \tag{16}$$

where f and g are the prior distributions of λ_i and τ; as noted above, such priors have been an inspiration for our model. The local shrinkage parameter λ_i and global shrinkage parameter τ are able to offer sufficient flexibility in high-dimensional settings. There exist many options for the priors on the Gaussian scales λ_i, for example in the three parameter beta normal $\mathcal{TPBN}(a,b,\tau)$ case [23],

$$x_i \sim \mathcal{N}(0,\tau\lambda_i), \lambda_i \sim \mathrm{Ga}(a,\gamma_i), \gamma_i \sim \mathrm{Ga}(b,1) \tag{17}$$

The horseshoe prior [22] is a special case of $\mathcal{TPBN}$ with $a = 1/2$, $b = 1/2$, $\tau = 1$.

GSM further extends to the LSM [47] by adding one hierarchy in (16): $\lambda_i \sim \mathrm{Ga}(1,\gamma_i)$, $\gamma_i \sim f$, which captures higher order dependences. As a special case of LSM, the normal exponential gamma (NEG) prior [48] provides a Bayesian analog of adaptive lasso [26]. The Dirichlet Laplace (DL) prior employed here can be interpreted as a non-factorial LSM prior which resembles point mass mixture prior in a joint sense, with frequentist optimality properties studied in [49].

None of the above shrinkage priors take into account the multiscale nature of many compressible coefficients, like those associated with wavelets and the block DCT. This is a key contribution of this paper.

B. Other forms of imposing tree structure

There has been recent work exploiting the wavelet tree structure within the context of compressive sensing (CS) [14], [17], [50]. In these models a two-state Markov tree is constituted, with one state corresponding to large-amplitude wavelet coefficients and the other to small coefficients. If a parent node is in a large-amplitude state, the Markov process is designed to encourage the children coefficients to also be in this state; a similar state-transition property is imposed for the small-amplitude state.

The tree-based model in [14] is connected to the construction in Section IV, with a compound Poisson Lévy measure. Specifically, consider the case for which $\delta = \Sigma = 0$, and ν is a compound Poisson process with $\nu = \nu_\ell^+ \mathcal{N}(0, \zeta_\ell^{-1}\boldsymbol{I})$, where ν_ℓ^+ a positive real number and ζ_ℓ a precision that depends on the wavelet layer ℓ. Priors may be placed on ν_ℓ^+ and ζ_ℓ. At layer ℓ, $J_\ell \sim$ Pois(ν_ℓ^+) "jumps" are drawn, and for each a random variable is drawn from $\mathcal{N}(0, \zeta_\ell^{-1}\boldsymbol{I})$; the jumps may be viewed as being placed uniformly at random between [0,1], defining $G_{\ell,i}$ (the T_i, with $\sum_i T_i = 1$ correspond to increments in the [0,1] window). Along the lines discussed in Section IV, at the root layer of the tree, each of the n_1 wavelet coefficients is assigned a distinct increment of length $1/n_1$. As discussed at the end of Section IV-C, here the $G_{\ell,i}$ are used to model the wavelet coefficients directly. If one or more of the J_1 aforementioned jumps falls within the window of coefficient i at layer $\ell = 1$ (*i.e.*, within window T_i), then the coefficient is non-zero, and is a Gaussian distributed; if no jump falls within width i, then the associated coefficient is exactly zero. If a coefficient is zero, then all of its decendent wavelet coefficients are zero, using the construction in Section IV. This may be viewed as a two-state model for the wavelet coefficients: each coefficient is either exactly zero, or has a value drawn from a Gaussian distribution. This is precisely the type of two-state model developed in [14], but viewed here from the perspective of a tree-based construction tied to infinitely divisible random variables (or, alternatively, but equivalently, Lévy processes), and particularly a compound Poisson process. In [14] a compound Poisson model was not

used, and rather a Markov setup was employed, but the basic concept is the same as above.

In [50] the authors developed a method similar to [14], but in [50] the negligible wavelet coeffecients were not shrunk exactly to zero. Such a model can also be viewed from the perspective of infinitely divisible random variables, ν again corresponding to a compound Poisson process. However, now $\delta = 0$ and $\Sigma \neq 0$. For an increment for which there are no jumps, the wavelet coefficient is only characterized by the Gaussian term (Brownian motion from the perspective of the Lévy process), and Σ is associated with small but non-zero wavelet coefficients.

This underscores that the proposed tree-based shrinkage prior, which employs the gamma process, may be placed within the context of the infinitely-divisible framework of Section IV, as can [14], [50], which use state-based models. This provides a significant generalization of [20], [21], which first demonstrated that many shrinkage and sparsity priors may be placed within the context of the increments of a Lévy process, but did not account for the tree-based structure of many representations, such as wavelets and the block DCT. Here we also emphasize that the perspective of infinitely divisible random variables is sufficient, without the need for the explicit Lévy processes (but, as discussed, the two are intimately related).

C. Connection to dictionary learning

In Section VIII we present experimental results on denoising images, and we make comparisons to an advanced dictionary-learning approach [8]. It is of interest to note the close connection of the proposed wavelet representation and a dictionary-learning approach. Within the wavelet decomposition, with n_1 root nodes, we have n_1 trees, and therefore the overall construction may be viewed as a "forest." Each tree has L layers, for an L-layer wavelet decomposition, and each tree characterizes a subregion of the image/data. Therefore, our proposed model can be viewed as dictionary learning, in which the dictionary building blocks are wavelets, and the tree-based shrinkage imposes form on the weights of the dictionary elements within one tree (corresponding to an image patch or subregion). Within the proposed gamma process construction, the "local" $\{\gamma_{\ell,i}\}$ account for statistical relationships between coefficients at layer ℓ, the parent-child encoding accounts for relationships between consecutive layers, and the "global" τ_ℓ also accounts for cross-tree statistical relationships.

VI. Generalized Noise Model

In (2) the additive noise was assumed to be i.i.d. zero-mean Gaussian with precision α_0. In many scenarios the assumption of purely Gaussian noise is inappropriate, and the noise may be represented more accurately as a sum of a Gaussian and spiky term, and in this case we modify the model as

$$\boldsymbol{y} = \boldsymbol{\Psi x} + \boldsymbol{w} + \boldsymbol{n}, \tag{18}$$

where the spiky noise $\boldsymbol{w}$ is also modeled by a shrinkage prior:

$$\begin{aligned} w_i|\mu,\zeta_i,\alpha_0 &\sim \mathcal{N}(0,\mu^{-1}\zeta_i^{-1}\alpha_0^{-1}),\\ \zeta_i|p_i &\sim \mathrm{InvGa}(1,(2p_i)^{-1}),\\ \boldsymbol{p} &\sim \mathrm{Dir}(1/n,\ldots,1/n),\\ \mu &\sim \mathrm{Ga}(e_0,f_0). \end{aligned} \tag{19}$$

The spiky noise $\boldsymbol{w}$ is characterized by a normal-inverse gamma-gamma hierarchy, as in (5), where again the Dirichlet distribution represents draws from a *normalized* gamma distribitition. Therefore, the prior on $\boldsymbol{w}$ is analogous to the compressible prior placed on the wavelet coefficients, with most components of $\boldsymbol{w}$ negligible, with a small number of spiky outliers. The Gaussian term is still modeled as $\boldsymbol{n}|\alpha_0 \sim \mathcal{N}(0,\alpha_0^{-1}\boldsymbol{I})$.

In this case the noise corresponds to infinitely-divisible random variables, with increments (now in two dimensions) defined by the size of the pixels: the term $\boldsymbol{n}$ corresponds to the Gaussian (Brownian) term, with $\boldsymbol{w}$ manifested by the compound Poisson term (with Lévy measure ν again corresponding to a gamma process). Interestingly, infinitely-divisible random variables are now playing two roles: (i) to model in a novel hierarchical manner the tree structure in the coefficients $\boldsymbol{x}$, and (ii) to separately model the Gaussian plus spiky noise $\boldsymbol{w}+\boldsymbol{n}$; the particular Lévy triplet (δ,Σ,ν) need not be the same for these two components of the model.

VII. Inference

We developed an MCMC algorithm for posterior inference, as well as a deterministic variational algorithm to approximate the posterior. The latter is easily adapted to develop an Expectation-Maximization (EM) algorithm to obtain MAP point estimates of the wavelet coefficients, and recalls frequentist algorithms like adaptive Lasso [26] and re-weighted ℓ_1 [46]. Details on the inference update equations are provided in the Appendix, and below we summarize unique aspects of the inference associated with the particular model considered.

A. MCMC inference

The MCMC algorithm involves a sequence of Gibbs updates, where each latent variable is resampled conditioned on the rest. Most conditional updates are conjugate, the important exception being the coefficients $\gamma_{\ell,i}$ at each level; these have conditionals

$$\begin{aligned} p(\tilde{\gamma}_{\ell,i}|-) \;\propto\; & \mathrm{InvGa}(\alpha_{\ell,i}|1,(2\tilde{\gamma}_{\ell,i}^{-1}))\mathrm{Dir}(\tilde{\boldsymbol{\gamma}}_\ell|\tilde{\boldsymbol{\gamma}}_{\ell-1})\\ & \times\mathrm{Dir}(\tilde{\boldsymbol{\gamma}}_{\ell+1}|\tilde{\boldsymbol{\gamma}}_\ell), \end{aligned} \tag{20}$$

$$\begin{aligned} p(\gamma_{\ell,i}|-) \;\propto\; & \mathrm{GIG}\left(2,\frac{\sum_{j\neq i}\gamma_{\ell,j}}{\alpha_{\ell,i}},\frac{\tilde{\gamma}_{\ell-1,pa(\ell,i)}}{n_c}-1\right)\\ & \times\mathrm{Dir}\left(\tilde{\boldsymbol{\gamma}}_{\ell+1}|\tilde{\boldsymbol{\gamma}}_\ell\right)\;\sum_j\gamma_{\ell,j}. \end{aligned} \tag{21}$$

Recall that $\tilde{\gamma}_{\ell,i}$ is just $\gamma_{\ell,i}$ normalized. We performed this update using a Metropolis step [51], proposing from the generalized-inverse-Gaussian (GIG) distribution specified by the first term, with the last two terms determining the acceptance probability. In our experiments, we observed acceptance rates of about 80%.

B. Variational posterior approximation

Here we deterministically approximate the posterior distribution $p(\cdot|\boldsymbol{y})$ by a parametric distribution $q(\cdot)$, which we then optimize to match the true posterior distribution. We use a mean-field factorized approximation, assuming a complete factorization across the latent variables, $q(\boldsymbol{\Theta}) = \prod_i q_i(\Theta_i)$. We approximate the distribution of each wavelet coefficient $x_{\ell,i}$ as a Gaussian distribution with mean $\mu_{x_{\ell,i}}$ and variance $\sigma^2_{\ell,i}$. The marginal distributions over α and τ_ℓ were set to exponential distributions. The $\tilde{\boldsymbol{\gamma}}_\ell$ vectors were set as Dirichlet distributions with parameter vector $(b_1, \cdots, b_{n_c})$. To optimize these parameters, we used a heuristic very closely related to expectation propagation [52]. For any variable (say x), the corresponding Gibbs update rule from the previous section gives its distribution conditioned on its Markov blanket. We look at this conditional distribution, plugging in the the average configuration of the Markov blanket (as specified by the current settings of the posterior approximations $q(\cdot)$). We then calculate the relevant moments of resulting conditional distribution over x, and update the variational distribution $q(x)$ to match these moments. We repeat this procedure, cycling over all variables, and seeking a compatible configuration of all parameters (corresponding to a fixed point of our algorithm). As mentioned earlier, the resulting update equations recall various frequentist algorithms, and we list them in the Appendix. Two important ones are:

$$\langle x_{\ell,i}\rangle = \langle\alpha_0\rangle\sigma^2_{x_{\ell,i}}\boldsymbol{\Psi}_k^T\left(\boldsymbol{y} - \sum_{l=1,l\neq k}^{n}\boldsymbol{\Psi}_l\langle x_l\rangle\right), \quad (22)$$

$$\sigma^2_{\ell,i} = \langle\alpha_0\rangle^{-1}(\langle\tau_\ell\rangle\langle\alpha_{\ell,i}\rangle + \boldsymbol{\Psi}_k^T\boldsymbol{\Psi}_k)^{-1}, \quad (23)$$

where k is the index of $\boldsymbol{x}$ corresponding to $x_{\ell,i}$, $\boldsymbol{\Psi}_k$ is the kth column of $\boldsymbol{\Psi}$, and $\langle\cdot\rangle$ denotes the expectation value of the entry in $\langle\,\rangle$. It is worth noting that updating the γ's requires calculating the mean of the distribution specified in equation (21) for the current averages of the posterior approximation. Unfortunately, this has no closed form. One approach is to calculate an importance sampling-based Monte Carlo estimate of this quantity, using the proposal distribution specified in the previous section. If s samples were used to produce this estimate, we call the resulting algorithm VB(s). The high acceptance rate of samples from the GIG suggests that this is a reasonably accurate approximation to the intractable distribution. This suggests a simpler approach where all terms are ignored except for the GIG, whose mean is used to update the γ's. In this case, the update rules become:

$$\langle\gamma_{\ell,i}\rangle = \frac{\sqrt{\langle\frac{1}{\alpha_{\ell,i}}\rangle\sum_{j\neq i}\langle\tilde{\gamma}_{\ell,j}\rangle}K_{\tilde{\gamma}_{\ell-1,pa(\ell,i)}/n_c}(\chi)}{\sqrt{2}K_{(\tilde{\gamma}_{\ell-1,pa(\ell,i)}/n_c-1)}(\chi)},$$

where $K_p(\chi)$ is the modified Bessel function of the second kind, and $\chi = \sqrt{2\langle\frac{1}{\alpha_{\ell,i}}\rangle\sum_{j\neq i}\langle\tilde{\gamma}_{\ell,j}\rangle}$. We call this approximation a-VB. This approximation can also be interpreted from the perspective of a wavelet hidden Markov (HM) tree [14], in which the dependence is always from the parent-node, rather than the child-nodes. Thus, the underlying dependence inside the HM tree is *single* direction, from parent to children.

C. MAP estimates via Expectation-Maximization

Closely related to the previous approach is an algorithm that returns MAP estimates of the wavelet coefficients, while (approximately) averaging out all other variables. This maximization over $\gamma_{\ell,i}$ must also be done approximately, but is a straightforward modification of the variational update of $\gamma_{\ell,i}$. From the properties of the GIG, the M-step becomes

$$\gamma_{\ell,i} = \frac{\eta + \sqrt{\eta^2 + \frac{2\sum_{j\neq i}\gamma_{\ell,j}}{\alpha_{\ell,i}}}}{2}, \quad (24)$$

where $\eta = \frac{\tilde{\gamma}_{\ell-1,pa(\ell,i)}}{n_c} - 2$. The approximate E-Step remains unaffected.

VIII. Numerical Results

Our code is implemented in MATLAB, with all results generated on a laptop with a 2.7 GHz CPU and 8 GB RAM. Parameters α_0, τ_ℓ and ν are all drawn form a broad gamma prior, Ga$(10^{-6}, 10^{-6})$. When performing inference, all parameters are initialized at random. No parameter tuning has been performed. We run VB and EM for 100 iterations, while our MCMC inferences were based on runs of 5000 samples with a discarded burn-in of 1000 samples. These number of iterations were not optimized, and typically excellent results are obtained with far fewer.

A. Details on CS implementation and comparisons

We apply the proposed approach, which we refer to as a shrinkage hierarchical model (denoted "s-HM" in the following figures) to compressive sensing (CS). In these experiments the elements of the projection matrix $\boldsymbol{H} \in \mathbb{R}^{m\times n}$ are drawn i.i.d. from $\mathcal{N}(0,1)$, and the "CS ratio," denoted CSR, is m/n.

In addition to considering s-HM, we consider a "flat" version of the shrinkage prior, ignoring the tree structure (analogous to models in [21]); for this case the prior on the wavelet coefficients is as in (5). Additionally, we make comparisons to the following algorithms (in all cases using publicly available code, with default settings):

- TSW-CS [14], [17], using code from http://people.ee.duke.edu/~lcarin/BCS.html
- OMP [11], using code from http://www.cs.technion.ac.il/~elad/software/
- ℓ_1 magic with TV norm (ℓ_1-TV) [4], using code from http://users.ece.gatech.edu/~justin/l1magic/
- Bayesian compressive sensing (BCS) [13], using code from http://people.ee.duke.edu/~lcarin/BCS.html
- linearized Bregman [53], using code from http://www.caam.rice.edu/~optimization/linearized_bregman/

For all algorithms to which we compared, the parameters were adjusted and optimized, to try to yield their best results. However, we found that the optimal performance in each case was with parameters near the default settings. Therefore, results for default settings are reported, for repeatability in subsequent comparisons. We also note that the proposed method was robust to parameter settings, with similar results found for a wide range of parameters. Consequently, it is felt that the presented comparisons are fair.

B. CS and wavelet-based representation

We consider images of size 128×128, and hence $n = 16,384$. We use the "Daubechies-4" wavelets [1], with scaling coefficients (LL channel) of size 8×8. In all Bayesian models, the posterior mean results are used for the reconstructed image.

Figure 2 plots the reconstruction peak signal-to-noise-ratio (PSNR) versus CS ratio for "Barbara". We see that a) the proposed s-HM tree model provides best results, and b) models (s-HM and TSW-CS) with tree structure are better than those without tree structure, consistent with the theory in [25].

Similar observations hold for other widely considered images, viz. "Lena" and "Cameraman" (omitted for brevity). The PSNR of the reconstructed images by the proposed s-HM tree model is about 1dB higher than TSW-CS. It is worth noting that even *without* the tree structure, the proposed local-global shrinkage prior performs well. The other tree-based model, TSW-CS, has a prior that, via a spike-slab construction, encourages explicit sparsity; this seems to undermine results relative to the proposed shrinkage-based tree model, which we examine next.

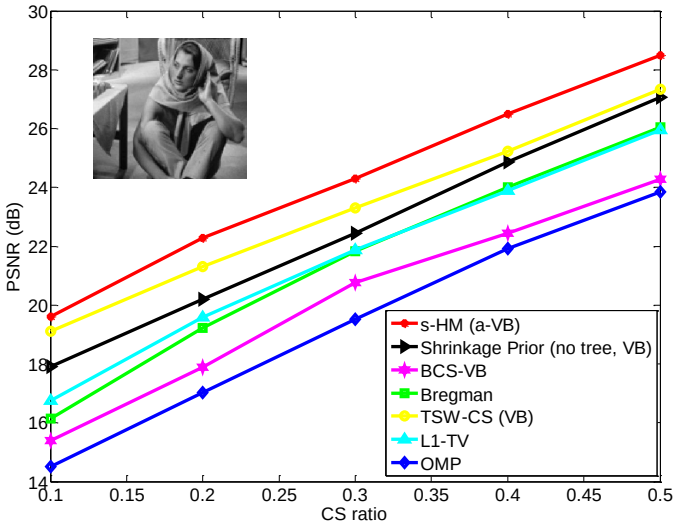

Fig. 2: PSNR comparison of various algorithms for compressive sensing with wavelet-based image representation.

1) Estimated Wavelet Coefficients: We compare the posterior distribution (estimated by MCMC) of the estimated wavelet coefficients from the proposed model versus TSW-CS [14], in which the tree-based Markovian spike-slab model is used. We consider "Barbara" with a CSR=0.4. An example of the posterior distribution for a typical wavelet coefficient is shown in Figure 3, comparing the proposed s-HM with TSW-CS. Note that s-HM puts much more posterior mass around the true answer than TSW-CS.

Because of the spike-slab construction in TSW-CS, each draw from the prior has explicit sparsity, while each draw from the prior for the continuous-shrinkage-based s-HM always has all coefficients being non-zero. While TSW-CS and s-HM both impose and exploit the wavelet tree structure, our experiments demonstrate the advantage of s-HM because it imposes compressibility rather than explicit sparsity. Specifically, note from Figure 3 that the coefficient inferred via TSW-CS is shrunk toward zero more than the s-HM, and TSW-CS has little posterior mass around the true coefficient value. This phenomenon has been observed in numerous comparisons. The over-shrinkage of TSW-CS implies an increased estimate of the noise level $\boldsymbol{n}$ for that algorithm (the small wavelet coefficients are shrunk to exactly zero, and the manifested residual is incorrectly attributed to $\boldsymbol{n}$). This is attributed as the key reason the s-HM algorithm consistently performs better than TSW-CS for CS inversion.

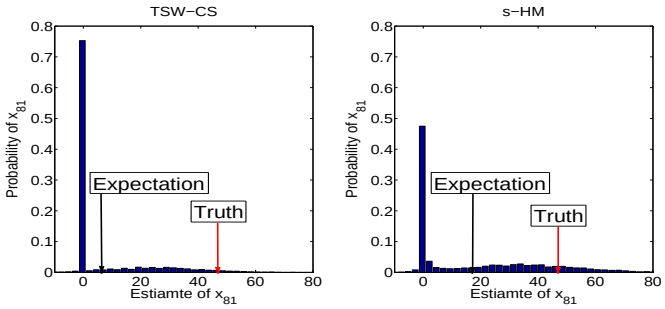

Fig. 3: Distribution of estimated x_{81} by proposed s-HM model (right) and TSW-CS (left).

2) Robustness to Noise: Based upon the above discussions, it is expected that s-HM will more accurately estimate the variance of additive noise $\boldsymbol{n}$, with TSW-CS over-estimating the noise variance. We study the robustness of CS reconstruction under different levels of measurement noise, by adding zero-mean measurement Gaussian noise when considering "Barbara"; the standard deviation here takes values in [0, 0.31] (the pixel values of the original image were normalized to lie in [0,1]). Figure 4 (right) plots the PSNR of the reconstruction results of the three best algorithms from Figure 2 (our tree and flat shrinkage models, and TSW-CS). We see that the proposed s-HM tree model is most robust to noise, with its difference from TSW-CS growing with noise level.

Figure 4 (left) plots the noise standard derivation inferred by the proposed models and TSW-CS (mean results shown). We see the tree structured s-HM model provides the best estimates of the noise variance, while the TSW-CS overestimates it. The noise-variance estimated by the flat shrinkage model is under-estimated (apparently the absence of the wavelet tree structure in the flat model causes the model to attribute some of the noise incorrectly to wavelet coefficients).

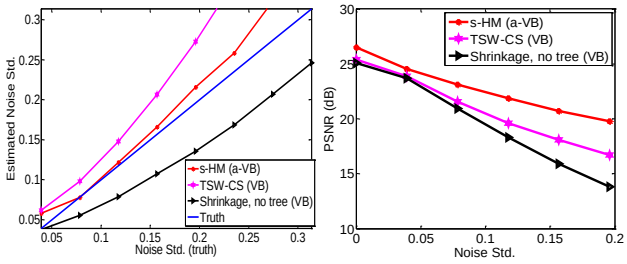

Fig. 4: Left: inferred noise standard deviation plotted versus truth. Right: reconstructed PSNR of "Barbara" with different algorithms, at CS ratio = 0.4 under various noise levels.

TABLE I: Reconstruction PSNR (dB) with different inferences of "Barbara" for the proposed s-HM model.

CS ratio	MCMC	VB(500)	VB(100)	VB(50)	a-VB	EM	VB(TSW-CS)
0.4	**26.83**	26.32	26.26	26.07	26.50	26.08	25.26
0.5	**28.51**	28.08	27.98	27.96	28.50	27.72	27.25

3) VB and EM Inference: The above results were based on an MCMC implementation. We next evaluate the accuracy and computation time of the different approximate algorithms. The computation time of our approximate VB and EM (one iteration 2 sec at CSR$=0.4$) is very similar to TSW-CS (one iteration 1.62 sec). The bottleneck in our methods is the evaluation of the Bessel function needed for inferences. This can be improved ($e.g.$, by pre-computing a table of Bessel function evaluations). Note, however, that our methods provide better results (Figure 2 and Table I) than TSW-CS.

Table I also compares the different variational algorithms. Recall that these involved solving an integral via Monte Carlo (MC), and we indicate the number of samples used by 'num', giving "VB(num)". As the number of samples increases, the PSNR increases, but at the cost of computation time (the algorithm with 20 MC samples takes 9.34 sec). It is interesting to note that using the approximate VB and approximate EM, the PSNR of the reconstructed image is higher than TSW-CS (VB) and even VB(500). Since the VB and EM provide similar results (VB is slightly better) and with almost the same computational time, we only show the results of a-VB.

C. CS and block-DCT representation

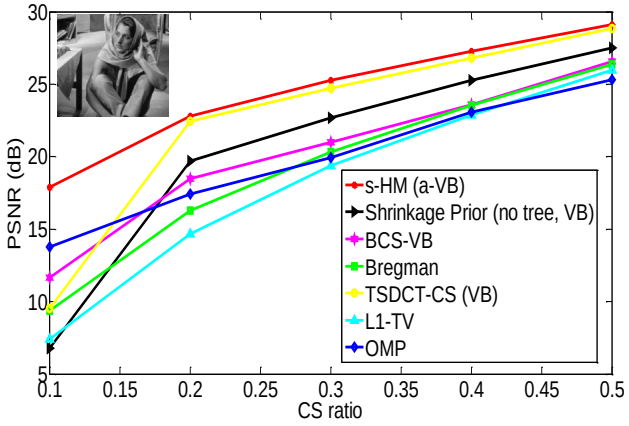

Fig. 5: PSNR comparison of various algorithms for compressive sensing with block-DCT image representation. The "TSDCT-CS" is the model implemented in [17], while the "s-HM" denotes proposed model with block-DCT tree structure, and "Shrinkage Prior (no tree)" symbolizes the flat model without tree-structure.

The principal focus of this paper has been on wavelet representations of images. However, we briefly demonstrate that the same model may be applied to other classes of tree structure. In particular, we consider an 8×8 block-DCT image representation, which is consistent with the JPEG image-compression standard. Details on how this tree structure is constituted are discussed in [3], and implementation details for CS are provided in [17]). Figure 5 shows CS results for "Barbara" using each of the methods discussed above, but now using a block-DCT representation. We again note that the proposed hierarchical shrinkage method performs best in this example, particularly for a small number of measurements ($e.g.$, CSR of 0.1).

As in [17], for most natural images we have found that the wavelet representation yields better CS recovery than the block-DCT representation. The results in Figures 2 and 5 are consistent with all results we have computed over a large range of the typical images used for such tests.

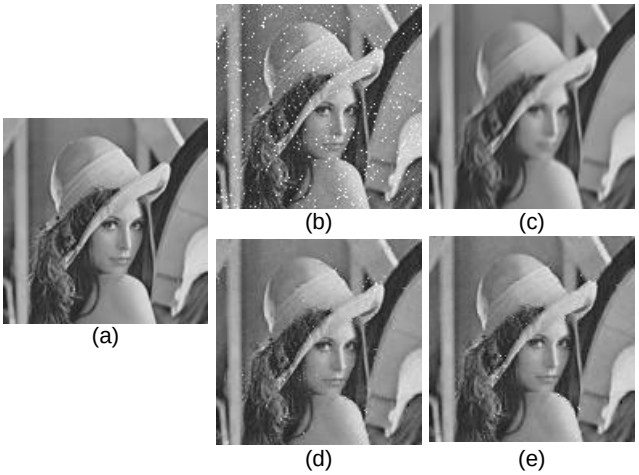

Fig. 6: Denoising. (a) original image, (b) noisy image (PSNR: 19.11dB), (c) dHBP (PSNR: 30.31dB), (d) shrinkage prior without tree (PSNR: 29.18dB), and (e) s-HM (PSNR: 30.40dB).

D. Denoising with spiky-plus-Gaussian noise

We now apply the proposed model to denoising, with measurement noise that is a superposition of a Gaussian component and a spiky component, as in Section VI. Inference is similar to Section VII. Figure 6 compares the performance of our model with the state-of-the-art dHBP (dependent Hierarchical Beta Process) model in [8], considering the same type of Gaussian+spiky noise considered there. Specifically, 3.2% of the pixels are corrupted by spiky noise with amplitudes distributed uniformly at random over [0,1] (scaled images as before), and the zero-mean Gaussian noise is added with standard deviation 0.02. While our results are similar to dHBP from the (imperfect) measure of PSNR, visually, the dHBP returns an image that appears to be more blurred than ours. For example, consider the detail in the eyes and hair in s-HM versus dHBP, Figures 6(c) and 6(e), respectively. Such detailed difference in local regions of the image are lost in the global PSNR measure. Other commonly used images, including "Barbara", "Cameraman", "Pepper" and "House" were also tested, with similar observations.

In the dictionary-learning method of dHBP in [8], neighboring patches are encouraged to share dictionary elements. This has the advantage of removing spiky noise, but it tends to remove high-frequency details (which often are not well shared between neighboring patches). In our work the spiky noise is removed because it is generally inconsistent with the wavelet-tree model developed here, that is characteristic of the wavelet coefficients of natural images but not of noise (Gaussian or spiky). Hence, our wavelet-tree model is encouraged to model the underling image, and the noise is attributed to the noise terms $\boldsymbol{w}$ and $\boldsymbol{n}$ in our model.

IX. Conclusions

We have developed a multiscale Bayesian shrinkage framework, accounting for the tree structure underlying a wavelet or block-DCT representation of a signal. The model that has been our focus is based on a gamma-distribution-based hierarchical construction. It was demonstrated that many shrinkage priors may be placed within the context of infinitely divisible random variables, with the gamma distribution but one example. The proposed approach yields state-of-the-art results, using both a wavelet or block-DCT based image representation, for the problems of CS image recovery and denoising with Gaussian plus spiky noise.

Concerning future research, it is of interest to examine other Lévy measure classes. For example, in applications for which there are only very few large coefficients, the symmetric alpha-stable distribution [40] may be most appropriate. Additionally, in this paper it has been assumed that the tree-based basis is known. In some applications one is interested in learning a tree-based dictionary matched to the data [54]. It is of interest to examine the utility of the proposed approach to this application, particularly because of its noted connection to dictionary learning. The proposed shrinkage priors have the advantage of, when viewed from the standpoint of the log posterior, being closely connected to many of the dictionary-learning methods used in the optimization literature [54].

References

[1] S. Mallat, *A wavelet tour of signal processing: The sparse way*. Academic Press, 2008.

[2] M. Shapiro, "Embedded image coding using zerotrees of wavelet coefficients," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3445–3462, Decemeber 1993.

[3] Z. Xiong, O. G. Gulerguz, and M. T. Orchard, "A DCT-based embedded image coder," *IEEE Signal Processing Letters*, vol. 3, no. 11, pp. 289–290, November 1996.

[4] E. J. Candès and T. Tao, "Decoding by linear programming," *IEEE Transactions on Information Theory*, vol. 51, no. 12, pp. 4203–4215, December 2005.

[5] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, April 2006.

[6] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 489–509, February 2006.

[7] J. Portilla, V. Strela, M. J. Wainwright, and E. P. Simoncelli, "Image denoising using scale mixtures of gaussians in the wavelet domain," *IEEE Transactions on Image Processing*, vol. 12, no. 11, pp. 1338–1351, November 2003.

[8] M. Zhou, H. Yang, G. Guillermo, D. Dunson, and L. Carin, "Dependent hierarchical beta process for image interpolation and denoising," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

[9] F. Couzinie-Devy, J. Sun, K. Alahari, and J. Ponce, "Learning to estimate and remove non-uniform image blur," *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[10] J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Transactions on Information Theory*, vol. 50, no. 10, pp. 2231–2242, October 2004.

[11] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Transactions on Information Theory*, vol. 53, no. 12, pp. 4655–4666, December 2007.

[12] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde, "Model-based compressive sensing," *IEEE Transactions on Information Theory*, vol. 56, no. 4, pp. 1982–2001, April 2010.

[13] S. Ji, Y. Xue, and L. Carin, "Bayesian compressive sensing," *IEEE Transactions on Signal Processing*, vol. 56, no. 6, pp. 2346–2356, June 2008.

[14] L. He and L. Carin, "Exploiting structure in wavelet-based bayesian compressive sensing," *IEEE Transactions on Signal Processing*, vol. 57, no. 9, pp. 3488–3497, September 2009.

[15] D. Baron, S. Sarvotham, and R. G. Baraniuk, "Bayesian compressive sensing via belief propagation," *IEEE Transactions on Signal Processing*, vol. 58, no. 1, pp. 269–280, January 2010.

[16] M. E. Tipping, "Sparse bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, vol. 1, pp. 211–244, September 2001.

[17] L. He, H. Chen, and L. Carin, "Tree-structured compressive sensing with variational bayesian analysis," *IEEE Signal Processing Letters*, vol. 17, no. 3, pp. 233–236, 2010.

[18] M. Seeger, "Bayesian inference and optimal design for the sparse linear model," *J. Machine Learning Res.*, vol. 9, June 2008.

[19] Z. Zhang, S. Wang, D. Liu, and M. I. Jordan, "EP-GIG priors and applications in bayesian sparse learning," *The Journal of Machine Learning Research*, vol. 13, pp. 2031–2061, 2012.

[20] N. G. Polson and J. G. Scott, "Shrink globally, act locally: Sparse bayesian regularization and prediction," *Bayesian Statistics*, 2010.

[21] ——, "Local shrinkage rules, Lévy processes and regularized regression," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 74, no. 2, pp. 287–311, 2012.

[22] C. M. Carvalho, N. G. Polson, and J. G. Scott, "The horseshoe estimator for sparse signals," *Biometrika*, vol. 97, no. 2, pp. 465–480, 2010.

[23] A. Armagan, M. Clyde, and D. B. Dunson, "Generalized beta mixtures of gaussians," *Advances in Neural Information Processing Systems (NIPS)*, pp. 523–531, 2011.

[24] Z. Zhang and B. Tu, "Nonconvex penalization, Lévy processes and concave conjugates," *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[25] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde, "Model-based compressive sensing," *IEEE Transactions on Information Theory*, vol. 56, no. 4, pp. 1982–2001, April 2010.

[26] H. Zou, "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1418–1429, December 2006.

[27] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *Journal of the ACM*, vol. 58, no. 3, pp. 1–37, May 2011.

[28] R. Baraniuk, "Compressive sensing," *IEEE Signal Processing Magazine*, vol. 24, no. 4, pp. 118–121, July 2007.

[29] D. V. D. Ville, T. Blu, and M. Unser, "Isotropic polyharmonic B-splines: scaling functions and wavelets," *IEEE Trans. Image Process.*, pp. 1798–1813, November 2005.

[30] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58, pp. 267–288, 1994.

[31] T. Park and G. Gasella, "The bayesian lasso," *Journal of the American Statistical Association*, vol. 103, pp. 681–686, 2008.

[32] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58, pp. 267–288, 1994.

[33] P. Bühlmann and T. Hothorn, "Spike and slab variable selection: Frequentist and bayesian strategies," *Statistical Science*, pp. 730–773, 2005.

[34] P. Schniter, L. Potter, and J. Ziniel, "Fast bayesian matching pursuit: Model uncertainty and parameter estimation for sparse linear models," *Proc. Information Theory and Applications Workshop, 2008, 326-333*, 2008.

[35] M. Crouse, R. Nowak, and R. Baraniuk, "Wavelet-based statistical signal processing using hidden markov models," *IEEE Transactions on Signal Processing*, vol. 46, no. 4, pp. 886–902, April 1998.

[36] D. Applebaum, *Lévy Processes and Stochastic Calculus*. Cambridge University Press, 2004.

[37] U. Kamilov, P. Pad, A. Amini, and M. Unser, "MMSE estimation of sparse Lévy processes," *IEEE Trans. Signal Process.*, pp. 137–147, January 2013.

[38] E. Bostan, U. Kamilov, M. Nilchian, and M. Unser, "Sparse stochastic processes and discretization of linear inverse problems," *IEEE Trans. Image Process.*, pp. 2699–2710, July 2013.

[39] M. Unser, P. Tafti, and Q. Sun, "A unified formulation of Gaussian versus sparse stochastic processes - Part I: Continuous-domain theory," *IEEE Trans. Information Theory*, pp. 1945–1962, March 2014.

[40] R. L. Wolpert, M. A. Clyde, and C. Tu, "Stochastic expansions using continuous dictionaries: Lévy Adaptive Regression Kernels," *Annals of Statistics*, 2011.

[41] T. J. Mitchell and J. J. Beauchamp, "Bayesian variable selection in linear regression," *Journal of the American Statistical Association*, vol. 83, no. 404, pp. 1023–1032, 1988.

[42] E. I. George and R. E. McCulloch, "Variable selection via gibbs sampling," *Journal of the American Statistical Association*, vol. 88, no. 423, pp. 881–889, 1993.

[43] V. Cevher, "Learning with compressible priors," *Advances in Neural Information Processing Systems (NIPS)*, pp. 261–269, 2009.

[44] C. M. Carvalho, N. G. Polson, and J. G. Scott, "Handling sparsity via the horseshoe," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2009, pp. 73–80.

[45] I. Daubechies, R. DeVore, M. Fornasier, and C. S. Güntürk, "Iteratively reweighted least squares minimization for sparse recovery," *Communications on Pure and Applied Mathematics*, vol. 63, no. 1, pp. 1–38, 2010.

[46] E. J. Candes, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5-6, pp. 877–905, 2008.

[47] P. Garrigues and B. A. Olshausen, "Group sparse coding with a laplacian scale mixture prior," *Advances in Neural Information Processing Systems (NIPS)*, pp. 676–684, 2010.

[48] J. E. Griffin and P. J. Brown, "Bayesian hyper-lassos with non-convex penalization," *Australian & New Zealand Journal of Statistics*, vol. 53, no. 4, pp. 423–442, 2011.

[49] A. Bhattacharya, D. Pati, N. S. Pillai, and D. B. Dunson, "Bayesian shrinkage," *arXiv:1212.6088*, 2012.

[50] S. Som and P. Schniter, "Compressive imaging using approximate message passing and a markov-tree prior," *IEEE Transactions on Signal Processing*, pp. 3439–3448, 2012.

[51] W. K. Hastings, "Monte carlo sampling methods using markov chains and their applications," *Biometrika*, pp. 97–109, 1970.

[52] T. P. Minka, "A family of algorithms for approximate Bayesian inference," *Ph.D. dissertation, MIT*, 2001.

[53] W. Yin, S. Osher, D. Goldfarb, and J. Darbon, "Bregman iterative algorithms for ℓ_1-minimization with applications to compressed sensing," *SIAM Journal on Imaging Sciences*, vol. 1, no. 1, pp. 143–168, 2008.

[54] R. Jenatton, J. Mairal, G. Obozinski, and F. Bach, "A multiscale framework for compressive sensing of proximal methods for sparse hierarchical dictionary learningvideo," *International Conference on Machine Learning (ICML)*, 2010.