Open access • Journal Article • DOI:10.1109/TSP.2009.2022003

# Exploiting Structure in Wavelet-Based Bayesian Compressive Sensing

— Source link

Lihan He, Lawrence Carin

**Institutions:** Duke University

**Published on:** 01 Sep 2009 - IEEE Transactions on Signal Processing (IEEE)

**Topics:** Wavelet packet decomposition, Stationary wavelet transform, Wavelet transform, Wavelet and Cascade algorithm

# Exploiting Structure in Wavelet-Based Bayesian Compressive Sensing

Lihan He and Lawrence Carin

Department of Electrical and Computer Engineering

Duke University, Durham, NC 27708-0291 USA

{lihan,lcarin}@ece.duke.edu

EDICS: DSP-RECO, MAL-BAYL

### Abstract

Bayesian compressive sensing (CS) is considered for signals and images that are sparse in a wavelet basis. The statistical structure of the wavelet coefficients is exploited explicitly in the proposed model, and therefore this framework goes beyond simply assuming that the data are compressible in a wavelet basis. The structure exploited within the wavelet coefficients is consistent with that used in wavelet-based compression algorithms. A hierarchical Bayesian model is constituted, with efficient inference via Markov chain Monte Carlo (MCMC) sampling. The algorithm is fully developed and demonstrated using several natural images, with performance comparisons to many state-of-the-art compressive-sensing inversion algorithms.

### Index Terms

Bayesian signal processing, wavelets, sparseness, compression

## I. Introduction

Over the last two decades there has been significant research directed toward development of transform codes, with the discrete-cosine and wavelet transforms [1] constituting two important examples. The discrete cosine transform (DCT) is employed in the JPEG standard [2], with wavelets employed in the JPEG2000 standard [3]. Wavelet-based transform coding [4] explicitly exploits the structure [5] manifested in the wavelet coefficients of typical data. Specifically, for most natural data (signals and images) the wavelet coefficients are compressible, implying that a large fraction of the coefficients may be set to zero with minimal impact on the signal-reconstruction accuracy.

A discrete wavelet transform may be implemented via a series of high- and low-pass filters, with decimation performed after each such filtering [1]. This naturally yields a quadtree structure of the wavelet coefficients for an image [1], with each wavelet coefficient generally serving as a “parent” for four “children” coefficients. The wavelet coefficients at the coarsest scale serve as “root nodes” for the quadtrees, with the finest scale of coefficients constituting the “leaf nodes”. For most natural images the negligible wavelet coefficients tend to be clustered together; specifically, if a wavelet coefficient at a particular scale is negligible, then its children are also generally (but not always) negligible. This leads to the concept of “zero trees” [4] in which a tree or subtree of wavelet coefficients are all collectively negligible. The structure of the wavelet coefficients, and specifically zero trees, are at the heart of most wavelet-based compression algorithms, and specifically JPEG2000.

Transform coding, particularly JPEG and JPEG2000, are now widely used in digital media. One observes, however, that after the digital data are measured and then transform compressed, one often “throws away” a large fraction of the transform coefficients, while still achieving accurate data reconstruction. This seems wasteful, since there are many applications for which data collection is expensive. For example, the collection of magnetic-resonance imagery (MRI) is time consuming and often uncomfortable for the patient, and hyperspectral cameras require measurement of images at a large number of spectral bands. Since collection of such data is expensive, and because after transform encoding a large fraction of the data are ultimately discarded in some sense, this suggests the following question: Is it possible to measure the informative part of the data directly, thereby reducing measurement costs, while still retaining all of the informative parts of the data? This goal has spawned the new field of compressive sensing (or compressed sensing) [6], [7], [8], in which it has been demonstrated that if the signal of interest is sparse in some basis, then with a relatively small number of appropriately designed projection measurements the underlying signal may be recovered *exactly*. If the data are compressible but not exactly sparse in a particular basis (many coefficients are negligibly small, but not exactly zero), one may still employ compressive sensing (CS) to recover the data up to an error proportional to the energy in the negligible coefficients [9]. Two of the early important applications of CS are in MRI [10] and in development of new hyperspectral cameras

[11].

Details on how to design the compressive-sensing projection vectors, and requirements on the (typically relatively small) number of such projections, may be found in [6], [7], [8], [9]. Assume that the set of $N$ CS projection measurements are represented by the vector $\boldsymbol{v}$, and that these measurements may be represented as $\boldsymbol{v} = \boldsymbol{\Phi}\boldsymbol{\theta}$, where $\boldsymbol{\theta}$ represents an $M \times 1$ vector of transform coefficients, and $\boldsymbol{\Phi}$ is an $N \times M$ matrix constituted via the compressive-sensing measurements; in CS it is desired that $N \ll M$. Given $\boldsymbol{\theta}$, one may recover the desired underlying signal via an inverse transform ($e.g.$, an inverse DCT or wavelet transform, depending on which basis is employed, in which $\boldsymbol{\theta}$ is sparse or compressible). Note that in CS we do not measure $\boldsymbol{\theta}$ directly, but rather projections on $\boldsymbol{\theta}$. One must infer $\boldsymbol{\theta}$ from $\boldsymbol{v}$, this generally an ill-posed inverse problem because $N < M$. To address this problem, many CS inversion algorithms seek to solve for $\boldsymbol{\theta}$ with the following $\ell_1$-regularized optimization problem:

$$\boldsymbol{\theta} = \arg\min_{\tilde{\boldsymbol{\theta}}} \|\tilde{\boldsymbol{\theta}}\|_{\ell_1} \quad \text{s.t.} \quad \boldsymbol{v} = \boldsymbol{\Phi}\tilde{\boldsymbol{\theta}}. \tag{1}$$

If $\boldsymbol{\theta}$ is sparse, with $S$ non-zero coefficients ($S \ll M$), then CS theory indicates that if $\boldsymbol{\Phi}$ is constructed properly (more on such constructions in Section III) then with "overwhelming probability" [6], [7], [8], [9] one may recover $\boldsymbol{\theta}$ *exactly* if $N > O(S \cdot \log(M/S))$; similar relationships hold when $\boldsymbol{\theta}$ is compressible but not exactly sparse. We note that there are many CS inversion algorithms that do not explicitly solve an $\ell_1$-based inversion but that are similarly motivated by sparseness; among these are [12], [13], [14].

The aforementioned $\ell_1$ inversion may be viewed as a maximum *a posteriori* estimate for $\boldsymbol{\theta}$ under the assumption that each component of $\boldsymbol{\theta}$ is drawn i.i.d. from a Laplace prior [15]. This i.i.d. assumption does not impose anticipated structure/correlation between transform coefficients. While this leads to development of many algorithms for CS inversion (see [15], [16], [12], [14], [17], among many others), such a formulation does not exploit all of the prior information available about the transform coefficients $\boldsymbol{\theta}$. For example, as discussed above with respect to the wavelet transform, there is anticipated structure in $\boldsymbol{\theta}$ that may be exploited to further constrain or regularize the inversion, ideally reducing the number of required CS measurements $N$. This concept has been made rigorous recently for sparse $\boldsymbol{\theta}$ [18], as well as for compressible $\boldsymbol{\theta}$ [19], [20], [21]; these papers demonstrate that one may achieve accurate CS inversions with substantially fewer projection measurements (smaller $N$) if known properties of the structure of

$\boldsymbol{\theta}$ are exploited properly.

The explicit use of the structure in wavelet coefficients has been used in many image analysis and processing applications. As indicated above, this structure is employed in image compression [4]. It is also employed in image denoising [22], as well as texture synthesis and image inpainting [23]. More directly related to CS, the wavelet tree structure has been employed in non-statistical CS inversion [24], and within more statistical settings via the hidden Markov tree (HMT) [25]. There have also been methods that augment the CS sensing structure, with linkage to the scales in a wavelet decomposition [26].

In addition to imposing and exploiting prior knowledge about structure in a wavelet decomposition of images, other forms of prior knowledge have been exploited in CS. For example the use of total variation (TV) [27] is generally a non-statistical approach that may be employed to account for prior knowledge about the properties of images. Researchers have also developed techniques that impose prior structure through learning the appropriate basis for sparse representation [28], [29]. Therefore prior knowledge about images and the CS sensing process has been used previously, with this prior knowledge not limited to wavelets.

While the above references impose various forms of prior information, in this paper that information is explicitly employed within a *Bayesian prior*. This allows a statistical CS inversion, unlike that in [24], [21], for example. As opposed to the work in [21], we do not make an explicit ("hard") imposition of the structure of the coefficients, but rather impose the structure statistically. The proposed technique is most closely related to recent Bayesian CS approaches for imposing prior belief about the signal of interest, usually in terms of a sparseness prior [15], [30] (this is closely related to more general research on imposing sparseness in Bayesian priors [31]). None of these previous Bayesian approaches explicitly imposed the known statistical structure of the wavelet decomposition of images, this constituting an important contribution of this paper. While a related statistical model was considered in [25], that framework required explicit *a priori* model training based on representative images. The method proposed here requires no *a priori* training imagery, and therefore it is robust to new image classes not observed previously. In addition, the model in [25] has "high" and "low" states, with the low states corresponding to the negligible wavelet coefficients; both the high and low states are characterized by Gaussian statistics. In the model proposed here a different but related construction is employed, in terms of a "spike-slab" prior [32], [33], [34], [35], [36], and here the coefficients in the "low" states are explicitly set to zero (sparseness is explicitly imposed). The proposed Bayesian approach yields “error bars” on the recovered image, which is a unique advantage of Bayesian approaches [15], [30], relative to all non-statistical approaches to CS inversion (the method in [25] also does not yield error bars). For example the methods in [19], [20], [21] also impose structure in the signal, but not in a Bayesian framework. In addition to developing the Bayesian model for exploiting wavelet structure in CS, a fast algorithm is presented, with empirical computational cost that is quite competitive (often superior) with existing CS algorithms.

The remainder of the paper is organized as follows. In Section II we review the structure inherent to wavelet coefficients in natural images, and in Section III we describe how this structure may be exploited in a Bayesian CS inversion framework. Example results are presented in Section IV, with comparisons to many of the state-of-the-art CS algorithms. Conclusions and discussions of future work are provided in Section V.

## II. Wavelet Tree Structure

The discrete wavelet transform may be represented in matrix form as [1]

$$\boldsymbol{x} = \boldsymbol{\Psi}\boldsymbol{\theta} \tag{2}$$

where $\boldsymbol{x}$ is an $M \times 1$ real vector of data, $\boldsymbol{\Psi}$ is an $M \times M$ matrix with columns corresponding to orthonormal scaling and wavelet basis vectors, and $\boldsymbol{\theta}$ represents the $M \times 1$ vector of wavelet-transform coefficients. The wavelet coefficients that constitute $\boldsymbol{\theta}$ may be represented in terms of a tree structure, as depicted in Figure 1 for an image. The coefficients at scale $s = 1$ correspond to “root nodes”, and the coefficients at the largest scale $s = L$ ($L = 3$ in Figure 1) correspond to “leaf nodes”; the top-left block in Figure 1 corresponds to the scaling coefficients, denoted as $s = 0$, which capture the coarse-scale representation of the image. Each wavelet coefficient at scales $1 \leq s \leq L - 1$ has four “children” coefficients at corresponding scale $s + 1$, and it is the statistical relationships between the parent and children coefficients that is exploited in the proposed CS inversion model.

The statistics of the wavelet coefficients may be represented via the hidden Markov tree (HMT), in which the structure of the wavelet tree is exploited explicitly. In an HMT model [5] each wavelet coefficient is assumed to be drawn from one of two zero-mean Gaussian distributions, these distributions defining the observation statistics for two hidden states. One of the states is

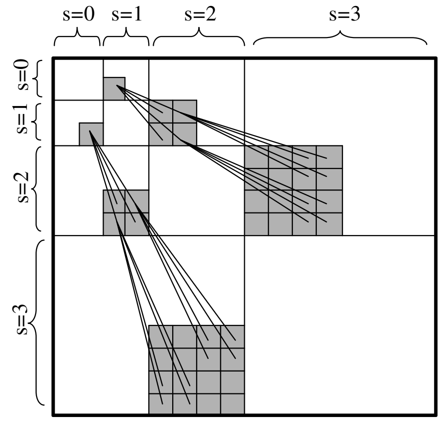


Fig. 1. Wavelet decomposition of an image, with the tree structure depicted across scales. The wavelet transform is performed with three wavelet decomposition levels, and two wavelet trees are shown in the figure. The top-left block ($s = 0$) represents scaling coefficients, and other regions are wavelet coefficients.

a "low" state, defined by a small Gaussian variance, and the "high" state is defined by a large variance. Intuitively, if a wavelet coefficient is relatively small it is more likely to reside in the "low" state; by contrast, a large wavelet coefficient has a high probability of coming from the "high" state. The probability of a given state is conditioned on the state of the parent coefficient, yielding a Markov representation across scales. The Markov transition property is represented by a $2 \times 2$ matrix $P$, with $P(i, j)$ representing the probability that children coefficients are in state $j$ given that the associated parent coefficient is in state $i$; $i = 1$ and $j = 1$ (arbitrarily) correspond to the "low" state, and $i = 2$ and $j = 2$ correspond to the "high" state. Typically $P(1, 1) = 1-\epsilon$ and $P(1, 2) = \epsilon$, where $\epsilon > 0$ satisfies $\epsilon \ll 1$. This form of $P$ imposes the belief that if a parent coefficient is small, its children are also likely to be small. We also note that $P$ generally varies between different scales and across different wavelet quadtrees. For the root nodes, $P$ is a $1 \times 2$ vector, representing an initial-state distribution.

When modeling the statistics of wavelet coefficients for a given signal, the observation is the wavelet coefficient, and for each of the two states the observation is drawn from a zero-mean Gaussian with associated variance (small variance for the "low" state and a relatively large variance for the "high" state). In compressive sensing we do not observe the wavelet coefficients directly, but rather observe projections of these coefficients. The form of the HMT will be employed within the compressive-sensing inversion, thereby explicitly imposing the belief that if a given coefficient is negligible, then its children coefficients are likely to be so as well. This imposes important structure into the form of the wavelet coefficients across scales, and it is consistent with state-of-the-art wavelet-based compression algorithms that are based upon "zero trees" (subtrees of wavelet coefficients that may all be set to zero with negligible effect on the reconstruction accuracy) [4], [37]. The motivation for the HMT construct is discussed in detail in [5].

## III. Tree-Structured Wavelet Compressive Sensing

### *A. Compressive Sensing with Wavelet-Transform Coefficients*

Assume a discrete signal/image is represented by the $M$-dimensional vector $\boldsymbol{x}$, and that it is compressible in a wavelet basis represented by the $M \times M$ matrix $\boldsymbol{\Psi}$ (defined as above). The CS measurements $\boldsymbol{v} = \boldsymbol{\Phi}\boldsymbol{\Psi}^T\boldsymbol{x} = \boldsymbol{\Phi}\boldsymbol{\theta}$, where $\boldsymbol{\Phi}$ is an $N \times M$ dimensional matrix ($N < M$), and $\boldsymbol{\theta}$ denotes an $M$-dimensional vector of wavelet-transform coefficients ($\boldsymbol{\theta} = \boldsymbol{\Psi}^T\boldsymbol{x}$). The rows of $\boldsymbol{\Phi}$ correspond to randomly defined projection vectors [7], [8]. For most natural signals $\boldsymbol{\theta}$ is compressible, meaning that a large fraction of the coefficients in $\boldsymbol{\theta}$ may be set to zero. Stated mathematically, for most natural images $||\boldsymbol{\theta} - \boldsymbol{\theta}_m||_{\ell_2}$ is proportional to $(m+1)^{-1/2}$, where $\boldsymbol{\theta}_m$ corresponds to $\boldsymbol{\theta}$ with all but the $m$ largest coefficients set to zero [7], [8]. This compressibility property makes it possible to infer $\boldsymbol{\theta}$ based on a small number of projection measurements, assuming that $\boldsymbol{\Phi}$ is designed properly.

Assume only $m$ transform coefficients in $\boldsymbol{\theta}$ are significant, and the other $M - m$ coefficients are negligibly small. We rewrite $\boldsymbol{\theta} = \boldsymbol{\theta}_m + \boldsymbol{\theta}_e$, where $\boldsymbol{\theta}_m$ represents the original $\boldsymbol{\theta}$ with the $M - m$ smallest coefficients set to zero, and $\boldsymbol{\theta}_e$ represents $\boldsymbol{\theta}$ with the largest $m$ coefficients set to zero. We therefore have

$$\boldsymbol{v} = \boldsymbol{\Phi}\boldsymbol{\theta} = \boldsymbol{\Phi}\boldsymbol{\theta}_m + \boldsymbol{\Phi}\boldsymbol{\theta}_e = \boldsymbol{\Phi}\boldsymbol{\theta}_m + \boldsymbol{n}_e, \tag{3}$$

where $\boldsymbol{n}_e = \boldsymbol{\Phi}\boldsymbol{\theta}_e$. According to Section II, each element of $\boldsymbol{\theta}_e$ can be modeled by a zero-mean Gaussian with small variance (as being drawn from a "low" state), and thus each element of $\boldsymbol{n}_e$, which is a linear combination of elements in $\boldsymbol{\theta}_e$, can also be modeled by a zero-mean Gaussian with appropriate variance. Further, if we also assume the CS measurements are noisy, with zero-mean Gaussian noise $\boldsymbol{n}_0$, we have

$$\boldsymbol{v} = \boldsymbol{\Phi}\boldsymbol{\theta}_m + \boldsymbol{n}_e + \boldsymbol{n}_0 = \boldsymbol{\Phi}\boldsymbol{\theta}_m + \boldsymbol{n}, \tag{4}$$

where the elements of $\boldsymbol{n}$ can be represented by a zero-mean Gaussian noise with unknown variance $\sigma^2$, or unknown precision $\alpha_n = \sigma^{-2}$ (to be inferred in the CS inversion).

For the wavelet-based CS reconstruction problem, given measurements $\boldsymbol{v}$ and the random projection matrix $\boldsymbol{\Phi}$, the objective is to estimate the values and the locations of the nonzero elements in the transform coefficients $\boldsymbol{\theta}_m$. For simplicity we henceforth use $\boldsymbol{\theta}$ to replace $\boldsymbol{\theta}_m$ in (4), with the understanding that $\boldsymbol{\theta}$ is now sparse (a large fraction of coefficients are exactly zero).

### *B. Tree-Structured Wavelet CS Model*

Baraniuk *et al.* [21] demonstrate that it is possible to improve compressive-sensing reconstruction performance by leveraging signal models (structure within the transform coefficients), by introducing dependencies between locations of the signal coefficients. Two greedy CS algorithms, CoSaMP [13] and iterative hard thresholding (IHT) [38], are implemented in [21], with the wavelet tree structure incorporated into the inversion models.

In this paper the proposed tree-structured wavelet compressive sensing (TSW-CS) model is constructed in a hierarchical Bayesian learning framework. In this setting we infer a full posterior density function on the wavelet coefficients, and therefore we may quantify our confidence in the inversion ($e.g.$, the variance about the mean inverted signal). Within the Bayesian framework we impose a prior belief for the model parameters, represented in terms of prior distributions on the model parameters. The posterior distribution for all model parameters and for the wavelet coefficients are inferred based on the observed data $\boldsymbol{v}$. The structural information embodied by the wavelet tree (the parent-children relationship and the propagation of small coefficients across scales) is incorporated in the prior, and is therefore imposed statistically.

We utilize a spike-and-slab prior, which has been used recently in Bayesian regression and factor models [32], [33], [34], [35], [36]. The prior for the $i$th element of $\boldsymbol{\theta}$ (corresponding to the $i$th transform coefficient) has the form

$$\theta_i \sim (1-\pi_i)\delta_0 + \pi_i \mathcal{N}(0, \alpha_i^{-1}), \quad i = 1, 2, ...M, \tag{5}$$

which is a mixture of two components. The first component $\delta_0$ is a point mass concentrated at zero, and the second component is a zero-mean Gaussian distribution with (relatively small) precision $\alpha_i$; the former represents the zero coefficients in $\boldsymbol{\theta}$ and the latter the non-zero coefficients. This is a two-component mixture model, and the two components are associated with the two states in the HMT. Related models of this type have been employed previously for wavelet-based clustering [39]. The form of this model is different from an HMT [5] in that the coefficient associated with the "low" state is now explicitly set to zero, such that the inferred wavelet coefficients are explicitly sparse (many coefficients exactly zero). However, like in the HMT, we impose the belief that if a parent coefficient is zero, its children coefficients are likely to also be zero. To achieve this goal, the key to the model is imposition of dependencies in the $\pi_i$ across scales, in the form discussed above.

The mixing weight $\pi_i$, the precision parameter $\alpha_i$, as well as the unknown noise precision $\alpha_n$, are learned from the data. The proposed Bayesian tree-structured wavelet (TSW) CS model is summarized as follows:

$$\boldsymbol{v}|\boldsymbol{\theta},\alpha_n \sim \mathcal{N}(\boldsymbol{\Phi\theta},\alpha_n^{-1}\boldsymbol{I}), \tag{6a}$$

$$\theta_{s,i} \sim (1-\pi_{s,i})\delta_0+\pi_{s,i}\mathcal{N}(0,\alpha_s^{-1}),\ \text{ with } \pi_{s,i}=\begin{cases}\pi_r, & \text{if } s=1,\\ \pi_s^0, & \text{if } 2\le s\le L, \theta_{pa(s,i)}=0,\\ \pi_s^1, & \text{if } 2\le s\le L, \theta_{pa(s,i)}\neq 0,\end{cases} \tag{6b}$$

$$\alpha_n \sim \text{Gamma}(a_0,b_0), \tag{6c}$$

$$\alpha_s \sim \text{Gamma}(c_0,d_0), \quad s=1,...,L, \tag{6d}$$

$$\pi_r \sim \text{Beta}(e_0^r,f_0^r), \tag{6e}$$

$$\pi_s^0 \sim \text{Beta}(e_0^{s0},f_0^{s0}), \quad s=2,...,L, \tag{6f}$$

$$\pi_s^1 \sim \text{Beta}(e_0^{s1},f_0^{s1}), \quad s=2,...,L, \tag{6g}$$

where $\theta_{s,i}$ denotes the $i$th wavelet coefficient (corresponding to the spatial location) at scale $s$, for $i=1,...,M_s$ ($M_s$ is the total number of wavelet coefficients at scale $s$), $\pi_{s,i}$ is the associated mixing weight, and $\theta_{pa(s,i)}$ denotes the parent coefficient of $\theta_{s,i}$. In (6b) it is assumed that all the nonzero coefficients at scale $s$ share a common precision parameter $\alpha_s$. It is also assumed that all the coefficients at scale $s$ with a zero-valued parent share a common mixing weight $\pi_s^0$, and the coefficients at scale $s$ with a nonzero parent share a mixing weight $\pi_s^1$. We may also let each coefficient maintain its own $\pi_{s,i}$, but we found from the experiments that the performance is very similar to that from the model presented in (6) (sharing common $\pi_s^0$ and $\pi_s^1$ for each scale), and (6) is much simpler because there are less parameters in the model. Gamma priors are placed on the noise precision parameter $\alpha_n$ and the nonzero coefficient precision parameter $\alpha_s$, and the posteriors of these precisions are inferred from the data. The mixing weights $\pi_r, \pi_s^0$ and $\pi_s^1$ are also inferred, by placing Beta priors on them. To impose the structural information, depending on the scale and the parent value of the coefficients, different Beta priors are imposed. For the coefficients at the root node, a prior preferring a value close to one is set in (6e), because at the low-resolution level many wavelet coefficients are nonzero; for the coefficients with a zero-valued parent, a prior preferring zero is considered in (6f), to represent the propagation of zero coefficients across scales; finally, (6g) is for the coefficients with a nonzero parent, and hence no particular preference is considered since zero or nonzero values are both possible (the hyperparameters $e_0^{s1}$ and $f_0^{s1}$ impose a uniform prior on $\pi_s^1$). The exact setting of hyper-parameters on the priors is discussed below when presenting results. Note that the model presented in (6) does not include the scaling coefficients (coefficients at scale $s = 0$); in Section III-D we extend the model to also estimate the scaling coefficients.

The prior imposed in (6f) implies that if a parent node is zero, with high (but not unity) probability its children coefficients will also be zero. The form of the model reduces the degrees of freedom *statistically* in the solution space $\mathcal{R}^M$, since we impose the belief that particular forms of wavelet coefficients are more probable. As opposed to the work in [21], we do not make an explicit ("hard") imposition of the structure of the coefficients, but rather impose the structure statistically.

Concerning setting hyperparameters in the proposed TSW-CS algorithm, while there are many parameters, their settings are relatively "standard", and results were found to be quite robust to reasonable variations in these parameters. The same settings were used in all examples considered below, and no tuning was performed. As an example of how these parameters were selected, $a_0$, $b_0$, $c_0$ and $d_0$ have been set in a non-informative manner consistent with related regression models, such as the relevance vector machine (see [40]). Concerning setting the parameters on the Beta distributions, our goal is to impose the belief that if a parent coefficient is zero, then it is likely that its children will also be zero. As indicated above, this is done using parameters like $1/M$, which imposes that this assumption will be violated rarely, with probability linked to the number of coefficients $M$ in the data. Any other reasonable variation of this setting has been found to yield very similar results to those reported here.

The desired structural information is naturally integrated into the proposed TSW-CS model. It can be seen that the two components in the spike-and-slab prior are analogous to the two states in the HMT model, and the zero-mean Gaussian distributions are analogous to the observation functions of the HMT. The transition-probability matrix $P$ at scale $s$ ($s > 1$) in the HMT is now represented by the mixing weights $\pi_s^0$ and $\pi_s^1$, with $P(1,1) = 1 - \pi_s^0, P(1,2) = \pi_s^0, P(2,1) = 1 - \pi_s^1$, and $P(2,2) = \pi_s^1$ (the initial-state distribution is represented by $[1 - \pi_r,\ \pi_r]$). Note that $\pi_s^0$ and $\pi_s^1$ represent the summary of the overall (Markovian) statistical properties for all the wavelet coefficients at scale $s$, while for each particular coefficient $\theta_{s,i}$, an associated posterior of mixing weight, $\tilde{\pi}_{s,i}$, will be inferred (see Section III-C for the inference).

We also note that the HMT has recently been employed explicitly within CS inversion, for wavelet-based CS [25]. In that previous work one must first train an HMT model on representative example data, and then that model is used within the CS inversion. The difficulty of such an approach is that one must have access to training data that is known *a priori* to be appropriate for the CS data under test. By contrast, in the proposed inference engine, in addition to inferring a posterior distribution on the wavelet coefficients, posterior distributions are jointly inferred on the underlying model parameters as well. There is therefore no need for *a priori* training data. In this sense the proposed method infers the wavelet coefficients *and* a statistical model for these coefficients, with the model consistent with the expected statistical structure typically inherent to the wavelet transform.

### C. MCMC Inference

We implement the posterior computation by a Markov chain Monte Carlo (MCMC) method [41] based on Gibbs sampling, where the posterior distribution is approximated by a sufficient number of samples. These samples are collected by iteratively drawing each random variable (model parameters and intermediate variables) from its conditional posterior distribution given the most recent values of all the other random variables. The priors of the random variables are set independently as

$$p(\alpha_n, \{\alpha_s\}_{s=1:L}, \pi_r, \{\pi_s^0, \pi_s^1\}_{s=2:L}) =$$
$$\text{Gamma}(a_0, b_0)\left\{\prod_{s=1}^{L}\text{Gamma}(c_0, d_0)\right\}\text{Beta}(e_0^r, f_0^r)\left\{\prod_{s=2}^{L}\text{Beta}(e_0^{s0}, f_0^{s0})\text{Beta}(e_0^{s1}, f_0^{s1})\right\}. \quad (7)$$

Under this setting the priors are conjugate to the likelihoods, and the conditional posteriors used to draw samples can be derived *analytically*. At each MCMC iteration, the samples are drawn from the following conditional posterior distributions:

- $p(\theta_{s,i}|-) = (1-\tilde{\pi}_{s,i})\delta_0 + \tilde{\pi}_{s,i}\mathcal{N}(\tilde{\mu}_{s,i}, \tilde{\alpha}_{s,i}^{-1})$.

  Assume $\theta_{s,i}$ is the $j$th element in the $M$-dimensional vector $\boldsymbol{\theta}$, denoted by $\theta_{(j)}$, then

  $$\begin{aligned} \tilde{\alpha}_{s,i} &= \alpha_s + \alpha_n \Phi_{(j)}^T \Phi_{(j)}, \\ \tilde{\mu}_{s,i} &= \tilde{\alpha}_{s,i}^{-1}\alpha_n \Phi_{(j)}^T \tilde{\boldsymbol{v}}_{(j)}, \ \text{ with } \tilde{\boldsymbol{v}}_{(j)} = \boldsymbol{v} - \sum_{\substack{k=1\\k\neq j}}^{M} \Phi_{(k)}\theta_{(k)}, \\ \frac{\tilde{\pi}_{s,i}}{1-\tilde{\pi}_{s,i}} &= \frac{\pi_{s,i}}{1-\pi_{s,i}} \frac{\mathcal{N}(0|0,\alpha_s^{-1})}{\mathcal{N}(0|\tilde{\mu}_{s,i},\tilde{\alpha}_{s,i})}, \end{aligned}$$

  where $\Phi_{(j)}$ denotes the $j$th column of the $N \times M$ random projection matrix $\boldsymbol{\Phi}$.

- $p(\alpha_s|-) = \text{Gamma}(c_0 + \frac{1}{2}\sum_{i=1}^{M_s} \mathbf{1}(\theta_{s,i} \neq 0),\ d_0 + \frac{1}{2}\sum_{i=1}^{M_s} \theta_{s,i}^2)$.

  where $\mathbf{1}(y)$ denotes an indicator function such that $\mathbf{1}(y) = 1$ if $y$ is true and $0$ otherwise.

- $p(\pi_r|-) = \text{Beta}(e_0^r + \sum_{i=1}^{M_s} \mathbf{1}(\theta_{s,i} \neq 0),\ f_0^r + \sum_{i=1}^{M_s} \mathbf{1}(\theta_{s,i} = 0))$, for $s = 1$.
- $p(\pi_s^0|-) = \text{Beta}(e_0^{s0} + \sum_{i=1}^{M_s} \mathbf{1}(\theta_{s,i} \neq 0, \theta_{pa(s,i)} = 0),\ f_0^{s0} + \sum_{i=1}^{M_s} \mathbf{1}(\theta_{s,i} = 0, \theta_{pa(s,i)} = 0))$, for $2 \leq s \leq L$.
- $p(\pi_s^1|-) = \text{Beta}(e_0^{s1} + \sum_{i=1}^{M_s} \mathbf{1}(\theta_{s,i} \neq 0, \theta_{pa(s,i)} \neq 0),\ f_0^{s1} + \sum_{i=1}^{M_s} \mathbf{1}(\theta_{s,i} = 0, \theta_{pa(s,i)} \neq 0))$, for $2 \leq s \leq L$.
- $p(\alpha_n|-) = \text{Gamma}(a_0 + \frac{N}{2},\ b_0 + \frac{1}{2}(\boldsymbol{v} - \boldsymbol{\Phi\theta})^T(\boldsymbol{v} - \boldsymbol{\Phi\theta}))$.

At each MCMC iteration, $\boldsymbol{\theta}$ can be sampled in a block manner (all of the components of $\boldsymbol{\theta}$ sampled jointly); alternatively, $\theta_{s,i}$ can also be sampled sequentially for all $s$ and $i$. We observed in our experiments that sequential sampling typically achieves faster convergence, *i.e.*, less iterations are required to achieve MCMC convergence compared to block sampling. This is because in block sampling, computing the conditional posterior of $\theta_{(j)}$ uses all the other elements of $\boldsymbol{\theta}$ ($\theta_{(k)}$ for $k \neq j$) from the last MCMC iteration; however, by sequential sampling, computing the conditional posterior of $\theta_{(j)}$ can use $\theta_{(k)}$ for $k < j$ from the current iteration (updated before $\theta_{(j)}$ in the current iteration) and $\theta_{(k)}$ for $k > j$ from the last iteration. We observed fast convergence of this model for the problems considered; typically a burn-in period of 200 iterations is enough for an image of size $128 \times 128$, and the collection period corresponds to 100 samples (typically, in an MCMC analysis of a model of this complexity, one runs thousands of burn-in iterations, with thousands or tens of thousands collection iterations, to yield an accurate representation of the full posterior [39], [35]). It is very unlikely that this small number of MCMC iterations is sufficient to accurately represent the full posterior on all model parameters; however, based on many experiments, the mean wavelet coefficients have been found to provide a good practical CS inversion, and the collection samples also provide useful "error bars" (discussed further below). Figure 2 shows the convergence curve for a $128 \times 128$ image with 5000 measurements, for the example considered in detail in Section IV; we employed the sequential sampling approach in this example, as well as in all results presented below.

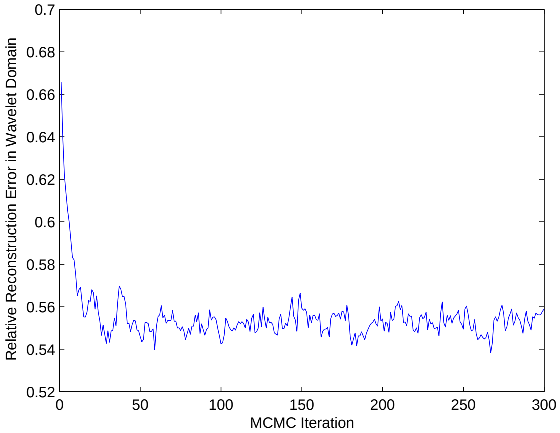


Fig. 2. Example of an MCMC convergence curve for an image of size $128 \times 128$. The vertical axis is evaluated as $||\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}||_2 / ||\boldsymbol{\theta}||_2$, where $\hat{\boldsymbol{\theta}}$ denotes the reconstructed wavelet coefficients.

The variational Bayesian (VB) method [42] is often considered for fast but approximate Bayesian inference. The MCMC solution employed here was found to be relatively accurate and fast, and therefore we did not implement VB inference. Given the fast MCMC convergence for the problems considered, with non-optimized programming in Matlab$^{TM}$, we feel that this may be a practical inference engine for the applications of interest. Specifically, as indicated below, the computational requirements of the TSW-CS model are competitive with many of the existing compressive-sensing inversion algorithms in the literature.

### D. *Extension with Scaling Coefficients*

The TSW-CS model presented in (6) only performs inversion for the wavelet coefficients, assuming that the scaling-function coefficients are measured separately (this has been assumed in many previous compressive-sensing studies [14]). However, it is also of interest in many applications to also infer the scaling coefficients. We may readily extend our TSW-CS model to include reconstruction of the scaling coefficients as follows. Specifically, the model is the same as (6), except with

$$\theta_{s,i} \sim (1-\pi_{s,i})\delta_0 + \pi_{s,i}\mathcal{N}(0,\alpha_s^{-1}), \text{ with } \pi_{s,i} = \begin{cases} \pi_{sc}, & \text{if } s=0, \\ \pi_r, & \text{if } s=1, \\ \pi_s^0, & \text{if } 2 \leq s \leq L, \theta_{pa(s,i)} = 0, \\ \pi_s^1, & \text{if } 2 \leq s \leq L, \theta_{pa(s,i)} \neq 0, \end{cases} \tag{8a}$$

$$\pi_{sc} \sim \text{Beta}(e_0^{sc}, f_0^{sc}). \tag{8b}$$

Compared to (6), the extended model in (8) includes the scale $s=0$ for the scaling coefficients, with an associated mixing weight $\pi_{sc}$, which is drawn from a prior distribution Beta$(e_0^{sc}, f_0^{sc})$. Considering that the scaling coefficients are usually nonzero, the hyperparameters $e_0^{sc}$ and $f_0^{sc}$ are specified such that $\pi_{sc}=1$ is almost always true (since $\pi_{sc}$ is only one more parameter, we perform inference on it, but our experience is that it may be set $\pi_{sc}=1$ with minimal change on the results, for the examples considered). All scaling coefficients share a common precision parameter $\alpha_0$, which is learned from the inference.

## IV. Experimental Results

We compared the performance of TSW-CS to seven recently developed CS reconstruction algorithms: basis pursuit (BP) [16], Bayesian compressive sensing (BCS) [15], fast-BCS[1] [15], orthogonal matching pursuit (OMP) [12], stagewise orthogonal matching pursuit (StOMP) [14], Lasso-modified least angle regression (LARS/Lasso) [17], and total variation (TV) [10]. For the TV and BP implementations, we used solvers from the $\ell_1$-*Magic* toolbox[2]; for the OMP, StOMP and LARS/Lasso algorithms, we use the solvers `SolveOMP`, `SolveStOMP`, and `SolveLasso`, respectively, from the *SparseLab* toolbox[3]. The BCS algorithm can be implemented via the relevance vector machines (RVM) [15]; we implemented it using a variational RVM [43]. All software are written in MATLAB$^{TM}$, and run on PCs with 3.6GHz CPU and 4GB memory.

[1] Code at $http://www.ece.duke.edu/\sim shji/BCS.html$

[2] $http://www.acm.caltech.edu/l1magic/$

[3] $http://sparselab.stanford.edu/$

As is often done in such tests, below we assume in all cases that the scaling coefficients are measured separately and are known. However, we note that using the approach discussed in Section III-D, in all examples considered, the TSW-CS results were essentially unchanged when scaling coefficients are inferred as well. The BP results were found to yield errors significantly larger than those associated with the other methods, with much larger CPU times, and therefore they are not explicitly presented in the results below.

The hyperparameters for the priors in the TSW-CS model are as follows: $a_0 = b_0 = c_0 = d_0 = 10^{-6}$, $[e_0^r,\ f_0^r] = [0.9,\ 0.1] \times M_1$, $[e_0^{s0},\ f_0^{s0}] = [\frac{1}{M},\ 1 - \frac{1}{M}] \times M_s$, and $[e_0^{s1},\ f_0^{s1}] = [0.5,\ 0.5] \times M_s$. Note that the form $[u,\ 1-u] \times V$ is used to represent the hyperparameters $e_0$ and $f_0$ in the Beta priors, where $u$ represents the prior mean of the mixing weight $\pi$, and $V$ represents the confidence of the prior (larger $V$ means more confidence). We set as $\frac{1}{M}$ the prior probability of the rare event that a child is not zero given that its parent is zero (recall that $M$ is the total number of estimated wavelet coefficients), and use $M_s$ (number of coefficients at scale $s$) for the confidence so that the strength of the prior is comparable to the likelihood. For the other CS algorithms, default parameters (if required as input arguments) are used. The StOMP algorithm with CFDR thresholding is not stable; consequently, we use the StOMP algorithm with CFAR thresholding, with the false-alarm rate specified as 0.01.

All examples considered below are for $128 \times 128$ images. The scaling coefficients constitute a block of size $8 \times 8$, and we here assume the scaling coefficients are measured directly. Our objective is to estimate the wavelet coefficients of size $128^2 - 8^2 = 16320$, based on a given number of CS measurements. We use the Haar wavelet in all examples considered below (but similar results are found with other wavelet families).

For each CS algorithm we produce a curve of relative reconstruction error as a function of number of measurements $N$. The relative reconstruction error is defined as $\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2 / \|\boldsymbol{x}\|_2$, where $\boldsymbol{x}$ is the original image, and $\hat{\boldsymbol{x}}$ is the recovered image based on the wavelet coefficients reconstructed by a particular CS algorithm.

We considered a subset of images from Microsoft Research in Cambridge[4]. Five image classes have been selected: flowers, cows, buildings, urban, and office; ten images were selected at random from each class. The images are depicted in Figure 3. In Figure 4 a comparison is

[4] Available at $http://research.microsoft.com/en-us/projects/objectclassrecognition/$

performed for all algorithms discussed above, using the left-most "flowers" image in Figure 3; for brevity we only plot such a figure for this example image, with results for all other images in Figure 3 quite similar (average results for all images in Figure 3 are tabulated below). Each evaluated point in the curves in Figure 4 is computed based on the average of five trials. For each trial, a random projection matrix $\mathbf{\Phi}$ is generated. For an experiment with $N$ measurements, the $N$ rows of $\mathbf{\Phi}$ are selected at random from an $M \times M$ matrix of DCT basis vectors, with the entry in row $i$ and column $j$ expressed as $C(i)\cos[\pi(i-1)(2j-1)/2M]$, where $C(i)$ is a normalization constant such that $C(i) = 1/\sqrt{M}$ for $i = 1$ and $C(i) = \sqrt{2/M}$ for $2 \leq i \leq M$. Note that the indexes of the row selection and the random permutation are stored, but the $\mathbf{\Phi}$ matrix itself is not stored.

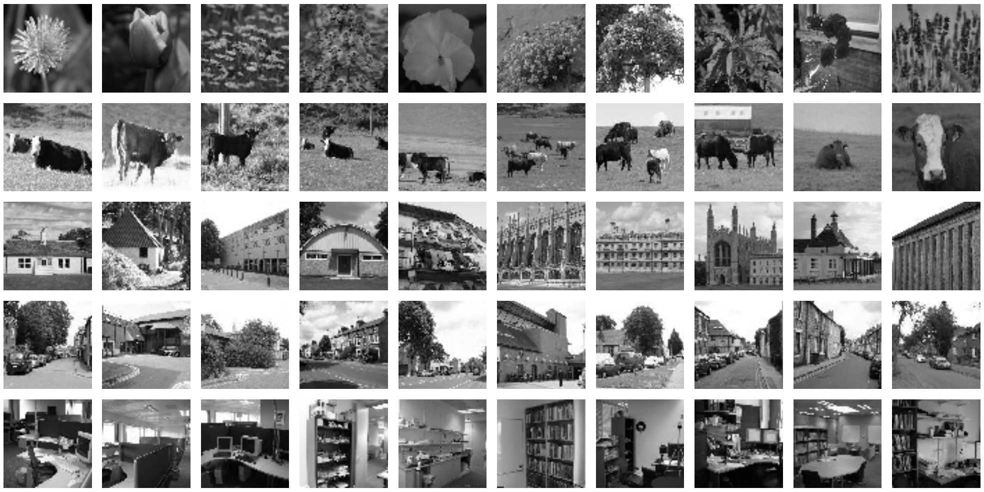
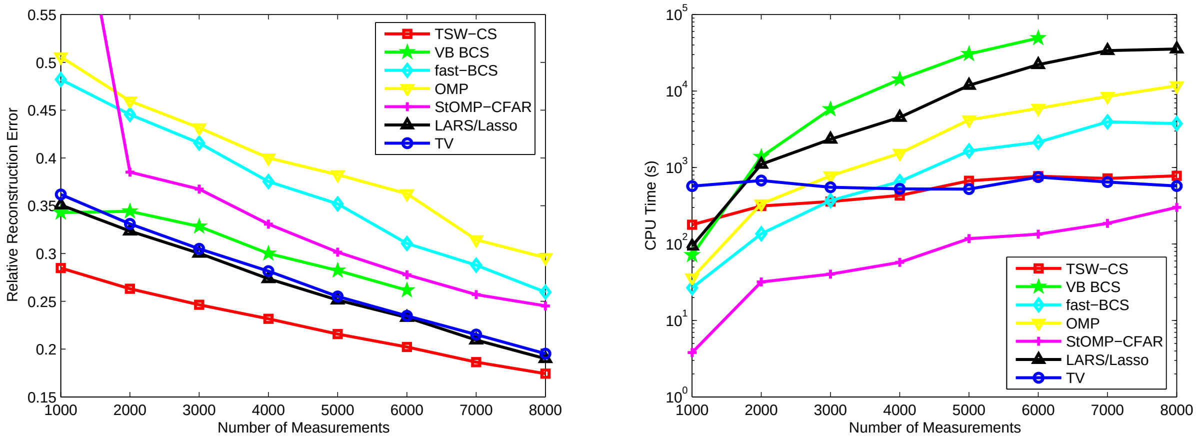


Fig. 3. Natural images from five classes. Each row represents one class, with ten images selected randomly from each class. From top to bottom, the five classes are flowers, cows, buildings, urban, and office.

In Table I is presented average performance of all algorithms, based on *all* of the images in Figure 3, for $N = 2000$ and $N = 6000$ CS measurements. Based on Figure 4, $N = 2000$ corresponds to a relatively small number of measurements and $N = 6000$ a large number of measurements. The utility of exploiting prior knowledge about the structure of the wavelet coefficients is particularly valuable with a small number of CS measurements, as predicted by the theory in [21].

To provide a better feel for why accounting for structure is valuable in the CS inversion, Figure 5 presents an example of the reconstructed wavelet coefficients $\hat{\boldsymbol{\theta}}$ for all CS algorithms



(a) Reconstruction accuracy

(b) CPU time

Fig. 4. Performance comparisons for the left-most "flowers" image in Figure 3. (a) Relative reconstruction error as a function of number of CS measurements, (b) associated CPU time.

(a) Number of measurements N=2000

| Algorithm / Class | | TSW-CS | VB BCS | fast-BCS | OMP | StOMP-CFAR | LARS/ Lasso | TV |
|---|---|---|---|---|---|---|---|---|
| Flowers | MEAN | **0.2296** | 0.3075 | 0.3921 | 0.4082 | 0.3465 | 0.2870 | 0.2908 |
| | STD | **0.0736** | 0.0914 | 0.1251 | 0.1313 | 0.1038 | 0.0868 | 0.0840 |
| Cows | MEAN | **0.1631** | 0.2162 | 0.2663 | 0.2776 | 0.2391 | 0.1988 | 0.2053 |
| | STD | **0.0556** | 0.0683 | 0.0920 | 0.0963 | 0.0758 | 0.0649 | 0.0609 |
| Buildings | MEAN | **0.2178** | 0.2814 | 0.3559 | 0.3723 | 0.3173 | 0.2618 | 0.2655 |
| | STD | **0.0606** | 0.0778 | 0.1031 | 0.1063 | 0.0878 | 0.0735 | 0.0722 |
| Urban | MEAN | **0.2003** | 0.2556 | 0.3254 | 0.3389 | 0.2885 | 0.2377 | 0.2463 |
| | STD | **0.0252** | 0.0341 | 0.0475 | 0.0492 | 0.0376 | 0.0321 | 0.0309 |
| Office | MEAN | **0.2360** | 0.3164 | 0.3969 | 0.4187 | 0.3553 | 0.2920 | 0.2958 |
| | STD | **0.0448** | 0.0642 | 0.0780 | 0.0842 | 0.0708 | 0.0592 | 0.0579 |

(b) Number of measurements N=6000

| Algorithm / Class | | TSW-CS | VB BCS | fast-BCS | OMP | StOMP-CFAR | LARS/ Lasso | TV |
|---|---|---|---|---|---|---|---|---|
| Flowers | MEAN | **0.1616** | 0.2120 | 0.2478 | 0.2794 | 0.2264 | 0.1874 | 0.1927 |
| | STD | 0.0717 | 0.0874 | 0.1102 | 0.1209 | 0.0941 | 0.0774 | **0.0708** |
| Cows | MEAN | **0.1082** | 0.1414 | 0.1556 | 0.1799 | 0.1454 | 0.1233 | 0.1356 |
| | STD | 0.0499 | 0.0605 | 0.0717 | 0.0810 | 0.0639 | 0.0541 | **0.0456** |
| Buildings | MEAN | **0.1475** | 0.1903 | 0.2187 | 0.2465 | 0.2014 | 0.1668 | 0.1729 |
| | STD | 0.0578 | 0.0705 | 0.0850 | 0.0954 | 0.0742 | 0.0623 | **0.0575** |
| Urban | MEAN | **0.1334** | 0.1728 | 0.1981 | 0.2238 | 0.1834 | 0.1508 | 0.1600 |
| | STD | 0.0204 | 0.0251 | 0.0305 | 0.0343 | 0.0268 | 00226 | **0.0192** |
| Office | MEAN | **0.1271** | 0.1788 | 0.1937 | 0.2258 | 0.1882 | 0.1573 | 0.1651 |
| | STD | 0.0283 | 0.0378 | 0.0415 | 0.0499 | 0.0378 | 0.0327 | **0.0276** |

TABLE I

MEAN AND STANDARD DEVIATION OF THE RECONSTRUCTION ERROR FOR EACH CLASS. THE BOLD NUMBER REPRESENTS THE BEST AMONG ALL THE CS ALGORITHMS UNDER COMPARISON.

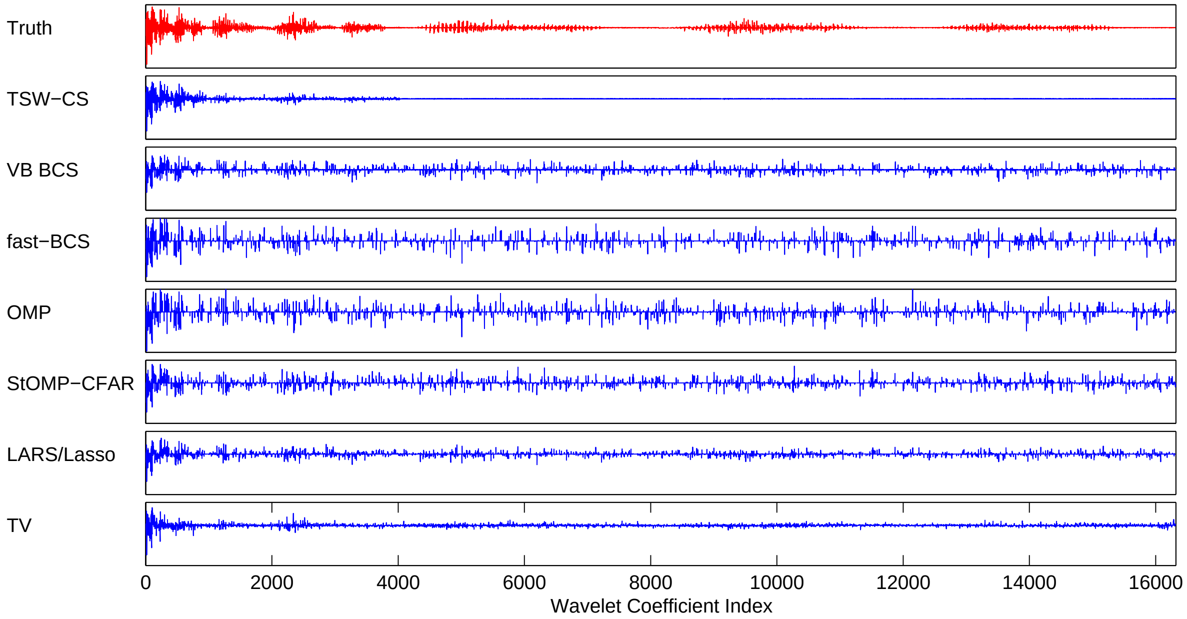
under comparison, for the example in Figure 4 and 2000 measurements. We observe that when the number of measurements is relatively small, the TSW-CS model concentrates more energy on the low-resolution scales, and so estimates the coefficients in the low-resolution bands better. To make this point clearer, Figure 5(b) shows zoom-in plots of the first 960 coefficients at the low-resolution scales $s = 1$ and $s = 2$. When the number of measurements increases, details of an image are then revealed. With the $\delta_0$ component and the parent-child relationships in the prior setting, the TSW-CS model provides a much sparser solution, in the sense of less high frequency noise in the reconstruction compared to the other algorithms, and so estimates large coefficients at higher-resolution scales more accurately. Figure 6 shows the comparisons of the reconstructed $\hat{\boldsymbol{\theta}}$ for the situation of more measurements, for the example in Figure 4 with 6000 measurements. It can be seen that in the regions of zero or small coefficients in the "truth" (usually at high-resolution scales), the TSW-CS model infers exact zero values, while the other algorithms often infer noisy estimations (small non-zero values). These noisy values impair the accurate estimation of large coefficients at high-resolution scales.

Recall that the TV algorithm also exploits known structure about natural images, and from Figures 5 and 6 we note that the TV results are also relatively good. However, TV does not infer the structured sparseness as well as TSW-CS. Further, TV is not a statistical method, and therefore it does not provide a measure of confidence in the CS inversion, this representing the principal advantage of the Bayesian formulation.

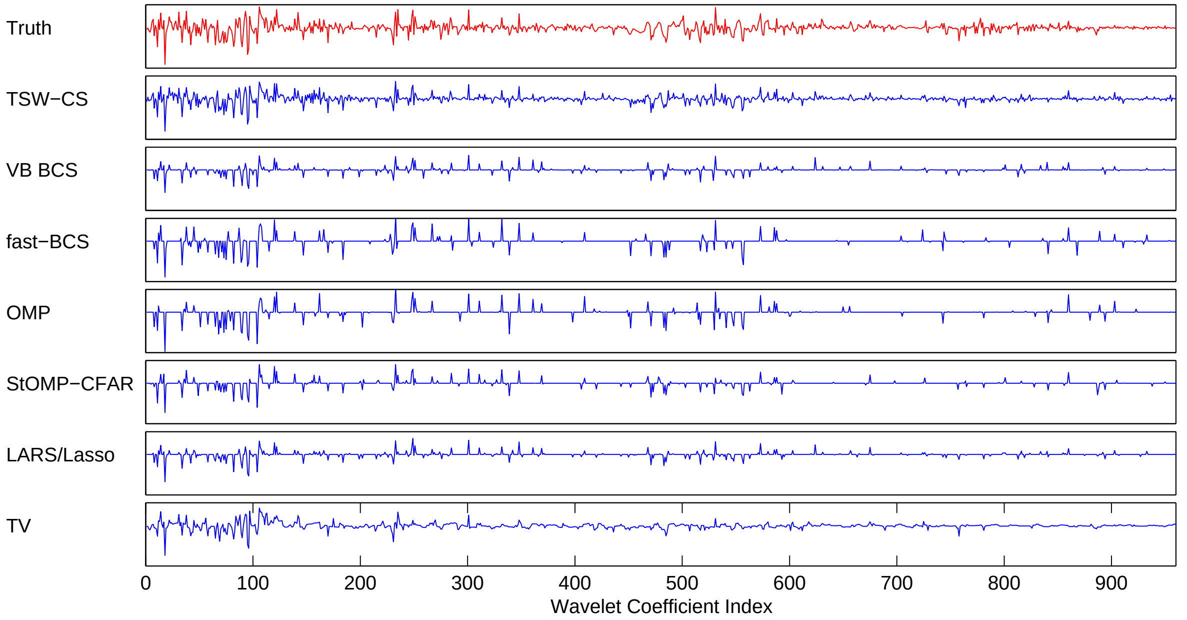
The Bayesian learning framework provided by TSW-CS infers a posterior distribution for the wavelet coefficients (and other model parameters), so it yields "error bars" for the estimated wavelet coefficients, indicating the confidence for the current estimation. This level of confidence may be of interest for placing confidences on inferences made from particular portions of the image. Further, if the TSW-CS may be constituted in fast software or (better) in hardware, it may be fast enough to adaptively infer when a sufficient number of CS measurements have been performed. As an example, Figure 7 plots the error bars of the first 50 estimated wavelet coefficients for the example in Figure 4; this subset of coefficients are selected to make the figure easy to read, with error bars inferred for all coefficients. From Figure 7 one observes that the error bars on the reconstructed wavelet coefficients become tighter (and the reconstructed coefficients approach the "truth") as the number of measurements $N$ increases.

## V. Conclusions and Future Work

A new statistical model has been developed for Bayesian inverse compressive sensing (CS), for situations in which the signal of interest is compressible in a wavelet basis. The formulation explicitly exploits the structure in the wavelet coefficients of typical/natural signals [5], and related structure is exploited in conventional wavelet-based compression algorithms [4]. The advantage of CS, relative to conventional measure-and-then-compress approaches [4], is that the number of (projection) measurements may be significantly smaller than the number of



(a) All wavelet coefficients



(b) Wavelet coefficients at scales $s = 1, 2$

Fig. 5. Comparison of the reconstructed wavelet coefficients by the CS algorithms, for the example in Figure 4 and 2000 measurements. (a) All the wavelet coefficients, $M = 16320$. (b) A zoom-in version of (a), showing the first 960 wavelet coefficients (*i.e.*, coefficients at scales $s = 1, 2$).

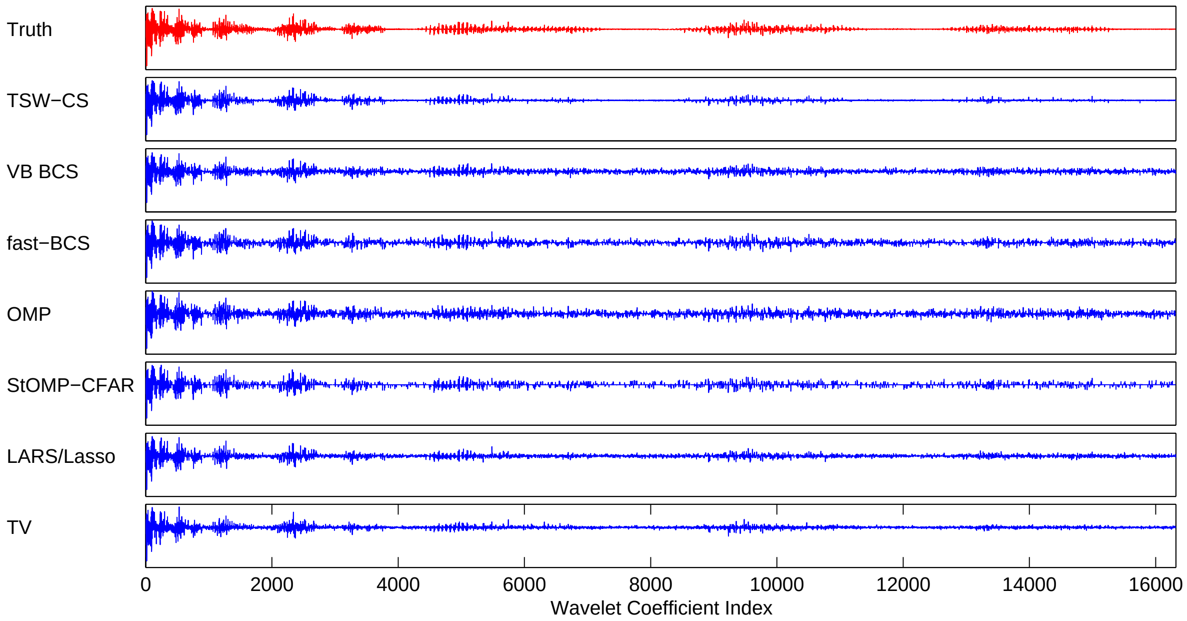


Fig. 6. Comparison of the reconstructed wavelet coefficients by the CS algorithms, for the example in Figure 4 and 6000 measurements.

measurements in traditional sampling methods.

Conventional CS research has assumed that the signal of interest is sparse or compressible in a particular basis ($e.g.$, wavelets), but it assumes no further structure in the transform coefficients. Recent research has demonstrated that if one exploits the structure in the transform coefficients characteristic of typical data or imagery, one often may significantly reduce the number of required CS measurements [18], [21]. In this paper we have assumed the signals of interest are compressible in a wavelet basis. The structure associated with typical wavelet coefficients has been utilized in a statistical setting, building on recent research on Bayesian CS [15].

The proposed method utilizes ideas related to the hidden Markov tree statistical representation of wavelet coefficients [5], and an efficient MCMC inference engine has been constituted. On all examples considered to date, considering real imagery, we have observed very fast convergence of the MCMC algorithm; the inference yields an estimate of the wavelet-transform coefficients as well as “error bars” on the coefficients, reflecting a level of confidence in the inference based on the available CS measurements. In this paper, using a set of canonical images that are widely used in the literature, the proposed method has demonstrated competitive computational cost,

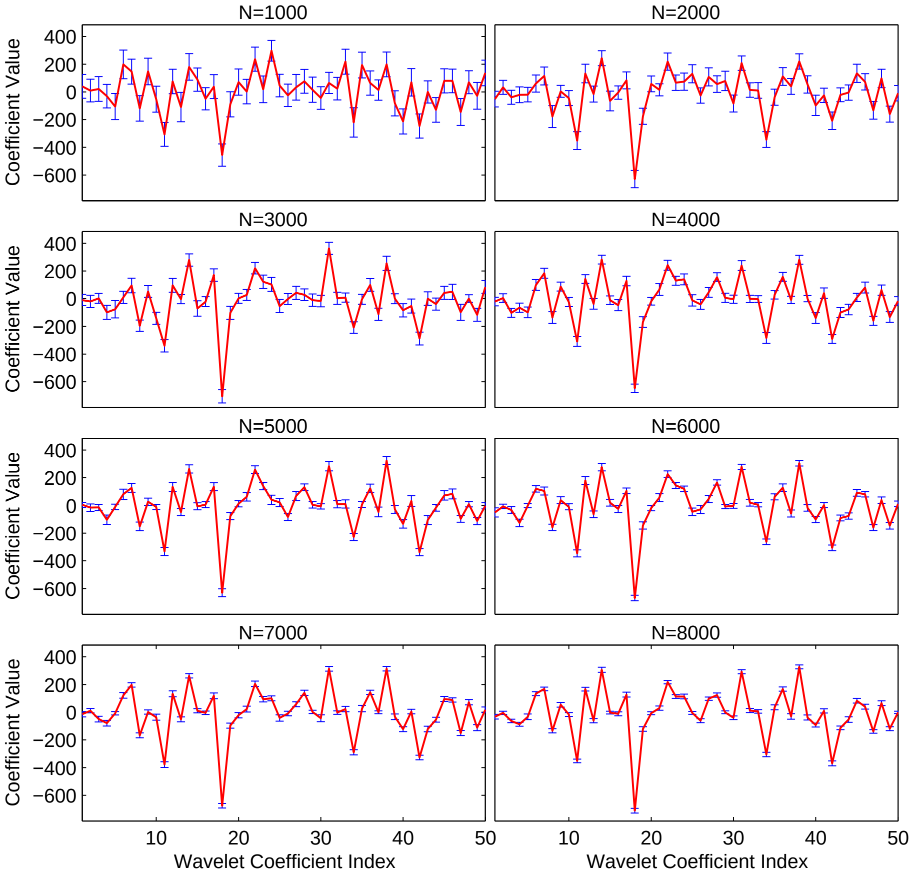


Fig. 7. Error bars of the first 50 estimated wavelet coefficients for the example in Figure 4. The error bars are computed as the standard deviation of the posterior distribution approximated by the MCMC samples for each estimated coefficient.

while consistently providing superior performance, as compared to traditional CS algorithms that do not exploit the structure inherent to the wavelet coefficients.

Concerning future research, there has recently been interest in the simultaneous inversion of multiple distinct CS measurements [44], [45] (by sharing information between these different measurements, the total number of CS measurements may be reduced). The Bayesian setting proposed here is particularly amenable to the joint processing of data from multiple images [46], and this will be investigated in future research. It is also of interest to examine the statistical leveraging of structure in other popular transforms, such as the DCT.

## References

[1] S. Mallat, *A Wavelet Tour of Signal Processing*, 2nd ed. Academic Press, 1998.

[2] G. K. Wallace, "The JPEG still picture compression standard," *IEEE Transactions on Consumer Electronics*, vol. 38, pp. 18–34, 1992.

[3] C. Christopoulos, "JPEG2000 tutorial," in *IEEE International Conference on Image Processing (ICIP)*, Kobe, Japan, 1999. [Online]. Available: http://www.dsp.toronto.edu/ dsp/JPEG2000/

[4] A. Said and W. A. Pearlman, "A new, fast, and efficient image codec based on set partitioning in hierarchical trees," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 6, pp. 243–250, 1996.

[5] M. S. Crouse, R. D. Nowak, and R. G. Baraniuk, "Wavelet-based statistical signal processing using hidden Markov model," *IEEE Transactions on Signal Processing*, vol. 46, pp. 886–902, 1998.

[6] E. Candès and J. Romberg, "Sparsity and incoherence in compressive sampling," *Inverse Problems*, vol. 23, pp. 969–985, 2007.

[7] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, pp. 489–509, 2006.

[8] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, pp. 1289–1306, 2006.

[9] E. J. Candès and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Processing Magazine*, vol. 25, pp. 21–30, 2008.

[10] M. Lustig, D. Donoho, and J. M. Pauly, "Sparse MRI: The application of compressed sensing for rapid MR imaging," *Magnetic Resonance in Medicine*, vol. 58, pp. 1182–1195, 2007.

[11] A. Wagadarikar, R. John, R. Willett, and D. Brady, "Single disperser design for coded aperture snapshot spectral imaging," *Applied Optics*, vol. 47, pp. B44–B51, 2008.

[12] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Transactions on Information Theory*, vol. 53, pp. 4655–4666, 2007.

[13] D. Needell and J. A. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Applied and Computational Harmonic Analysis*, June 2008.

[14] D. L. Donoho, Y. Tsaig, I. Drori, and J.-L. Starck, "Sparse solution of underdetermined linear equations by stagewise orthogonal matching pursuit," *Stanford Statistics Technical Report 2006-2*, April 2006.

[15] S. Ji, Y. Xue, and L. Carin, "Bayesian compressive sensing," *IEEE Transactions on Signal Processing*, vol. 56, pp. 2346–2356, 2008.

[16] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, pp. 33–61, 1999.

[17] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Annals of Statistics (with discussion)*, vol. 32, pp. 407–499, 2004.

[18] T. Blumensath and M. E. Davies, "Sampling theorems for signals from the union of linear subspaces," *IEEE Transactions on Information Theory*, 2008, submitted.

[19] A. Hormati and M. Vetterli, "Annihilating flter-based decoding in the compressed sensing framework," *Proc. SPIE Wavelets XII*, 2007.

[20] Y. C. Eldar and M. Mishali, "Robust recovery of signals from a union of subspaces," *Preprint*, 2008.

[21] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde, "Model-based compressive sensing," *IEEE Transactions on Information Theory*, 2008, submitted.

[22] J. Portilla, V. Strela, M. J. Wainwright, and E. P. Simoncelli, "Image denoising using gaussian scale mixtures in the wavelet domain," *IEEE Transactions on Image Processing*, vol. 12, pp. 1338–1351, 2002.

[23] J. Portilla and E. P. Simoncelli, “A parametric texture model based on joint statistics of complex wavelet coefficients,” *International Journal of Computer Vision*, vol. 40, no. 1, pp. 49–70, 2000.

[24] C. La and M. Do, “Signal reconstruction using sparse tree representations,” *Proc. SPIE Wavelets XI*, 2005.

[25] M. F. Duarte, M. B. Wakin, and R. G. Baraniuk, “Wavelet-domain compressive signal reconstruction using a hidden Markov tree model,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE Press, 2008, pp. 5137–5140.

[26] Y. T. D.L. and Donoho, “Extensions of compressed sensing,” *Signal Processing*, vol. 86, no. 3, pp. 549–571, 2006.

[27] E. J. Cands and J. Romberg, “Practical signal recovery from random projections,” *Wavelet Applications in Signal and Image Processing XI, Proc. SPIE Conf. 5914*, 2004.

[28] J. Duarte-Carvajalino and G. Sapiro, “Learning to sense sparse signals: simultaneous sensing matrix and sparsifying dictionary optimization,” *IMA Preprint Series 2211*, 2008.

[29] G. Peyr, “Best basis compressed sensing,” in *Proc. Int. Conf. Scale Space Variational Methods Computer Vision (SSVM07)*, 2007.

[30] M. Seeger and H. Nickisch, “Compressed sensing and bayesian experimental design,” in *ICML ’08: Proceedings of the 25th international conference on Machine learning*. New York, NY, USA: ACM, 2008, pp. 912–919.

[31] D. Wipf, J. Palmer, and B. Rao, “Perspectives on sparse bayesian learning,” in *Advances in Neural Information Processing Systems 16*. MIT Press, 2004.

[32] H. Ishwaran and J. S. Rao, “Spike and slab variable selection : Frequentist and Bayesian strategies,” *Annals of Statistics*, vol. 33, pp. 730–773, 2005.

[33] E. I. George and R. E. McCulloch, “Variable selection via Gibbs sampling,” *Journal of the American Statistical Association*, vol. 88, pp. 881–889, 1993.

[34] H. Chipman, “Bayesian variable selection with related predictors,” *Canadian Journal of Statistics*, vol. 24, pp. 17–36, 1996.

[35] C. Carvalho, J. Chang, J. Lucas, Q. Wang, J. Nevins, and M. West, “High-dimensional sparse factor modelling: Applications in gene expression genomics,” *Journal of the American Statistical Association*, 2008.

[36] M. West, “Bayesian factor regression models in the “large p, small n” paradigm,” in *Bayesian Statistics 7*, J. M. Bernardo, A. P. Dawid, J. O. Berger, M. West, D. Heckerman, M. J. Bayarri, and A. F. M. Smith, Eds. Oxford University Press, 2003, pp. 723–732.

[37] J. M. Shapiro, “Embedded image coding using zerotrees of wavelet coefficients,” *IEEE Transactions on Signal Processing*, vol. 41, pp. 3445–3462, 1993.

[38] T. Blumensath and M. E. Davies, “Iterative hard thresholding for compressed sensing,” July 2008, preprint.

[39] S. Ray and B. Mallick, “Functional clustering by Bayesian wavelet methods,” *Journal of the Royal Statistical Society. Series B, statistical methodology*, vol. 68, pp. 305–332, 2006.

[40] M. E. Tipping, “Sparse bayesian learning and the relevance vector machine,” *Journal of Machine Learning Research*, pp. 211–244, 2000.

[41] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*, 2nd ed. New York: Springer, 2004.

[42] M. J. Beal, “Variational algorithms for approximate Bayesian inference,” Ph.D. dissertation, Gatsby Computational Neuroscience Unit, University College London, 2003.

[43] C. Bishop and M. Tipping, “Variational relevance vector machines,” in *Proceedings of the 16th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*. San Francisco, CA: Morgan Kaufmann, 2000, pp. 46–53.

[44] D. Baron, M. B. Wakin, M. F. Duarte, S. Sarvotham, and R. G. Baraniuk, "Distributed compressed sensing," in *Thirty-Ninth Asilomar Conference on Signals, Systems and Computers*. IEEE Press, 2005, pp. 1537–1541.

[45] S. Ji, D. Dunson, and L. Carin, "Multi-task compressive sensing," *IEEE Transactions on Signal Processing*, vol. 57, Jan. 2009.

[46] Y. Qi, D. Liu, L. Carin, and D. Dunson, "Multi-task compressive sensing with dirichlet process priors," in *International Conference on Machine Learning (ICML)*, 2008.