Reconstructing Structured Sparse Signals from Compressive Samples via a Max-Product EM Algorithm

Zhao Song and Aleksandar Dogandžić
ECpE Department, Iowa State University
3119 Coover Hall, Ames, IA 50011
{zhaosong,ald}@iastate.edu

Abstract

We propose a Bayesian expectation-maximization (EM) algorithm for reconstructing structured approximately sparse signals via belief propagation. The measurements follow an underdetermined linear model where the regression-coefficient vector is the sum of an unknown approximately sparse signal and a zero-mean white Gaussian noise with an unknown variance. The signal is composed of large- and small-magnitude components identified by binary state variables whose probabilistic dependence structure is described by a hidden Markov tree. Gaussian priors are assigned to the signal coefficients given their state variables and the Jeffreys' noninformative prior is assigned to the noise variance. Our signal reconstruction scheme is based on an EM iteration that aims at maximizing the posterior distribution of the signal and its state variables given the noise variance. We employ a max-product algorithm to implement the maximization (M) step of our EM iteration. The proposed EM algorithm estimates the vector of state variables *as well as* solves iteratively a linear system of equations to obtain the corresponding signal estimate. We select the noise variance so that the corresponding estimated signal and state variables (obtained upon convergence of the EM iteration) have the largest marginal posterior distribution. Our numerical examples show that the proposed algorithm achieves better reconstruction performance compared with the state-of-the-art methods.

Index Terms

Belief propagation, expectation maximization (EM) algorithm, hidden Markov tree (HMT), max-product algorithm, structured sparsity, sparse signal reconstruction.

I. Introduction

The advent of compressive sampling (compressed sensing) in the past few years has sparked research activity in sparse signal reconstruction, whose main goal is to estimate the *sparsest* $p \times 1$ signal coefficient vector $\boldsymbol{s}$ from the $N \times 1$ measurement vector $\boldsymbol{y}$ satisfying the following underdetermined system of linear equations: $\boldsymbol{y} = H\,\boldsymbol{s}$, where H is an $N \times p$ *sensing matrix* and $N \leq p$.

A tree dependency structure is exhibited by the wavelet coefficients of many natural images [1]–[5], see also Fig. 1(a) and [2, Fig. 2]. A probabilistic Markov tree structure has been employed to model the statistical dependency between the state variables of wavelet coefficients [1]. An approximate belief propagation algorithm has been first applied to compressive sampling in [6], which employs sparse Rademacher sensing matrices for Bayesian signal reconstruction. Donoho *et al.* [7] simplified the sum-product algorithm by approximating messages with using a Gaussian distribution specified by two

The material in this manuscript was presented in part at the SPIE Optics+Photonics Symposium, San Diego, CA, August 2012.

scalar parameters, leading to their *approximate message passing (AMP)* algorithm. Following the AMP framework, [8] proposed a *turbo-AMP* structured sparse signal recovery method based on loopy belief propagation and turbo equalization and applied it to reconstruct one-dimensional signals; [5] applied the turbo-AMP approach to reconstruct compressible images. However, the above references do not employ the exact form of the messages and also have the following limitations: Baron *et al.* [6] rely on sparsity of the sensing matrix, the methods by Baron *et al.* [6] and Donoho *et al.* [7] apply to unstructured signals only, and the turbo-AMP approach in [5] and [8] needs columns of the sensing matrix to be normalized, see [5, eq. (22)] and [8, Sec. IV.A].

In this paper, we combine the hierarchical measurement model in [9] with a Markov tree prior on the binary state variables that identify the large- and small-magnitude signal coefficients and develop a Bayesian maximum *a posteriori* (MAP) expectation-maximization (EM) signal reconstruction scheme that aims at maximizing the posterior distribution of the signal and its state variables given the noise variance, where the maximization (M) step employs a max-product belief propagation algorithm. Unlike the previous work, we *do not* approximate the message form in our belief propagation scheme. Unlike the turbo-AMP scheme in [5] and [8], our reconstruction scheme *does not* require the columns of the sensing matrix to be normalized. Since there are no loops in the graphical model behind our M-step objective function, the M step of our EM algorithm is exact. In [10], we proposed a similar EM algorithm for a random signal model [11] with a purely sparse deterministic signal component and a noninformative prior on this component given the binary state variables. We apply a grid search to select the noise variance so that the estimated signal and state variables have the largest marginal posterior distribution.

In Section II, we introduce our measurement and prior models. Section III describes the proposed EM algorithm, where the M step implementation via the max-product algorithm is presented in Section III-A. The selection of the noise variance parameter is discussed in Section IV. Numerical simulations in Section V compare reconstruction performances of the proposed and existing methods.

We introduce the notation: I_n and $\mathbf{0}_{n\times 1}$ denote the identity matrix of size n and the $n \times 1$ vector of zeros, respectively; "^{T}" and $\| \cdot \|_p$ are the transpose and ℓ_p norm, respectively; $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \varSigma)$ denotes the probability distribution function (pdf) of a multivariate Gaussian random vector $\boldsymbol{x}$ with mean $\boldsymbol{\mu}$ and covariance matrix $\varSigma$; Inv-$\chi^2(\sigma^2;\, \nu, \sigma_0^2)$ denotes the pdf of a scaled inverse chi-square distribution with ν degrees of freedom and a scale parameter σ_0^2, see [12, p. 50 and App. A]; $|\mathcal{T}|$ is the cardinality of the set $\mathcal{T}$; $\upsilon(\cdot)$ is an invertible operator that transforms the two-dimensional matrix element indices into one-dimensional vector element indices. Finally, ρ_H denotes the largest singular value of a matrix H, also known as the spectral norm of H, and "$\odot$" denotes the Hadamard (elementwise) product.

II. Measurement and Prior Models

We model an $N \times 1$ real-valued measurement vector $\boldsymbol{y}$ using the standard additive white Gaussian noise measurement model with the likelihood function given by the following pdf [2], [5]:

$$p_{\boldsymbol{y} \mid \boldsymbol{s}, \sigma^2}(\boldsymbol{y} \mid \boldsymbol{s}, \sigma^2) = \mathcal{N}(\boldsymbol{y}\,;\, H\,\boldsymbol{s}, \sigma^2\, I_p) \tag{1}$$

where H is an $N \times p$ real-valued sensing matrix with $\operatorname{rank}(H) = N$ satisfying (without loss of generality)

$$\rho_H = 1 \tag{2}$$

$\boldsymbol{s} = [s_1, s_2, \ldots, s_p]^T$ is an unknown $p \times 1$ real-valued signal coefficient vector, and σ^2 is the unknown noise variance.

We adopt the Jeffreys' noninformative prior for the variance component σ^2:

$$p_{\sigma^2}(\sigma^2) \propto (\sigma^2)^{-1}. \tag{3}$$

Define the vector of binary state variables $\boldsymbol{q} = [q_1, q_2, \ldots, q_p]^T \in \{0, 1\}^p$ that determine if the magnitudes of the signal components s_i, $i = 1, 2, \ldots, p$ are small ($q_i = 0$) or large ($q_i = 1$). Assume that s_i are conditionally independent given q_i and assign the following prior pdf to the signal coefficients:

$$p_{\boldsymbol{s} \mid \boldsymbol{q}, \sigma^2}(\boldsymbol{s} \mid \boldsymbol{q},\, \sigma^2) = \prod_{i=1}^{p} [\mathcal{N}(s_i\,;\, 0, \gamma^2\,\sigma^2)]^{q_i}\, [\mathcal{N}(s_i\,;\, 0, \epsilon^2\,\sigma^2)]^{1-q_i} \tag{4a}$$

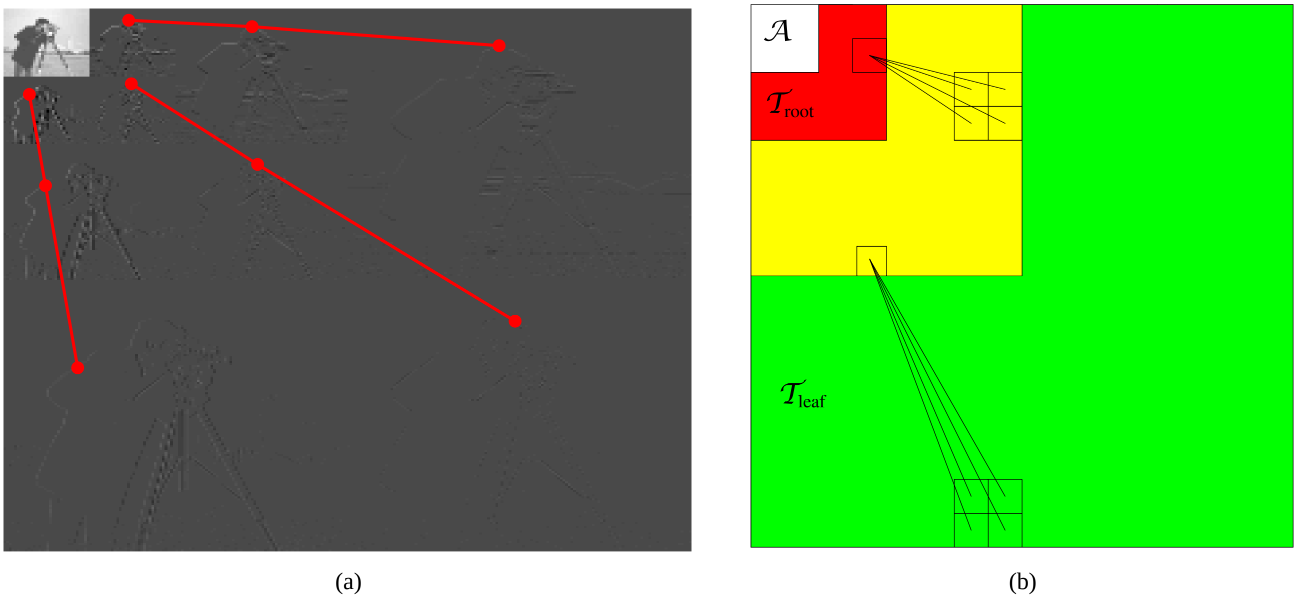
where γ^2 and ϵ^2 are known positive constants and, typically, $\gamma^2 \gg \epsilon^2$. Hence, the large- and small-magnitude signal coefficients s_i corresponding to $q_i = 1$ and $q_i = 0$ are modeled as zero-mean Gaussian random variables with variances $\gamma^2\,\sigma^2$ and $\epsilon^2\,\sigma^2$, respectively. Consequently, γ^2 and ϵ^2 are relative variances (to the noise variance σ^2) of the large- and small-magnitude signal coefficients. Equivalently,

$$p_{\boldsymbol{s} \mid \boldsymbol{q}, \sigma^2}(\boldsymbol{s} \mid \boldsymbol{q},\, \sigma^2) = \mathcal{N}(\boldsymbol{s}\,;\, \boldsymbol{0}_{p \times 1}, \sigma^2\, D(\boldsymbol{q})) \tag{4b}$$

where

$$D(\boldsymbol{q}) = \operatorname{diag}\{(\gamma^2)^{q_1}\,(\epsilon^2)^{1-q_1}, (\gamma^2)^{q_2}\,(\epsilon^2)^{1-q_2}, \ldots, (\gamma^2)^{q_p}\,(\epsilon^2)^{1-q_p}\}. \tag{4c}$$

We now introduce the Markov tree prior probability mass function (pmf) on the state variables q_i [1], [5]. To make this probability model easier to understand, we introduce two-dimensional signal element indices (i_1, i_2). Recall that the conversion operator $v(\cdot)$ is invertible; hence, there is a one-to-one correspondence between the corresponding one- and two-dimensional signal element indices. A parent



(a)

(b)

Fig. 1. (a) Clustering of significant discrete wavelet transform coefficients of a compressed ‘Cameraman’ image and (b) types of wavelet decomposition coefficients: approximation, root, and leaf, whose sets are denoted by $\mathcal{A}, \mathcal{T}_{\text{root}}$, and $\mathcal{T}_{\text{leaf}}$, respectively.

wavelet coefficient with a two-dimensional position index (i_1, i_2) has four children in the finer wavelet decomposition level with two-dimensional indices $(2\,i_1 - 1, 2\,i_2 - 1)$, $(2\,i_1 - 1, 2\,i_2)$, $(2\,i_1, 2\,i_2 - 1)$ and $(2\,i_1, 2\,i_2)$, see Fig. 1(b). The parent-child dependency assumption implies that, if a parent coefficient in a certain wavelet decomposition level has small (large) magnitude, then its children coefficients in the next finer wavelet decomposition level tend to have small (large) magnitude as well. Denote by ρ and c the numbers of rows and columns of the image, and by L the number of wavelet decomposition levels (tree depth).

We set the prior pmf $p_{\boldsymbol{q}}(\boldsymbol{q})$ as follows. In the first wavelet decomposition level ($l = 1$), assign

$$p_{q_i}(1) = \Pr\{q_i = 1\} = \begin{cases} 1, & i \in \mathcal{A} \\ P_{\text{root}}, & i \in \mathcal{T}_{\text{root}} \end{cases} \tag{5a}$$

where

$$\mathcal{A} = \upsilon\big(\{1, 2, \dots, \frac{\rho}{2^L}\} \times \{1, 2, \dots, \frac{c}{2^L}\}\big) \tag{5b}$$

$$\mathcal{T}_{\text{root}} = \upsilon\big(\{1, 2, \dots, \frac{\rho}{2^{L-1}}\} \times \{1, 2, \dots, \frac{c}{2^{L-1}}\}\big) \setminus \mathcal{A} \tag{5c}$$

are the sets of indices of the approximation and root node coefficients and $P_{\text{root}} \in (0, 1)$ is a known constant denoting the prior probability that a root node signal coefficient has large magnitude, see Fig. 1(b). In the levels $l = 2, 3, \dots, L$, assign

$$p_{q_i \,|\, q_{\pi(i)}}(1 \,|\, q_{\pi(i)}) = \begin{cases} P_{\text{H}}, & q_{\pi(i)} = 1 \\ P_{\text{L}}, & q_{\pi(i)} = 0 \end{cases} \tag{5d}$$

where $\pi(i)$ denotes the index of the parent of node i. Here, $P_{\mathrm{H}} \in (0,1)$ and $P_{\mathrm{L}} \in (0,1)$ are known constants denoting the probabilities that the signal coefficient s_i is large if the corresponding parent signal coefficient is large or small, respectively.

Our wavelet tree structure consists of $|\mathcal{T}_{\mathrm{root}}|$ trees and spans all signal wavelet coefficients except the approximation coefficients; hence, the set of indices of the wavelet coefficients within the trees is

$$\mathcal{T} = v\big(\{1,2,\ldots,\rho\} \times \{1,2,\ldots,c\}\big) \setminus \mathcal{A} \tag{5e}$$

Define also the set of leaf variable node indices within the tree structure as

$$\mathcal{T}_{\mathrm{leaf}} = v\big([\{1,2,\ldots,\rho\} \times \{1,2,\ldots,c\}] \setminus [\{1,2,\ldots,\frac{\rho}{2}\} \times \{1,2,\ldots,\frac{c}{2}\}]\big) \tag{5f}$$

see Fig. 1(b). More complex models are possible; see e.g., [3] and [5], which, however, need at least 10 hyperparameters to specify the prior for the same wavelet tree and did not report large-scale examples. Here, we only need 5 tuning parameters P_{root}, P_{H}, P_{L}, γ^2, and ϵ^2, each with a clear meaning. A fairly crude choice of these parameters is sufficient for achieving good reconstruction performance, see Section V.

The logarithm of the prior pmf $p_{\boldsymbol{q}}(\boldsymbol{q})$ is

$$\begin{aligned}
\ln p_{\boldsymbol{q}}(\boldsymbol{q}) = \text{const} &+ \Big[\sum_{i\in\mathcal{A}} \ln \mathbb{1}(q_i = 1)\Big] + \Big[\sum_{i\in\mathcal{T}_{\mathrm{root}}} q_i \, \ln P_{\mathrm{root}} + (1-q_i)\, \ln(1-P_{\mathrm{root}})\Big] \\
&+ \Big[\sum_{i\in\mathcal{T}\setminus\mathcal{T}_{\mathrm{root}}} q_i \, q_{\pi(i)} \, \ln P_{\mathrm{H}} + (1-q_i)\, q_{\pi(i)} \, \ln(1-P_{\mathrm{H}}) \\
&+ q_i \, (1-q_{\pi(i)}) \, \ln P_{\mathrm{L}} + (1-q_i)\,(1-q_{\pi(i)}) \, \ln(1-P_{\mathrm{L}})\Big]
\end{aligned} \tag{5g}$$

where const denotes the terms that are not functions of $\boldsymbol{q}$.

A. Bayesian Inference

Define the vectors of state variables and signal coefficients

$$\boldsymbol{\theta} = \begin{bmatrix}\boldsymbol{\theta}_1^T & \boldsymbol{\theta}_2^T & \cdots & \boldsymbol{\theta}_p^T\end{bmatrix}^T, \quad \boldsymbol{\theta}_i = \begin{bmatrix} q_i, & s_i\end{bmatrix}^T. \tag{6}$$

The joint posterior distribution of $\boldsymbol{\theta}$ and σ^2 is

$$\begin{aligned}
& p_{\boldsymbol{\theta},\sigma^2\,|\,\boldsymbol{y}}(\boldsymbol{\theta},\, \sigma^2 \,|\, \boldsymbol{y}) \propto p_{\boldsymbol{y}\,|\,\boldsymbol{s},\sigma^2}(\boldsymbol{y}\,|\,\boldsymbol{s}, \sigma^2)\, p_{\boldsymbol{s}\,|\,\boldsymbol{q},\sigma^2}(\boldsymbol{s}\,|\,\boldsymbol{q},\, \sigma^2)\, p_{\boldsymbol{q}}(\boldsymbol{q})\, p_{\sigma^2}(\sigma^2) \\
& \propto (\sigma^2)^{-(p+N+2)/2} \, \exp[-0.5\, \|\boldsymbol{y} - H\,\boldsymbol{s}\|_2^2/\sigma^2 - 0.5\, \boldsymbol{s}^T\, D^{-1}(\boldsymbol{q})\, \boldsymbol{s}/\sigma^2] \left(\frac{\epsilon^2}{\gamma^2}\right)^{0.5\,\sum_{i=1}^{p} q_i} p_{\boldsymbol{q}}(\boldsymbol{q})
\end{aligned} \tag{7}$$

which implies

$$p_{\sigma^2\,|\,\boldsymbol{\theta},\boldsymbol{y}}(\sigma^2\,|\,\boldsymbol{\theta},\, \boldsymbol{y}) = \text{Inv-}\chi^2\left(\sigma^2 \,\middle|\, p+N, \frac{\|\boldsymbol{y} - H\,\boldsymbol{s}\|_2^2 + \boldsymbol{s}^T\, D^{-1}(\boldsymbol{q})\,\boldsymbol{s}}{p+N}\right) \tag{8a}$$

$$\begin{aligned}
p_{\boldsymbol{\theta}\,|\,\boldsymbol{y}}(\boldsymbol{\theta}\,|\,\boldsymbol{y}) &= \frac{p_{\boldsymbol{\theta},\sigma^2\,|\,\boldsymbol{y}}(\boldsymbol{\theta},\, \sigma^2\,|\,\boldsymbol{y})}{p_{\sigma^2\,|\,\boldsymbol{\theta},\boldsymbol{y}}(\sigma^2\,|\,\boldsymbol{\theta},\, \boldsymbol{y})} \\
&\propto p_{\boldsymbol{q}}(\boldsymbol{q}) \left(\frac{\epsilon^2}{\gamma^2}\right)^{0.5\,\sum_{i=1}^{p} q_i} \Big/ \Big[\frac{\|\boldsymbol{y} - H\,\boldsymbol{s}\|_2^2 + \boldsymbol{s}^T\, D^{-1}(\boldsymbol{q})\,\boldsymbol{s}}{p+N}\Big]^{(p+N)/2}
\end{aligned} \tag{8b}$$

and

$$p_{\boldsymbol{\theta}\,|\,\sigma^2,\boldsymbol{y}}(\boldsymbol{\theta}\,|\,\sigma^2,\boldsymbol{y}) \propto \exp\Big[-0.5\frac{\|\boldsymbol{y}-H\boldsymbol{s}\|_2^2+\boldsymbol{s}^T\,D^{-1}(\boldsymbol{q})\boldsymbol{s}}{\sigma^2}\Big]\Big(\frac{\epsilon^2}{\gamma^2}\Big)^{0.5\sum_{i=1}^p q_i}\,p_{\boldsymbol{q}}(\boldsymbol{q}). \tag{8c}$$

For a fixed $\boldsymbol{q}$, (8b) is maximized with respect to $\boldsymbol{s}$ at

$$\widehat{\boldsymbol{s}}(\boldsymbol{q}) = D(\boldsymbol{q})\,H^T\,[I_N + H\,D(\boldsymbol{q})\,H^T]^{-1}\,\boldsymbol{y}. \tag{9}$$

which is the Bayesian linear-model minimum mean-square error (MMSE) estimator of $\boldsymbol{s}$ for a given $\boldsymbol{q}$ [13, Theorem 11.1]. As ϵ^2 decreases to zero, $\widehat{\boldsymbol{s}}(\boldsymbol{q})$ becomes more sparse (becoming exactly sparse for $\epsilon^2 = 0$); as ϵ^2 increases, $\widehat{\boldsymbol{s}}(\boldsymbol{q})$ becomes less sparse.

Substituting (9) into (8b) yields the following *concentrated (profile) marginal posterior*:

$$\max_{\boldsymbol{s}} p_{\boldsymbol{\theta}\,|\,\boldsymbol{y}}(\boldsymbol{\theta}\,|\,\boldsymbol{y}) \propto p_{\boldsymbol{q}}(\boldsymbol{q})\,\Big(\frac{\epsilon^2}{\gamma^2}\Big)^{0.5\,\sum_{i=1}^p q_i}\Big/\Big\{\frac{\boldsymbol{y}^T\,[I_N + H\,D(\boldsymbol{q})\,H^T]^{-1}\,\boldsymbol{y}}{p+N}\Big\}^{(p+N)/2} \tag{10}$$

which is a function of the state variables $\boldsymbol{q}$ only.

We wish to maximize (8b) with respect to $\boldsymbol{\theta}$, but cannot perform this task directly. Consequently, we adopt the following indirect approach: We first develop an EM algorithm for maximizing $p_{\boldsymbol{\theta}\,|\,\sigma^2,\boldsymbol{y}}(\boldsymbol{\theta}\,|\,\sigma^2,\boldsymbol{y})$ in (8c) for a given σ^2 (Section III) and then propose a grid search scheme for selecting the best regularization parameter σ^2 so that the estimated signal and state variables have the largest marginal posterior distribution (8b) (Section IV).

III. An EM Algorithm for Maximizing $p_{\boldsymbol{\theta}\,|\,\sigma^2,\boldsymbol{y}}(\boldsymbol{\theta}\,|\,\sigma^2,\boldsymbol{y})$

Motivated by [9, Sec. V.A], we introduce the following hierarchical two-stage model:

$$p_{\boldsymbol{y}\,|\,\boldsymbol{z},\sigma^2}(\boldsymbol{y}\,|\,\boldsymbol{z},\sigma^2) = \mathcal{N}\big(\boldsymbol{y}\,;\,H\,\boldsymbol{z},\sigma^2\,(I_N - H\,H^T)\big) \tag{11a}$$

$$p_{\boldsymbol{z}\,|\,\boldsymbol{s}}(\boldsymbol{z}\,|\,\boldsymbol{s}) = \mathcal{N}(\boldsymbol{z}\,;\,\boldsymbol{s},\sigma^2\,I_p) \tag{11b}$$

where $\boldsymbol{z}$ is an $p \times 1$ vector of *missing data*. Observe that the assumption (2) guarantees that the covariance matrix $\sigma^2\,(I_N - H\,H^T)$ in (11a) is positive semidefinite.

Our EM algorithm for maximizing $p_{\boldsymbol{\theta}\,|\,\sigma^2,\boldsymbol{y}}(\boldsymbol{\theta}\,|\,\sigma^2,\boldsymbol{y})$ in (8c) consists of iterating between the following expectation (E) and maximization (M) steps:

$$\text{E step:}\quad \boldsymbol{z}^{(j)} = [z_1^{(j)}, z_2^{(j)}, \ldots, z_p^{(j)}]^T = \boldsymbol{s}^{(j)} + H^T\,(\boldsymbol{y} - H\,\boldsymbol{s}^{(j)}) \tag{12}$$

and

$$\text{M step:}\ \boldsymbol{\theta}^{(j+1)} = \arg\max_{\boldsymbol{\theta}}\Big\{-0.5\frac{\|\boldsymbol{z}^{(j)}-\boldsymbol{s}\|_2^2+\boldsymbol{s}^T D^{-1}(\boldsymbol{q})\boldsymbol{s}}{\sigma^2} + \ln[p_{\boldsymbol{q}}(\boldsymbol{q})] + 0.5\ln\Big(\frac{\epsilon^2}{\gamma^2}\Big)\sum_{i=1}^p q_i\Big\} \tag{13a}$$

$$= \arg\max_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}\,|\,\sigma^2,\boldsymbol{z}}(\boldsymbol{\theta}\,|\,\sigma^2,\boldsymbol{z}^{(j)}) \tag{13b}$$

where j denotes the iteration index. For any two consecutive iterations j and $j+1$, our EM algorithm ensures that the objective posterior function *does not* decrease [14], i.e.

$$p_{\boldsymbol{\theta}\,|\,\sigma^2,\boldsymbol{y}}(\boldsymbol{\theta}^{(j+1)}\,|\,\sigma^2,\boldsymbol{y}) \geq p_{\boldsymbol{\theta}\,|\,\sigma^2,\boldsymbol{y}}(\boldsymbol{\theta}^{(j)}\,|\,\sigma^2,\boldsymbol{y}). \tag{14}$$

To simplify the notation, we omit the dependence of the iterates $\boldsymbol{\theta}^{(j)}$ on σ^2 in this section. Denote by $\boldsymbol{\theta}^{(+\infty)}$, $\boldsymbol{s}^{(+\infty)}$, and $\boldsymbol{q}^{(+\infty)}$ the estimates of $\boldsymbol{\theta}$, $\boldsymbol{s}$, and $\boldsymbol{q}$ obtained upon convergence of the above EM iteration.

The above EM iteration provides an estimate $\boldsymbol{q}^{(+\infty)}$ of the vector of state variables $\boldsymbol{q}$ *as well as* finds the solution (9) of the underlying linear system to obtain the corresponding signal estimate:

$$\boldsymbol{s}^{(+\infty)} = \widehat{\boldsymbol{s}}(\boldsymbol{q}^{(+\infty)}). \tag{15}$$

As ϵ^2 decreases to zero, $\boldsymbol{s}^{(+\infty)}$ becomes more sparse; as ϵ^2 increases, $\boldsymbol{s}^{(+\infty)}$ becomes less sparse.

Appendix A presents the derivation of the E and M steps in (12) and (13) and the proofs of the monotonicity property (14) and the property (15) of the signal estimate upon convergence.

Note that the M step in (13b) is equivalent to maximizing $p_{\boldsymbol{\theta}\,|\,\sigma^2,\boldsymbol{z}}(\boldsymbol{\theta}\,|\,\sigma^2,\boldsymbol{z})$ for the missing data vector $\boldsymbol{z} = \boldsymbol{z}^{(j)}$. In the following section, we describe efficient maximization of $p_{\boldsymbol{\theta}\,|\,\sigma^2,\boldsymbol{z}}(\boldsymbol{\theta}\,|\,\sigma^2,\boldsymbol{z})$.

A. M Step: Maximizing $p_{\boldsymbol{\theta}\,|\,\sigma^2,\boldsymbol{z}}(\boldsymbol{\theta}\,|\,\sigma^2,\boldsymbol{z})$

Before we proceed, define

$$\widehat{s}_i(0) = \frac{\epsilon^2}{1+\epsilon^2}\, z_i, \quad \widehat{s}_i(1) = \frac{\gamma^2}{1+\gamma^2}\, z_i \tag{16}$$

where we omit the dependence of $\widehat{s}_i(0)$ and $\widehat{s}_i(1)$ on z_i to simplify the notation.

Observe that

$$p_{\boldsymbol{\theta}\,|\,\sigma^2,\boldsymbol{z}}(\boldsymbol{\theta}\,|\,\sigma^2,\boldsymbol{z}) \propto p_{\boldsymbol{\theta}_{\mathcal{A}}\,|\,\sigma^2,\boldsymbol{z}}(\boldsymbol{\theta}_{\mathcal{A}}\,|\,\sigma^2,\boldsymbol{z})\, p_{\boldsymbol{\theta}_{\mathcal{T}}\,|\,\sigma^2,\boldsymbol{z}}(\boldsymbol{\theta}_{\mathcal{T}}\,|\,\sigma^2,\boldsymbol{z}) \tag{17}$$

where $\boldsymbol{\theta}_{\mathcal{A}}$ and $\boldsymbol{\theta}_{\mathcal{T}}$ consist of $\boldsymbol{\theta}_i, i \in \mathcal{A}$ and $\boldsymbol{\theta}_i, i \in \mathcal{T}$, respectively, and

$$p_{\boldsymbol{\theta}_{\mathcal{A}}\,|\,\sigma^2,\boldsymbol{z}}(\boldsymbol{\theta}_{\mathcal{A}}\,|\,\sigma^2,\boldsymbol{z}) \propto \Big\{ \prod_{i\in\mathcal{A}} \mathcal{N}(z_i\,;\, s_i, \sigma^2)\, \mathcal{N}(s_i\,;\, 0, \gamma^2\,\sigma^2)\, \mathbb{1}(q_i = 1) \Big\} \tag{18a}$$

$$p_{\boldsymbol{\theta}_{\mathcal{T}}\,|\,\sigma^2,\boldsymbol{z}}(\boldsymbol{\theta}_{\mathcal{T}}\,|\,\sigma^2,\boldsymbol{z}) \propto \Big\{ \prod_{i\in\mathcal{T}} \mathcal{N}(z_i; s_i, \sigma^2)\, [\mathcal{N}(s_i; 0, \gamma^2\,\sigma^2)]^{q_i}\, [\mathcal{N}(s_i; 0, \epsilon^2\,\sigma^2)]^{1-q_i} \Big\} p_{\boldsymbol{q}_{\mathcal{T}}}(\boldsymbol{q}_{\mathcal{T}}). \tag{18b}$$

Here, (18a) follows from (5a) and (18b) corresponds to the hidden Markov tree (HMT) probabilistic model that contains no loops. Fig. 2 depicts an HMT that is a part of the probabilistic model (18b). Maximizing $p_{\boldsymbol{\theta}_{\mathcal{A}}\,|\,\sigma^2,\boldsymbol{z}}(\boldsymbol{\theta}_{\mathcal{A}}\,|\,\sigma^2,\boldsymbol{z}^{(j)})$ in (18a) with respect to $\boldsymbol{\theta}_i$, $i \in \mathcal{A}$ yields

$$\widehat{\boldsymbol{\theta}}_i = \big[1, \;\; \widehat{s}_i(1)\big]^T, \quad i \in \mathcal{A} \tag{19}$$

where we have used the identity (B1a) in Appendix B.

We now apply the max-product belief propagation algorithm [15]–[17] to each tree in our wavelet tree structure, with the goal to find the mode of $p_{\boldsymbol{\theta}_{\mathcal{T}}\,|\,\sigma^2,\boldsymbol{z}}(\boldsymbol{\theta}_{\mathcal{T}}\,|\,\sigma^2,\boldsymbol{z})$. We represent the HMT probabilistic model for $p_{\boldsymbol{\theta}_{\mathcal{T}}\,|\,\sigma^2,\boldsymbol{z}}(\boldsymbol{\theta}_{\mathcal{T}}\,|\,\sigma^2,\boldsymbol{z})$ via *potential functions* as [see (18b)]

$$p_{\boldsymbol{\theta}_{\mathcal{T}}\,|\,\sigma^2,\boldsymbol{z}}(\boldsymbol{\theta}_{\mathcal{T}}\,|\,\sigma^2,\boldsymbol{z}) \propto \Big[\prod_{i\in\mathcal{T}\backslash\mathcal{T}_{\text{root}}} \psi_i(\boldsymbol{\theta}_i)\, \psi_{i,\pi(i)}(q_i, q_{\pi(i)}) \Big] \Big[\prod_{i\in\mathcal{T}_{\text{root}}} \psi_i(\boldsymbol{\theta}_i) \Big] \tag{20}$$

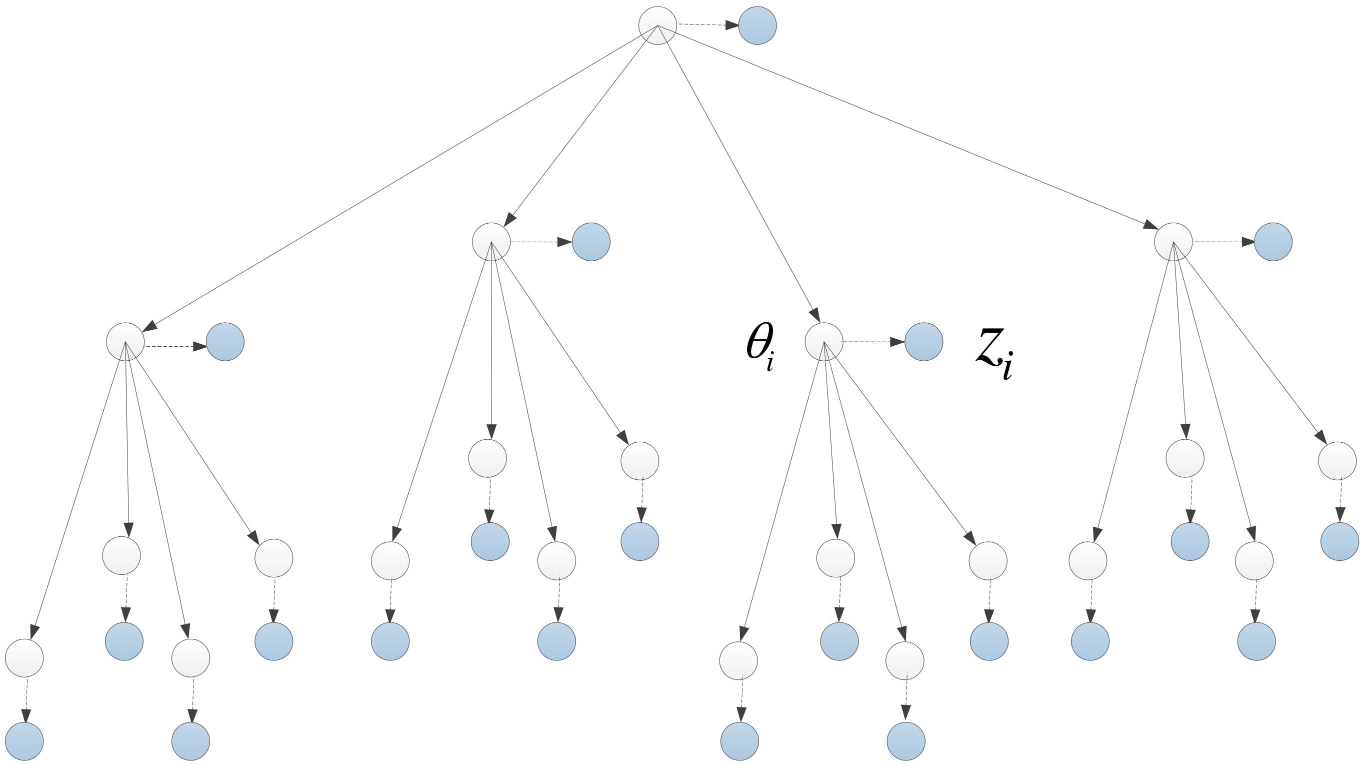


Fig. 2. A hidden Markov tree, part of the probabilistic model (18b).

where

$$\psi_i(\boldsymbol{\theta}_i) = \begin{cases} \mathcal{N}(z_i\,;\, s_i, \sigma^2)\,[\mathcal{N}(s_i\,;\, 0, \gamma^2\,\sigma^2)]^{q_i}\,[\mathcal{N}(s_i\,;\, 0, \epsilon^2\,\sigma^2)]^{1-q_i}, & i \in \mathcal{T}\backslash\mathcal{T}_{\text{root}} \\ \mathcal{N}(z_i\,;\, s_i, \sigma^2)\,[P_{\text{root}}\,\mathcal{N}(s_i\,;\, 0, \gamma^2\,\sigma^2)]^{q_i}\,[(1-P_{\text{root}})\,\mathcal{N}(s_i\,;\, 0, \epsilon^2\,\sigma^2)]^{1-q_i}, & i \in \mathcal{T}_{\text{root}} \end{cases} \tag{21a}$$

and, for $i \in \mathcal{T}\backslash\mathcal{T}_{\text{root}}$,

$$\psi_{i,\pi(i)}(q_i, q_{\pi(i)}) = [{P_{\text{H}}}^{q_i}\,(1-P_{\text{H}})^{1-q_i}]^{q_{\pi(i)}}\,[{P_{\text{L}}}^{q_i}\,(1-P_{\text{L}})^{1-q_i}]^{1-q_{\pi(i)}}. \tag{21b}$$

Our algorithm for maximizing (20) consists of computing and passing upward and downward messages and calculating and maximizing beliefs.

1) Computing and Passing Upward Messages: We propagate the upward messages from the lowest decomposition level (i.e., the leaves) towards the root of the tree. Fig. 3(a) depicts the computation of the upward message from variable node $\boldsymbol{\theta}_i$ to its parent node $\boldsymbol{\theta}_{\pi(i)}$ wherein we also define a *child* of $\boldsymbol{\theta}_i$ as a variable node $\boldsymbol{\theta}_k$ with index $k \in \text{ch}(i)$, where $\text{ch}(i)$ is the index set of the children of i: for $i = \upsilon(i_1, i_2)$, $\text{ch}(i) = \{\upsilon\big((2\,i_1-1, 2\,i_2-1), (2\,i_1-1, 2\,i_2), (2\,i_1, 2\,i_2-1), (2\,i_1, 2\,i_2)\big)\}$. Here, we use a circle and an edge with an arrow to denote a variable node and a message, respectively. The upward messages have the following general form [16]:

$$m_{i\to\pi(i)}(q_{\pi(i)}) = \alpha \max_{\boldsymbol{\theta}_i} \Big\{\psi_i(\boldsymbol{\theta}_i)\,\psi_{i,\pi(i)}(q_i, q_{\pi(i)}) \prod_{k\in\text{ch}(i)} m_{k\to i}(q_i)\Big\} \tag{22}$$

where $\alpha > 0$ denotes a normalizing constant used for computational stability [16]. For nodes that have no children (corresponding to the level L, i.e., $i \in \mathcal{T}_{\text{leaf}}$), we set the multiplicative term $\prod_{k\in\text{ch}(i)} m_{k\to i}(\boldsymbol{\theta}_i)$ in (22) to one.

In Appendix B-I, we show that the only two candidates for $\boldsymbol{\theta}_i$ in the maximization of (22) are $[0, \widehat{s}_i(0)]^T$ and $[1, \widehat{s}_i(1)]^T$, see (16).

Substituting these candidates into (22) and normalizing the messages yields (see Appendix B-I)

$$m_{i\to\pi(i)}(q_{\pi(i)}) = [\mu_i^{\text{u}}(0)]^{1-q_{\pi(i)}}\,[\mu_i^{\text{u}}(1)]^{q_{\pi(i)}} \tag{23a}$$

where $[\mu_i^{\text{u}}(0), \mu_i^{\text{u}}(1)]^T = \boldsymbol{\mu}_i^{\text{u}}$,

$$\begin{aligned}\boldsymbol{\mu}_i^{\text{u}} &= \frac{[\max\{\boldsymbol{\nu}_{0,i}^{\text{u}} \odot \boldsymbol{\eta}_i^{\text{u}}\},\ \max\{\boldsymbol{\nu}_{1,i}^{\text{u}} \odot \boldsymbol{\eta}_i^{\text{u}}\}]^T}{\max\{\boldsymbol{\nu}_{0,i}^{\text{u}} \odot \boldsymbol{\eta}_i^{\text{u}}\} + \max\{\boldsymbol{\nu}_{1,i}^{\text{u}} \odot \boldsymbol{\eta}_i^{\text{u}}\}}\\ &= \frac{\big[\exp(\ln\max\{\boldsymbol{\nu}_{0,i}^{\text{u}} \odot \boldsymbol{\eta}_i^{\text{u}}\} - \ln\max\{\boldsymbol{\nu}_{1,i}^{\text{u}} \odot \boldsymbol{\eta}_i^{\text{u}}\}),\quad 1\big]^T}{1 + \exp(\ln\max\{\boldsymbol{\nu}_{0,i}^{\text{u}} \odot \boldsymbol{\eta}_i^{\text{u}}\} - \ln\max\{\boldsymbol{\nu}_{1,i}^{\text{u}} \odot \boldsymbol{\eta}_i^{\text{u}}\})}\end{aligned} \tag{23b}$$

$$\boldsymbol{\nu}_{0,i}^{\text{u}} = \big[1 - P_{\text{L}},\quad P_{\text{L}}\big]^T \odot \boldsymbol{\phi}(z_i) \tag{23c}$$

$$\boldsymbol{\nu}_{1,i}^{\text{u}} = \big[1 - P_{\text{H}},\quad P_{\text{H}}\big]^T \odot \boldsymbol{\phi}(z_i) \tag{23d}$$

$$\boldsymbol{\eta}_i^{\text{u}} = \begin{cases} \bigodot_{k\in\text{ch}(i)} \boldsymbol{\mu}_k^{\text{u}}, & i \in \mathcal{T}\backslash\mathcal{T}_{\text{leaf}} \\ \big[1,\quad 1\big]^T, & i \in \mathcal{T}_{\text{leaf}} \end{cases} \tag{23e}$$

$$\boldsymbol{\phi}(z) = \Big[\exp(-0.5\,\tfrac{z^2}{\sigma^2+\sigma^2\epsilon^2})/\epsilon,\quad \exp(-0.5\,\tfrac{z^2}{\sigma^2+\sigma^2\gamma^2})/\gamma\Big]^T \tag{23f}$$

and $\epsilon = \sqrt{\epsilon^2} > 0$ and $\gamma = \sqrt{\gamma^2} > 0$. A numerically stable implementation of (23b) that we employ is illustrated in the second expression in (23b). Similarly, the elementwise products in (23c)–(23e) are implemented as exponentiated sums of logarithms of the product terms.

2) Computing and Passing Downward Messages: Upon obtaining all the upward messages, we now compute the downward messages and propagate them from the root towards the lowest level (i.e., the leaves). Fig. 3(b) depicts the computation of the downward message from the parent $\boldsymbol{\theta}_{\pi(i)}$ to the variable node $\boldsymbol{\theta}_i$, which involves upward messages to $\boldsymbol{\theta}_{\pi(i)}$ from its other children, i.e. the *siblings* of $\boldsymbol{\theta}_i$, marked as $\boldsymbol{\theta}_k$, $k \in \text{sib}(i)$. This downward message also requires the message sent to $\boldsymbol{\theta}_{\pi(i)}$ from its parent node, which is the *grandparent* of $\boldsymbol{\theta}_i$, denoted by $\boldsymbol{\theta}_{\text{gp}(i)}$. The downward messages have the following general form [16]:

$$m_{\pi(i)\to i}(q_i) = \alpha \max_{\boldsymbol{\theta}_{\pi(i)}} \Big\{\psi_{\pi(i)}(\boldsymbol{\theta}_{\pi(i)})\,\psi_{i,\pi(i)}(q_i, q_{\pi(i)})\,m_{\text{gp}(i)\to\pi(i)}(q_{\pi(i)}) \prod_{k\in\text{sib}(i)} m_{k\to\pi(i)}(q_{\pi(i)})\Big\} \tag{24}$$

where $\alpha > 0$ denotes a normalizing constant used for computational stability. For the variable nodes i in the second decomposition level that have no grandparents (i.e., $\pi(i) \in \mathcal{T}_{\text{root}}$), we set the multiplicative term $m_{\text{gp}(i)\to\pi(i)}(q_{\pi(i)})$ in (24) to one.

In Appendix B-II, we show that the only two candidates for $\boldsymbol{\theta}_{\pi(i)}$ in the maximization of (24) are $[0, \widehat{s}_{\pi(i)}(0)]^T$ and $[1, \widehat{s}_{\pi(i)}(1)]^T$, see also (16). Substituting these candidates into (24) and normalizing the messages yields (see Appendix B-II)

$$m_{\pi(i)\to i}(q_i) = [\mu_i^{\text{d}}(0)]^{1-q_i}\,[\mu_i^{\text{d}}(1)]^{q_i} \tag{25a}$$

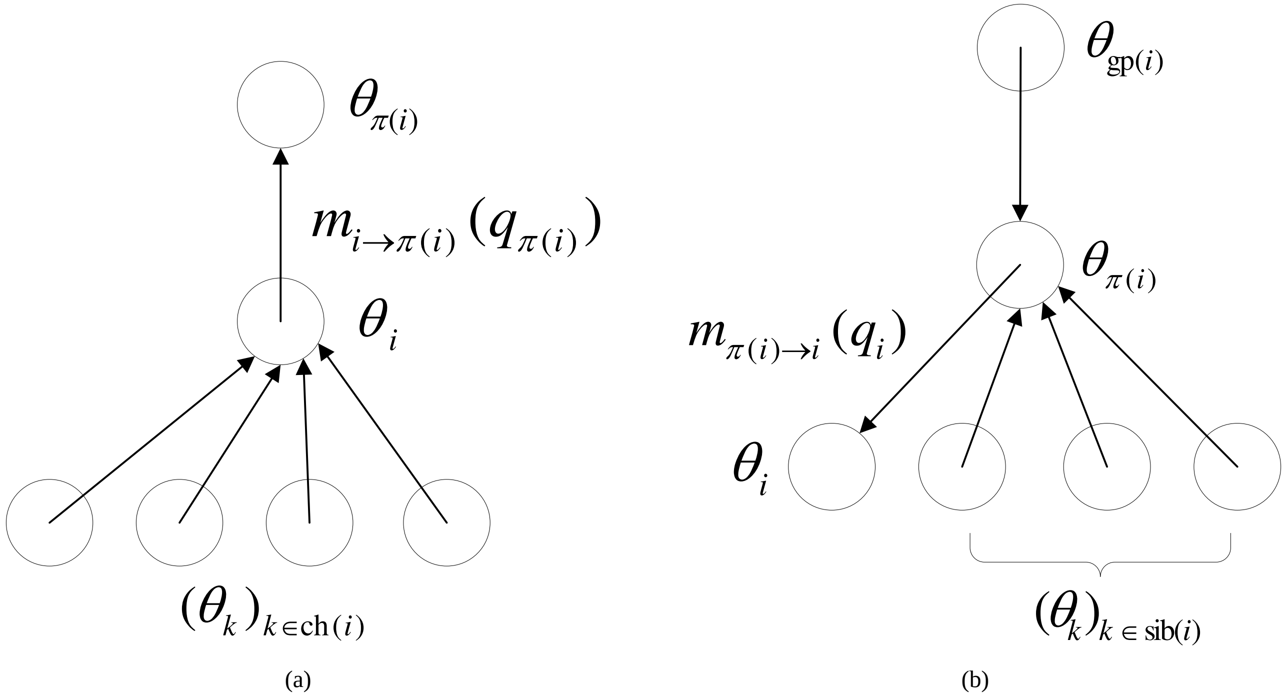

$\theta_{\pi(i)}$
$m_{i\to\pi(i)}(q_{\pi(i)})$
θ_i
$(\theta_k)_{k\in\mathrm{ch}(i)}$
(a)
$\theta_{\mathrm{gp}(i)}$
$\theta_{\pi(i)}$
$m_{\pi(i)\to i}(q_i)$
θ_i
$(\theta_k)_{k\in\mathrm{sib}(i)}$
(b)

Fig. 3. Computing and passing (a) upward and (b) downward messages.

for $\pi(i) \in \mathcal{T}\backslash\mathcal{T}_{\text{leaf}}$, where $[\mu_i^{\mathrm{d}}(0), \mu_i^{\mathrm{d}}(1)]^T = \boldsymbol{\mu}_i^{\mathrm{d}}$ and

$$\begin{aligned}
\boldsymbol{\mu}_i^{\mathrm{d}} &= \frac{[\max\{\boldsymbol{\nu}_{0,i}^{\mathrm{d}} \odot \boldsymbol{\eta}_i^{\mathrm{d}}\},\, \max\{\boldsymbol{\nu}_{1,i}^{\mathrm{d}} \odot \boldsymbol{\eta}_i^{\mathrm{d}}\}]^T}{\max\{\boldsymbol{\nu}_{0,i}^{\mathrm{d}} \odot \boldsymbol{\eta}_i^{\mathrm{d}}\} + \max\{\boldsymbol{\nu}_{1,i}^{\mathrm{d}} \odot \boldsymbol{\eta}_i^{\mathrm{d}}\}} \\
&= \frac{\big[\exp(\ln\max\{\boldsymbol{\nu}_{0,i}^{\mathrm{d}} \odot \boldsymbol{\eta}_i^{\mathrm{d}}\} - \ln\max\{\boldsymbol{\nu}_{1,i}^{\mathrm{d}} \odot \boldsymbol{\eta}_i^{\mathrm{d}}\}), \quad 1\big]^T}{1 + \exp(\ln\max\{\boldsymbol{\nu}_{0,i}^{\mathrm{d}} \odot \boldsymbol{\eta}_i^{\mathrm{d}}\} - \ln\max\{\boldsymbol{\nu}_{1,i}^{\mathrm{d}} \odot \boldsymbol{\eta}_i^{\mathrm{d}}\})} &\text{(25b)} \\
\boldsymbol{\nu}_{0,i}^{\mathrm{d}} &= \big[1 - P_{\mathrm{L}}, \quad 1 - P_{\mathrm{H}}\big]^T \odot \boldsymbol{\phi}(z_{\pi(i)}) \odot \big[\bigodot_{k\in\mathrm{sib}(i)} \boldsymbol{\mu}_k^{\mathrm{u}}\big] &\text{(25c)} \\
\boldsymbol{\nu}_{1,i}^{\mathrm{d}} &= \big[P_{\mathrm{L}}, \quad P_{\mathrm{H}}\big]^T \odot \boldsymbol{\phi}(z_{\pi(i)}) \odot \big[\bigodot_{k\in\mathrm{sib}(i)} \boldsymbol{\mu}_k^{\mathrm{u}}\big] &\text{(25d)} \\
\boldsymbol{\eta}_i^{\mathrm{d}} &= \begin{cases} \big[1 - P_{\mathrm{root}}, \quad P_{\mathrm{root}}\big]^T, & \pi(i) \in \mathcal{T}_{\mathrm{root}} \\ \boldsymbol{\mu}_{\pi(i)}^{\mathrm{d}}, & \pi(i) \in (\mathcal{T}\backslash\mathcal{T}_{\mathrm{root}})\backslash\mathcal{T}_{\mathrm{leaf}} \end{cases} . &\text{(25e)}
\end{aligned}$$

A numerically stable implementation of (25b) that we employ is illustrated in the second expression in (25b).

The above upward and downward messages have discrete representations, which is practically important and is a consequence of the fact that we use a Gaussian prior on the signal coefficients, see (4). Indeed, in contrast with the existing message passing algorithms for compressive sampling [5]–[8], our max-product scheme employs *exact* messages.

3) Maximizing Beliefs: Upon computing and passing all the upward and downward messages, we maximize the beliefs, which have the following general form [16]:

$$b(\boldsymbol{\theta}_i) = \alpha\, \psi_i(\boldsymbol{\theta}_i)\, m_{\pi(i)\to i}(q_i) \prod_{k\in \mathrm{ch}(i)} m_{k\to i}(q_i) \tag{26}$$

for each $i \in \mathcal{T}$, where $\alpha > 0$ is a normalizing constant. [In (26), we set $m_{\pi(i)\to i}(q_i) = 1$ if $i \in \mathcal{T}_{\mathrm{root}}$ and $\prod_{k\in \mathrm{ch}(i)} m_{k\to i}(q_i) = 1$ if $i \in \mathcal{T}_{\mathrm{leaf}}$.] We then use these beliefs to obtain the mode

$$\widehat{\boldsymbol{\theta}}_{\mathcal{T}} = \arg\max_{\boldsymbol{\theta}_{\mathcal{T}}} p_{\boldsymbol{\theta}_{\mathcal{T}}\,|\,\sigma^2,\,\boldsymbol{z}}(\boldsymbol{\theta}_{\mathcal{T}}\,|\,\sigma^2,\,\boldsymbol{z}) \tag{27}$$

where the elements of $\widehat{\boldsymbol{\theta}}_{\mathcal{T}}$ are [see (16)]

$$\widehat{\boldsymbol{\theta}}_i = \big[\widehat{q}_i,\ \ \widehat{s}_i(\widehat{q}_i)\big]^T = \arg\max_{\boldsymbol{\theta}_i} b(\boldsymbol{\theta}_i) = \begin{cases} \big[1,\ \ \widehat{s}_i(1)\big]^T, & \beta_i(1) \ge \beta_i(0) \\ \big[0,\ \ \widehat{s}_i(0)\big]^T, & \text{otherwise} \end{cases}, \quad i \in \mathcal{T} \tag{28a}$$

and

$$\boldsymbol{\beta}_i = \big[\beta_i(0),\ \ \beta_i(1)\big]^T = \begin{cases} \alpha_1 \big[1 - P_{\mathrm{root}},\ \ P_{\mathrm{root}}\big]^T \odot \boldsymbol{\phi}(z_i) \odot \boldsymbol{\eta}_i^{\mathrm{u}}, & i \in \mathcal{T}_{\mathrm{root}} \\ \alpha_1 \boldsymbol{\phi}(z_i) \odot \boldsymbol{\mu}_i^{\mathrm{d}} \odot \boldsymbol{\eta}_i^{\mathrm{u}}, & i \in \mathcal{T} \setminus \mathcal{T}_{\mathrm{root}} \end{cases}. \tag{28b}$$

Here, $\alpha_1 > 0$ is a normalizing constant. The detailed derivation for the forms of $\widehat{\boldsymbol{\theta}}_i$ and $\boldsymbol{\beta}_i$ in (28) is provided in Appendix B-III.

In summary,

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}\,|\,\sigma^2,\boldsymbol{z}}(\boldsymbol{\theta}\,|\,\sigma^2, \boldsymbol{z}) \tag{29}$$

where $\widehat{\boldsymbol{\theta}} = [\,\widehat{\boldsymbol{\theta}}_1^T\ \widehat{\boldsymbol{\theta}}_2^T \cdots \widehat{\boldsymbol{\theta}}_p^T\,]^T$ and

$$\widehat{\boldsymbol{\theta}}_i = \big[\widehat{q}_i,\ \ \widehat{s}_i(\widehat{q}_i)\big]^T = \begin{cases} \big[1,\ \ \widehat{s}_i(1)\big]^T, & i \in \mathcal{A} \\ \big[1,\ \ \widehat{s}_i(1)\big]^T, & \beta_i(1) \ge \beta_i(0),\ i \in \mathcal{T} \\ \big[0,\ \ \widehat{s}_i(0)\big]^T, & \beta_i(1) < \beta_i(0),\ i \in \mathcal{T} \end{cases} \tag{30}$$

follows by combining (19) and (28a) and we have omitted the dependence of $\widehat{\boldsymbol{\theta}}$ on $\boldsymbol{z}$ and $\widehat{\boldsymbol{\theta}}_i$ on z_i to simplify the notation.

IV. SELECTING σ^2

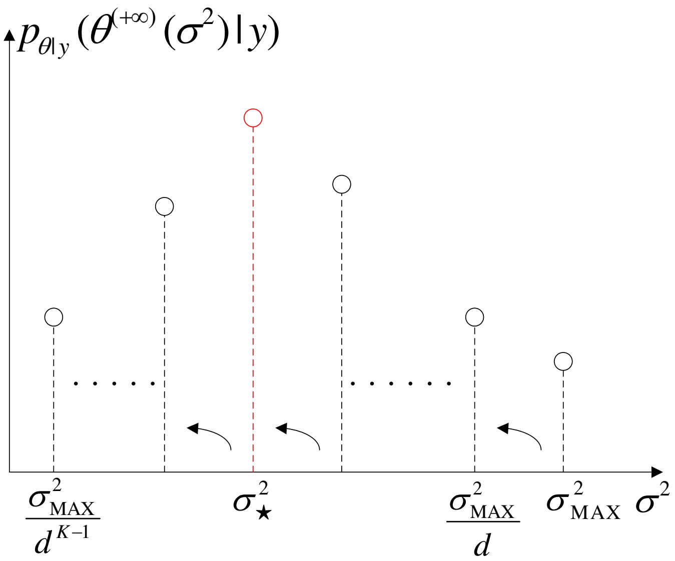
We can integrate σ^2 out, yielding the marginal posterior of $\boldsymbol{\theta}$ in (8b), and derive an ’outer’ EM iteration for maximizing $p_{\boldsymbol{\theta}\,|\,\boldsymbol{y}}(\boldsymbol{\theta}\,|\,\boldsymbol{y})$:

(i) Fix σ^2 and apply the EM iteration proposed in Section III to obtain an estimate $\boldsymbol{\theta}^{(+\infty)}(\sigma^2)$ of $\boldsymbol{\theta}$;

(ii) Fix $\boldsymbol{\theta}$ to the value obtained in (i) and estimate σ^2 as

$$\widehat{\sigma}^2(\boldsymbol{\theta}) = \frac{\|\boldsymbol{y} - H\,\boldsymbol{s}\|_2^2 + \boldsymbol{s}^T D^{-1}(\boldsymbol{q})\,\boldsymbol{s}}{p + N}. \tag{31}$$

Even though it guarantees monotonic increase of the marginal posterior $p_{\boldsymbol{\theta}\,|\,\boldsymbol{y}}(\boldsymbol{\theta}\,|\,\boldsymbol{y})$, the ’outer’ EM iteration (i)–(ii) does not work well in practice because it gets stuck in an undesirable local maximum of



$p_{\theta|y}(\theta^{(+\infty)}(\sigma^2)\,|\,y)$
$\frac{\sigma^2_{\text{MAX}}}{d^{K-1}}$
$\sigma^2_\star$
$\frac{\sigma^2_{\text{MAX}}}{d}$
σ^2_{MAX}
σ^2

Fig. 4. Grid search in selecting σ^2.

$p_{\boldsymbol{\theta}\,|\,\boldsymbol{y}}(\boldsymbol{\theta}\,|\,\boldsymbol{y})$. To find a better (generally local) maximum of $p_{\boldsymbol{\theta}\,|\,\boldsymbol{y}}(\boldsymbol{\theta}\,|\,\boldsymbol{y})$, we apply a grid search over σ^2 as follows.

We apply the EM algorithm in Section III using a range of values of the regularization parameter σ^2. We traverse the grid of K values of σ^2 sequentially and use the signal estimate from the previous grid point to initialize the signal estimation at the current grid point: in particular, we move from a larger σ^2 (say σ^2_{old}) to the next smaller $\sigma^2_{\text{new}}(<\sigma^2_{\text{old}})$ and use $\boldsymbol{s}^{(+\infty)}(\sigma^2_{\text{old}})$ (obtained upon convergence of the EM iteration in Section III for $\sigma^2 = \sigma^2_{\text{old}}$) to initialize the EM iteration at σ^2_{new}. The largest σ^2 on the grid and the initial signal estimate at this grid point are selected as

$$\sigma^2_{\text{MAX}} = \frac{\|\boldsymbol{y}\|_2^2}{p+N}, \qquad \boldsymbol{\theta}^{(0)}(\sigma^2_{\text{MAX}}) = \mathbf{0}_{2p\times 1}. \tag{32a}$$

The consecutive grid points σ^2_{new} and σ^2_{old} satisfy

$$\sigma^2_{\text{new}} = \frac{\sigma^2_{\text{old}}}{d} \tag{32b}$$

where $d > 1$ is a constant determining the search resolution. Finally, we select the σ^2 from the above grid of candidates that yields the largest marginal posterior distribution (8b):

$$\sigma^2_\star = \arg \max_{\sigma^2 \in \{\sigma^2_{\text{MAX}}, \sigma^2_{\text{MAX}}/d, \ldots, \sigma^2_{\text{MAX}}/d^{K-1}\}} p_{\boldsymbol{\theta}\,|\,\boldsymbol{y}}(\boldsymbol{\theta}^{(+\infty)}(\sigma^2)\,|\,\boldsymbol{y}) \tag{33}$$

and the final estimates of $\boldsymbol{\theta}$ and $\boldsymbol{s}$ as $\boldsymbol{\theta}^{(+\infty)}(\sigma^2_\star)$ and $\boldsymbol{s}^{(+\infty)}(\sigma^2_\star)$, respectively, see Fig. 4.

V. Numerical Examples

We compare the reconstruction performances of the following methods:

- our proposed *max-product EM* algorithm in Section III with the variance parameter σ^2 selected using the marginal-posterior based criterion in Section IV (labeled MP-EM), search resolution $d = 2$, and MATLAB implementations available at http://home.eng.iastate.edu/~ald/MPEM.html;
- our max-product EM algorithm in Section III with σ^2 *tuned manually* for good performance (labeled MP-EM$_{\text{OPT}}$) with $d = 2$;
- the turbo-AMP approach [5] with a MATLAB implementation at http://www.ece.osu.edu/~schniter/turboAMPimaging and the tuning parameters chosen as the default values in this implementation;
- the fixed-point continuation active set algorithm [18] (labeled FPC$_{\text{AS}}$) that aims at minimizing the Lagrangian cost function

$$0.5\,\|\boldsymbol{y} - H\,\boldsymbol{s}\|_2^2 + \tau\,\|\boldsymbol{s}\|_1 \tag{34a}$$

with the regularization parameter τ computed as

$$\tau = 10^a\,\|H^T\,\boldsymbol{y}\|_\infty \tag{34b}$$

where a is a tuning parameter chosen manually to achieve good reconstruction performance;
- the Barzilai-Borwein version of the gradient-projection for sparse reconstruction method with debiasing in [19, Sec. III.B] (labeled GPSR) with the convergence threshold $\texttt{tolP} = 10^{-5}$ and tuning parameter a in (34b) chosen manually to achieve good reconstruction performance;
- the double overrelaxation (DORE) thresholding method in [11, Sec. III] or its approximation (DORE$_{\text{app}}$) where the $(H\,H^T)^{-1}$ term is approximated by a diagonal matrix, initialized by the zero sparse signal estimate:

$$\boldsymbol{s}^{(0)} = \boldsymbol{0}_{p\times 1}; \tag{35}$$

- the normalized iterative hard thresholding (NIHT) scheme [20] initialized by the zero $\boldsymbol{s}^{(0)}$ in (35);
- the model-based iterative hard thresholding (MB-IHT) algorithmn [4] using a greedy tree approximation [21], initialized by the zero $\boldsymbol{s}^{(0)}$ in (35).

For the MP-EM, DORE, NIHT, and MB-IHT iterations, we use the following convergence criterion:

$$\frac{\|\boldsymbol{s}^{(j+1)} - \boldsymbol{s}^{(j)}\|_2^2}{p} < \delta \tag{36}$$

where $\delta > 0$ is the convergence threshold selected in the following examples so that the performances of the above methods do not change significantly by further decreasing δ.

The sensing matrix H has the following structure:

$$H = \frac{1}{\rho_\Phi}\,\Phi\,\Psi \tag{37}$$

where Φ is the $N \times p$ sampling matrix and Ψ is the $p \times p$ orthogonal sparsifying transform matrix (satisfying $\Psi\,\Psi^T = I_p$). Note that H in (37) satisfies (2). In the following examples, the sensing matrices Φ are either random Gaussian (Sections V-A and V-B) or structurally random [22] (Section V-C) and the sparsifying transform matrices Ψ are either identity (Section V-A) or inverse Haar wavelet transform matrices (Sections V-B and V-C). We set the tree depth $L = 4$.

A. Small-scale Structured Sparse Signal Reconstruction

We generated the binary state variables $\boldsymbol{q}$ of length $p = 1024$ using the Markov tree model in Section II with $P_{\rm L} = 10^{-5}$. Conditional on q_i, s_i are generated according to (4b). Here, the matrix-to-vector conversion operator $v(\cdot)$ corresponds to simple columnwise conversion. The entries of the sampling matrix Φ in (37) are independent, identically distributed (i.i.d.) standard Gaussian random variables and the transform matrix Ψ in (37) is identity: $\Psi = I_p$.

We vary the values of γ^2, ϵ^2, σ^2, $P_{\rm H}$, and $P_{\rm root}$ to test the performances of various methods under different conditions. Our performance metric is the *average* mean-square error (MSE) of an estimate $\widetilde{\boldsymbol{s}}$ of the signal coefficient vector:

$$\text{MSE}\{\widetilde{\boldsymbol{s}}\} = \frac{\mathrm{E}_{\Phi,\boldsymbol{s},\boldsymbol{y}}[\|\widetilde{\boldsymbol{s}} - \boldsymbol{s}\|_2^2]}{p} \tag{38}$$

computed using 500 Monte Carlo trials, where *averaging* is performed over the random Gaussian sampling matrices Φ, signal $\boldsymbol{s}$, and measurements $\boldsymbol{y}$. The expected number of large-magnitude signal coefficients is

$$\mathrm{E}\Big[\sum_{i=1}^{p} q_i\Big] = \frac{p}{4^L}\Big(1 + 3\sum_{l=0}^{L-1} 4^l\,P_l\Big) \tag{39a}$$

where P_l is the marginal probability that a state variable in the lth tree level is equal to one, computed recursively as follows:

$$P_l = P_{l-1}P_{\rm H} + (1 - P_{l-1})P_{\rm L} \tag{39b}$$

initialized by $P_0 = P_{\rm root}$.

NIHT, DORE, and MB-IHT require knowledge of the signal sparsity level r; in this example, we set r for these methods to the true signal support size. For $\sigma^2 = 1$, we select the convergence threshold in (36) to $\delta = 10^{-4}$ and for $\sigma^2 = 10^{-6}$, we select this convergence threshold to $\delta = 10^{-10}$. For GPSR and FPC$_{\text{AS}}$, we vary a within the set $\{-1, -2, -3, -4, -5, -6, -7, -8, -9\}$ and, for each N/p and each of the two methods, we use the optimal a that achieves the smallest MSE. For MP-EM, we set the grid length $K = 16$.

Recall that the turbo-AMP approach needs normalized columns of the sensing matrix, see [5, eq. (22)]. When applying the turbo-AMP method, we scale the sensing matrix as $H_{\text{scale}} = (1/\sqrt{N})\,\Phi\,\Psi$ so that it has approximately normalized columns. With measurements $\boldsymbol{y}$ and scaled sensing matrix H_{scale}, turbo-AMP returns the scaled signal estimate $\boldsymbol{s}_{\text{scale}}$, and we compute the final turbo-AMP signal estimate as $(\rho_{\Phi}/\sqrt{N})\,\boldsymbol{s}_{\text{scale}}$, whose performance is evaluated using (38).

Figs. 5 and 6 show the MSEs of different methods for several choices of γ^2, ϵ^2, and σ^2 where we fix $P_{\text{H}} = P_{\text{root}} = 0.5$ (corresponding to $\operatorname{E}[\sum_{i=1}^{p} q_i]/p = 0.0918$) and consider $\sigma^2 \in \{1, 10^{-6}\}$, $\epsilon^2 \in \{0.1, 10\}$, and $\gamma^2 \in \{10^3, 10^5\}$. Here, a larger value of the low-signal relative variance ϵ^2 implies that the signal coefficient vector $\boldsymbol{s}$ is less (approximately) sparse and a larger value of the high-signal relative variance γ^2 implies a relatively higher signal-to-noise (SNR). Observe that the noise variance $\sigma^2 = 10^{-6}$ corresponds to the noise precision $1/\sigma^2 = 10^6$, which is the mean of the prior pdf for $1/\sigma^2$ used in [5, Sec. IV, p. 3444].

In Fig. 5, we show the MSEs of various methods as functions of the subsampling factor N/p for more sparse signals ($\epsilon^2 = 0.1$), relatively lower SNR ($\gamma^2 = 10^3$), and variable noise variance $\sigma^2 \in \{1, 10^{-6}\}$. Observe that turbo-AMP is sensitive to the choice of the noise variance σ^2: It has the largest MSE for $\sigma^2 = 1$ and $N/p < 0.4$, but becomes the second best method for $\sigma^2 = 10^{-6}$ and most N/p. In contrast, MP-EM keeps the best reconstruction performance as σ^2 varies: The MSE of MP-EM is up to 4.6 times smaller than its closest competitor for both $\sigma^2 = 1$ and $\sigma^2 = 10^{-6}$.

The MSEs of most methods are roughly 10^6 times smaller in Fig. 5(b) where $\sigma^2 = 10^{-6}$ than the corresponding MSEs in Fig. 5(a) where $\sigma^2 = 1$. However, this is not true for turbo-AMP, which is very sensitive to the selection of its prior pdf for the noise precision $1/\sigma^2$. For the noise variance $\sigma^2 = 10^{-6}$, turbo-AMP performs significantly better than for $\sigma^2 = 1$ (upon taking into account the scaling adjustment by the factor 10^{-6}), which is facilitated by the fact that $1/\sigma^2 = 10^6$ is the mean of the prior pdf for $1/\sigma^2$ used in [5, Sec. IV, p. 3444] and in the corresponding MATLAB implementation at http://www.ece.osu.edu/~schniter/turboAMPimaging that we employ.

The approximate invariance of MP-EM to scaling of the measurements can be explained by the fact that the shape of the concentrated marginal posterior distribution (10) (which is a function of state variables $\boldsymbol{q}$ only) does not change as we scale the measurements $\boldsymbol{y}$ by a constant.

In Fig. 6, we fix $\sigma^2 = 10^{-6}$, focus on *less (approximately) sparse signals* with $\epsilon^2 = 10$, and show the MSEs of various methods as functions of the subsampling factor N/p for $\gamma^2 = 10^5$ (relatively higher SNR) and $\gamma^2 = 10^3$ (lower SNRs). When $\gamma^2 = 10^5$, turbo-AMP and MP-EM clearly outperform all other methods: turbo-AMP has the smallest MSE for $N/p < 0.3$. The MSE of turbo-AMP is larger than that of MP-EM when $N/p \geq 0.3$. When $\gamma^2 = 10^3$, MP-EM outperforms all the other methods except MP-EM$_{\text{OPT}}$

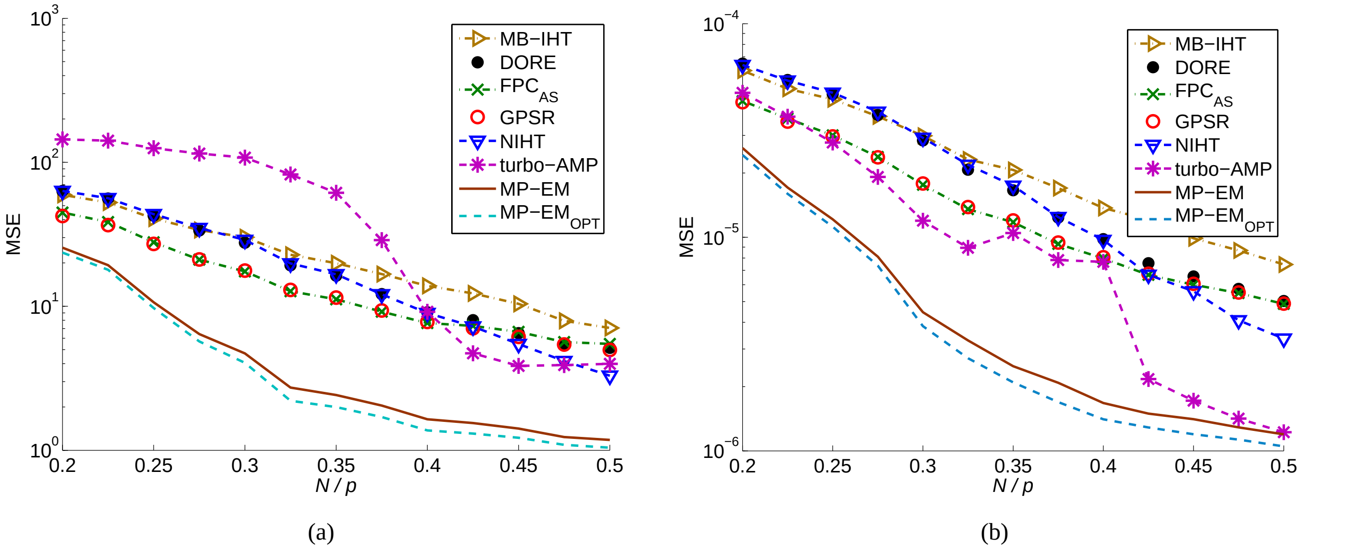


Fig. 5. MSEs as functions of the subsampling factor N/p for $P_{\rm H} = P_{\rm root} = 0.5$, $\gamma^2 = 10^3$, $\epsilon^2 = 0.1$ and (a) $\sigma^2 = 1$ and (b) $\sigma^2 = 10^{-6}$.

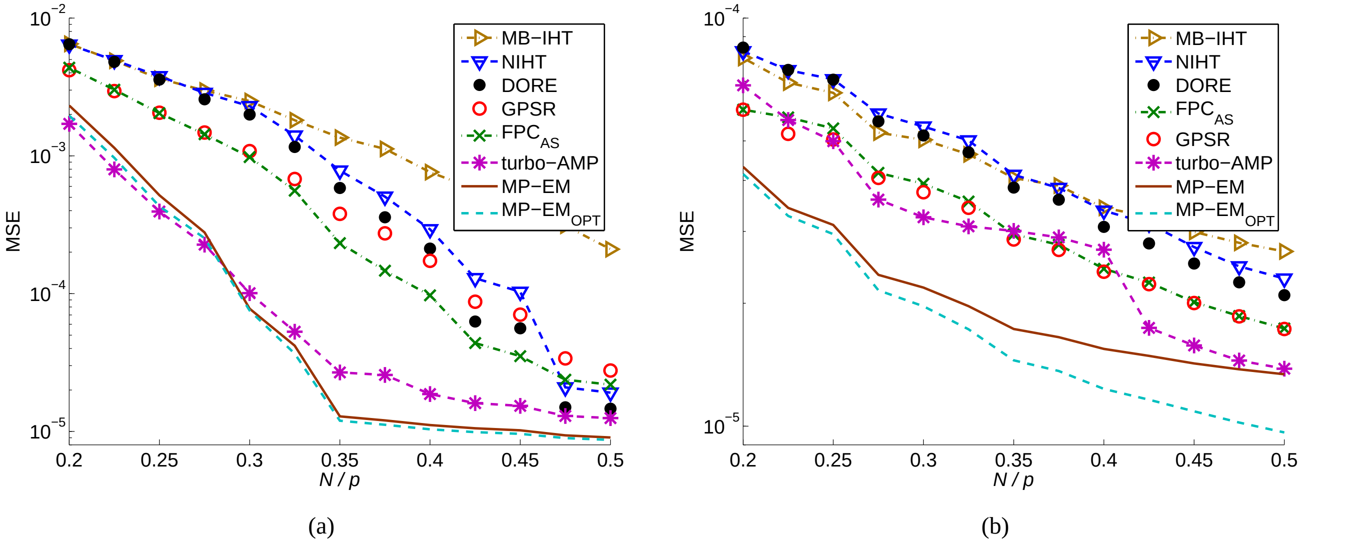


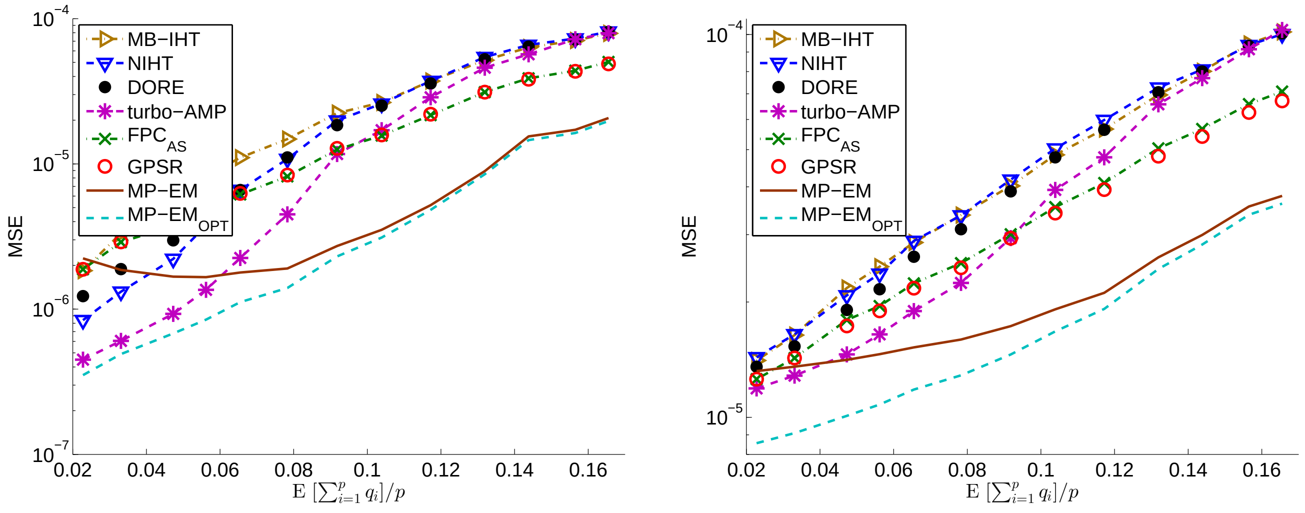
Fig. 6. MSEs as functions of the subsampling factor N/p for $P_{\rm H} = P_{\rm root} = 0.5$, $\sigma^2 = 10^{-6}$, $\epsilon^2 = 10$ and (a) $\gamma^2 = 10^5$ and (b) $\gamma^2 = 10^3$.

for all the subsampling factors.

Parts (b) of Figs. 5 and 6 show the MSE performances of various methods for reconstructing signals that are *more* and *less (approximately) sparse*, respectively, with all other simulation parameters being the same. For each method, the more sparse signals can be reconstructed with a smaller MSE than the less sparse signals at each subsampling factor N/p: Compare Figs. 5(b) and 6(b).

In both Figs. 5 and 6, the MSE of MP-EM is close to that of MP-EM$_{\rm OPT}$, which implies that the marginal-posterior based criterion in Section IV selects the variance parameter well in this example.

Both MP-EM and turbo-AMP yield generally non-sparse signal estimates, particularly when the underlying signal $\boldsymbol{s}$ is less (approximately) sparse, i.e., $\epsilon^2 = 10$.



(a)

(b)

Fig. 7. MSEs as functions of the expected significant coefficient ratio $\text{E}\,[\sum_{i=1}^{p} q_i]/p$ for $\sigma^2 = 10^{-6}$, $\gamma^2 = 10^3$, $N/p = 0.35$ and (a) $\epsilon^2 = 0.1$ and (b) $\epsilon^2 = 10$.

Fig. 7 shows the MSEs of different methods as functions of the normalized expected number of large-magnitude signal coefficients $\text{E}\,[\sum_{i=1}^{p} q_i]/p$ (corresponding to the *expected significant coefficient ratio*), obtained by varying $P_{\text{H}} = P_{\text{root}}$, where we fix $\sigma^2 = 10^{-6}$, $\gamma^2 = 10^3$, $N/p = 0.35$ and consider $\epsilon^2 \in \{0.1, 10\}$. MP-EM$_{\text{OPT}}$ has the smallest MSE for all expected significant coefficient ratios in Fig. 7. MP-EM provides a relatively poor performance compared with other methods when $\text{E}\,[\sum_{i=1}^{p} q_i]$ is small, implying that the marginal-posterior based criterion in Section IV does not select the variance parameter σ^2 well for very small expected significant coefficient ratios and that manual tuning of σ^2 is needed in this case.

For more (approximately) sparse signals with $\epsilon^2 = 0.1$ in Fig. 7(a), MP-EM outperforms all other methods except MP-EM$_{\text{OPT}}$ when $\text{E}\,[\sum_{i=1}^{p} q_i]/p \geq 0.0655$. For less sparse signals with $\epsilon^2 = 10$ in Fig. 7(b), MP-EM becomes the closest competitor to MP-EM$_{\text{OPT}}$ for $\text{E}\,[\sum_{i=1}^{p} q_i]/p \geq 0.0473$. For both more and less sparse signals, the gap between the MSEs of MP-EM and MP-EM$_{\text{OPT}}$ becomes smaller as $\text{E}\,[\sum_{i=1}^{p} q_i]$ increases. Turbo-AMP is the second best method when $\text{E}\,[\sum_{i=1}^{p} q_i]/p < 0.0655$ and $\text{E}\,[\sum_{i=1}^{p} q_i]/p < 0.0473$ for $\epsilon^2 = 0.1$ and $\epsilon^2 = 10$, respectively. However, it achieves a relatively fair performance for larger $\text{E}\,[\sum_{i=1}^{p} q_i]$.

For more (approximately) sparse signals with $\epsilon^2 = 0.1$ in Fig. 7(a), the convex approaches (GPSR and FPC$_{\text{AS}}$) outperform the hard thresholding methods (DORE, MB-IHT, NIHT) when $\text{E}\,[\sum_{i=1}^{p} q_i]/p \geq 0.0655$. For less sparse signals with $\epsilon^2 = 10$ in Fig. 7(b), the convex approaches outperform the hard thresholding methods over the entire range of expected significant coefficient ratios. With the exception of MP-EM and MP-EM$_{\text{OPT}}$, GPSR and FPC$_{\text{AS}}$ have smaller MSEs than all the other methods in Fig. 7(a) when $\mathrm{E}\,[\sum_{i=1}^{p} q_i]/p \geq 0.104$.

MB-IHT, which employs a greedy tree approximation and deterministic tree structure, achieves quite a poor MSE performance in Figs. 5–7. A relatively poor performance of MB-CoSaMP (which employs the same deterministic tree structure) has also been reported in [5, Sec. IV.B].

B. Image Reconstruction Using Gaussian I.I.D. Sampling Matrices

We reconstruct the 128×128 'Cameraman' image from compressive samples generated using random sampling matrices Φ with i.i.d. standard normal elements and the $p \times p$ orthogonal inverse Haar wavelet transform matrix Ψ. Here, the matrix-to-vector conversion operator $\upsilon(\cdot)$ is based on the MATLAB wavelet decomposition function `wavedec2` with Haar wavelet, which has also been used in [3] and [5]. Our performance metric is the average MSE of a signal coefficient vector estimate $\widetilde{\boldsymbol{s}}$:

$$\mathrm{MSE}\{\widetilde{\boldsymbol{s}}\} = \frac{\mathrm{E}_{\,\Phi}[\|\widetilde{\boldsymbol{s}} - \boldsymbol{s}\|_2^2]}{p} \tag{40}$$

computed using 10 Monte Carlo trials, where averaging is performed over the random Gaussian sampling matrices Φ.

Here, we employ DORE$_{\text{app}}$ that approximates the $(H\,H^T)^{-1} = \rho_\Phi^2\,(\Phi\,\Phi^T)^{-1}$ term by $(\rho_\Phi^2/p)\,I_N$, which is justified by the fact that $\mathrm{E}_{\,\Phi}[\Phi\Phi^T] = p\,I_N$ holds in this example, see also (37). For DORE$_{\text{app}}$, we apply the following empirical Bayesian estimate of random signal vector $\boldsymbol{z}$ [11, eq. (16)]:

$$\boldsymbol{z}^{(+\infty)} = \boldsymbol{s}^{(+\infty)} + H^T(HH^T)^{-1}(\boldsymbol{y} - H\boldsymbol{s}^{(+\infty)}) \tag{41}$$

where $\boldsymbol{s}^{(+\infty)}$ denotes the sparse signal estimates obtained upon convergence of DORE$_{\text{app}}$ iteration and the $(H\,H^T)^{-1}$ term is approximated by $(\rho_\Phi^2/p)\,I_N$. We set the sparsity level r for NIHT and DORE$_{\text{app}}$ as $2000\,N/p$ and $2500\,N/p$ for MB-IHT, tuned for good MSE performance. The convergence threshold in (36) is set to $\delta = 10^{-5}$. The grid length in MP-EM is set as $K = 12$ and the tuning parameters for MP-EM are chosen as

$$\gamma^2 = 1000, \quad \epsilon^2 = 0.1, \quad P_{\text{root}} = P_{\text{H}} = 0.2, \quad P_{\text{L}} = 10^{-5}. \tag{42}$$

For GPSR and FPC$_{\text{AS}}$, we tuned the regularization parameter τ manually by varying a with the set $\{-1, -2, -3, -4, -5, -6, -7, -8, -9\}$: the best reconstruction performances are achieved for $a = -3$. When applying the turbo-AMP method, we scale the sensing matrix as $H_{\text{scale}} = (1/\sqrt{N})\,\Phi\,\Psi$ and apply the same scaling correction as in the example in Section V-A.

Fig. 8 shows the MSE performances of different algorithms as functions of the normalized number of measurements (subsampling factor) N/p. MP-EM achieves the best MSE when $N/p \leq 0.35$. The MSEs

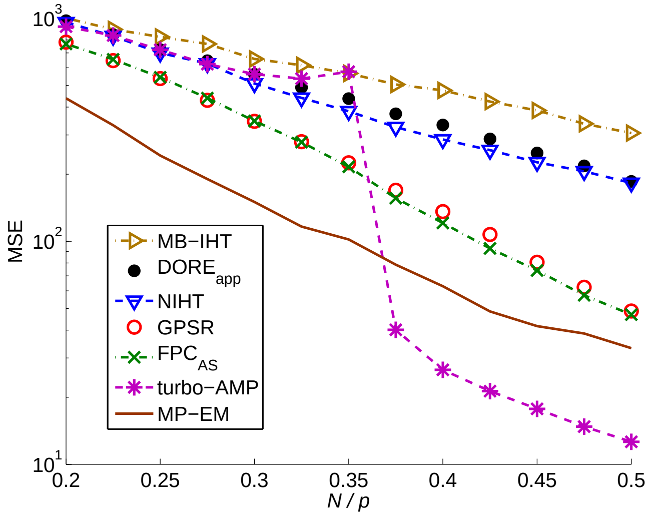


Fig. 8. MSEs as functions of the subsampling factor N/p.

of GPSR and FPC$_{\text{AS}}$ are close to each other and smaller than those of DORE$_{\text{app}}$, NIHT, and MB-IHT for all N/p and the MSE of MP-EM is 1.4 to 2.4 times smaller than that of GPSR and FPC$_{\text{AS}}$, see Fig. 8.

MB-IHT has the largest MSE for most N/p, which is likely due to the fact that it employs the deterministic tree structure, as discussed earlier.

For $N/p \leq 0.35$, turbo-AMP performs similarly to DORE$_{\text{app}}$, NIHT, and MB-IHT, but it outperforms all other methods for $N/p > 0.35$. The reasons why turbo-AMP performs well for large N/p, outperforming all competitors, are likely the followings:

- it uses a more general prior on the binary state variables than our MP-EM method, which allows the tree probability parameters P_{H}, P_{L}, γ^2, and ϵ^2 to vary between the signal decomposition levels, and
- *learns* the tree probability parameters parameters from the measurements.

In contrast, our MP-EM method employs the crude choices of the tree and other tuning parameters in (42).

C. Large-scale Image Reconstruction Using a Structurally Random Sampling Matrix

We now reconstruct the standard 256×256 'Lena' and 'Cameraman' images. As in Section V-B, the matrix-to-vector conversion operator $\upsilon(\cdot)$ is based on the MATLAB wavelet decomposition function `wavedec2` with Haar wavelet. The sampling matrix Φ is generated from structurally random compressive samples [22] and the transform matrix $\varPsi$ in (37) is the $p \times p$ orthogonal inverse Haar wavelet transform matrix, which implies that the sensing matrix H has orthonormal rows: $H\,H^T = I_N$ and, consequently, $\rho_\Phi = \rho_H = 1$. Our performance metric is the peak signal-to-noise ratio (PSNR) of an estimated signal $\widetilde{\boldsymbol{s}}$:

$$\text{PSNR (dB)} = 10\,\log_{10}\left\{\frac{[(\varPsi\boldsymbol{s})_{\text{MAX}} - (\varPsi\boldsymbol{s})_{\text{MIN}}]^2}{\|\widetilde{\boldsymbol{s}} - \boldsymbol{s}\|_2^2/p}\right\}. \tag{43}$$

Here, we employ the exact DORE and the exact random signal estimate in (41), which are computationally tractable because H has orthonormal rows. We set the sparsity level r for NIHT and DORE as $10000\,N/p$ and $15000\,N/p$ for MB-IHT, tuned for good PSNR performance. The convergence threshold in (36) is set to $\delta = 0.1$. The tuning parameters for MP-EM are given in (42) and the grid length in MP-EM is set as $K = 12$, the same as in Section V-B. We tuned the regularization parameters τ in (34b) for FPC_{AS} and GPSR manually and found that the best performance is achieved when $a = -3$ for both algorithms.

When applying the turbo-AMP method, we scale the sensing matrix as $H_{\text{scale}} = (\sqrt{p/N})\,\Phi\,\Psi$. With measurements $\boldsymbol{y}$ and scaled sensing matrix H_{scale}, turbo-AMP returns the scaled signal estimate $\boldsymbol{s}_{\text{scale}}$, and we compute the final turbo-AMP signal estimate as $(\sqrt{p/N})\,\boldsymbol{s}_{\text{scale}}$, whose performance is evaluated using (43). Our empirical experience shows that scaling the sensing matrix improves the reconstruction performance of the turbo-AMP algorithm in this example.

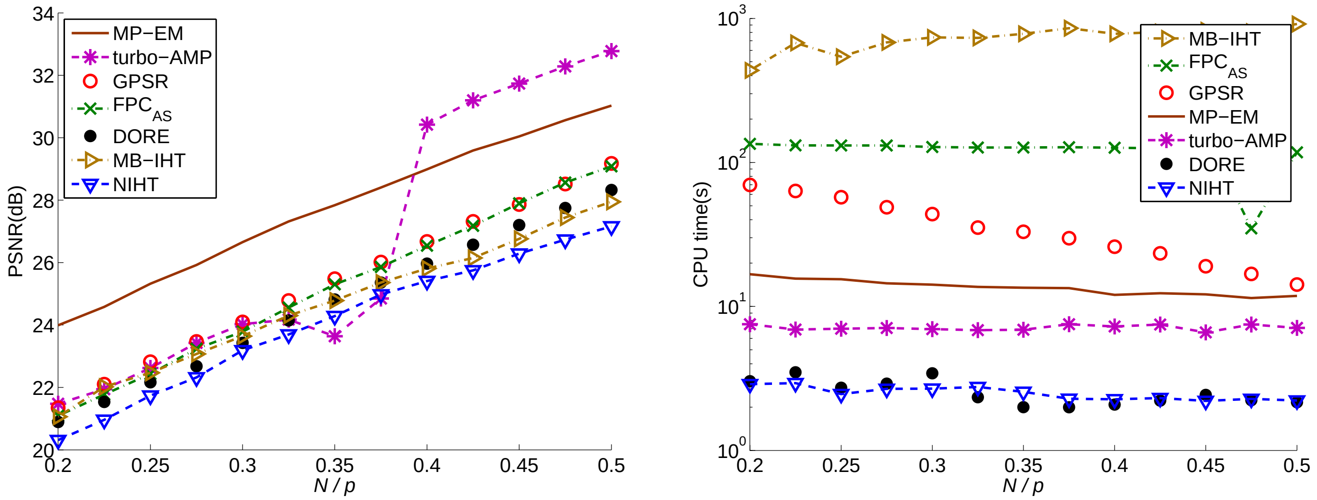
Fig. 9 shows the PSNRs and CPU times achieved by various methods when reconstructing the 256×256 'Lena' image. For $N/p < 0.4$, the proposed MP-EM method outperforms all other methods, where the performance improvement compared with the closest competitor varies between 2.4 dB and 2.6 dB. For $N/p \geq 0.4$, turbo-AMP outperforms all other methods. In terms of CPU time, DORE and NIHT are the fastest among all the methods compared. It takes around 7 seconds as the runtime for turbo-AMP at each measurement point. MP-EM is 1.5 to 2.3 times slower than turbo-AMP, but obviously faster than GPSR, FPC_{AS}, and MB-IHT.[1]

Fig. 10 shows the PSNRs and CPU times achieved by various methods when reconstructing the 256×256 'Cameraman' image. For $N/p < 0.4$, the proposed MP-EM method outperforms all other methods by at least 2.6 dB. For $N/p \geq 0.4$, turbo-AMP outperforms all other methods, but performs quite poorly for $N/p < 0.35$: a similar pattern that occurs also in Fig. 9. According to Fig. 10(b), both DORE and NIHT consume less than 4 s in terms of CPU time. It takes around 7 s for turbo-AMP at every measurement point. MP-EM is still consistently faster than GPSR, FPC_{AS}, and MB-IHT, and requires 4.0 to 10.8 s more than turbo-AMP.

In Figs. 9 and 10, MB-IHT achieves a fair performance and consumes the longest CPU time.

Figs. 11 and 12 show the reconstructed 256×256 'Lena' and 'Cameraman' images by different methods for $N/p = 0.375$, respectively: The MP-EM algorithm achieves better reconstructed image quality compared with the other methods.

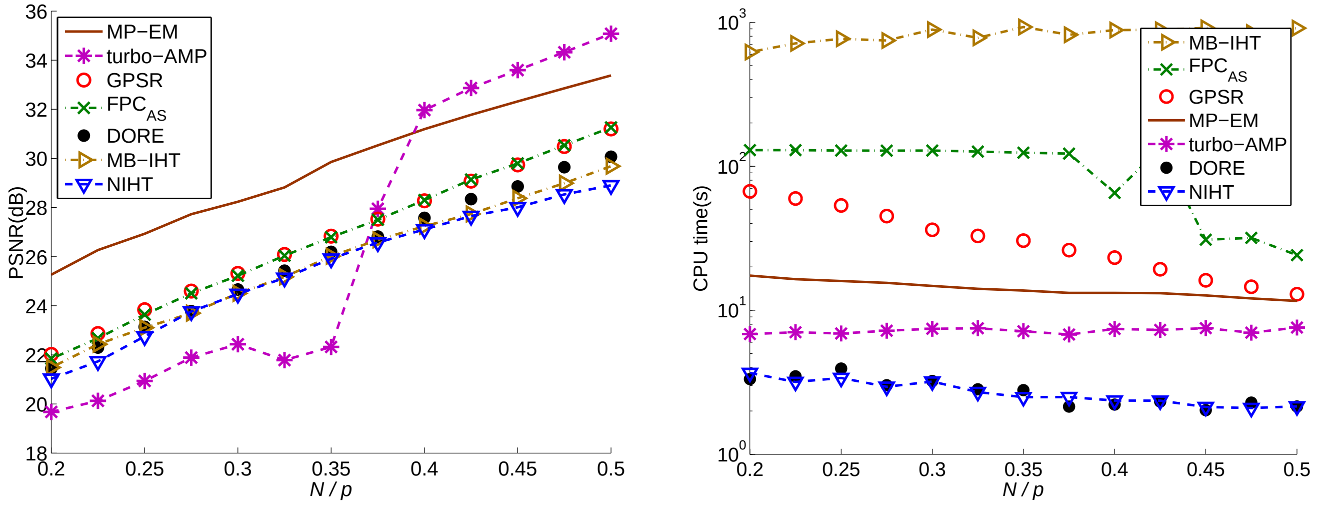
[1] Regarding the reported CPU time, note that the turbo-AMP code does not use MATLAB only, but combines MATLAB and JAVA codes.



(a)

(b)

Fig. 9. (a) PSNRs and (b) CPU times as functions of the subsampling factor N/p for the 256×256 'Lena' image.



(a)

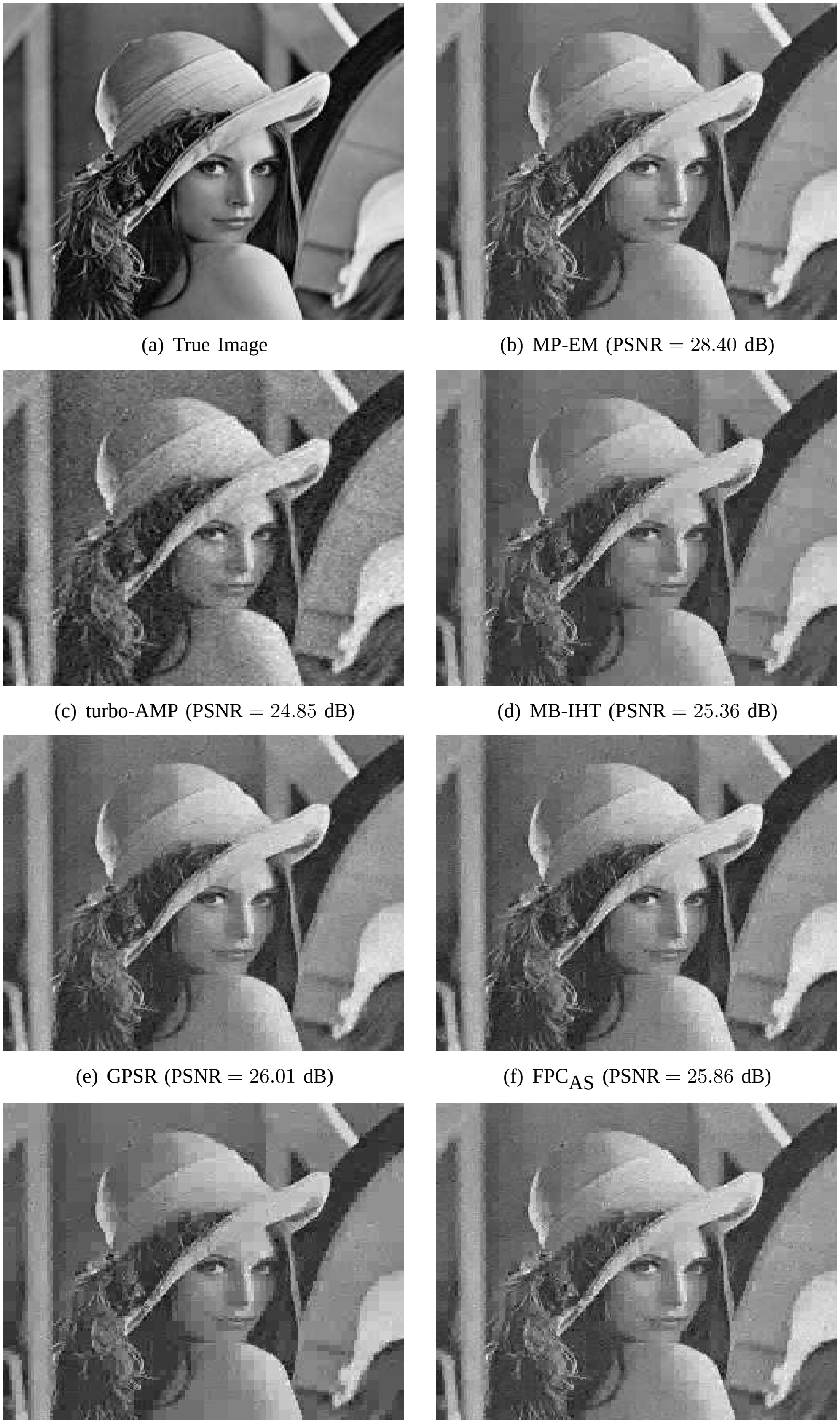
(b)

Fig. 10. (a) PSNRs and (b) CPU times as functions of the subsampling factor N/p for the 256×256 'Cameraman' image.

VI. Concluding Remarks

We presented a Bayesian EM algorithm for reconstructing approximately sparse signal from compressive samples using a Markov tree prior for the signal coefficients. We employed the max-product belief propagation algorithm to implement the M step of the proposed EM iteration. Compared with the existing message passing algorithms in the compressive sampling area, our method does not approximate the message form. The simulation results show that our algorithm often outperforms existing algorithms for simulated signals and standard test images with different sampling operators.

Our future work will include the convergence analysis of the MP-EM algorithm, incorporating other measurement models, using a more general prior on the binary state variables, and designing schemes for



(a) True Image

(b) MP-EM (PSNR $= 28.40$ dB)

(c) turbo-AMP (PSNR $= 24.85$ dB)

(d) MB-IHT (PSNR $= 25.36$ dB)

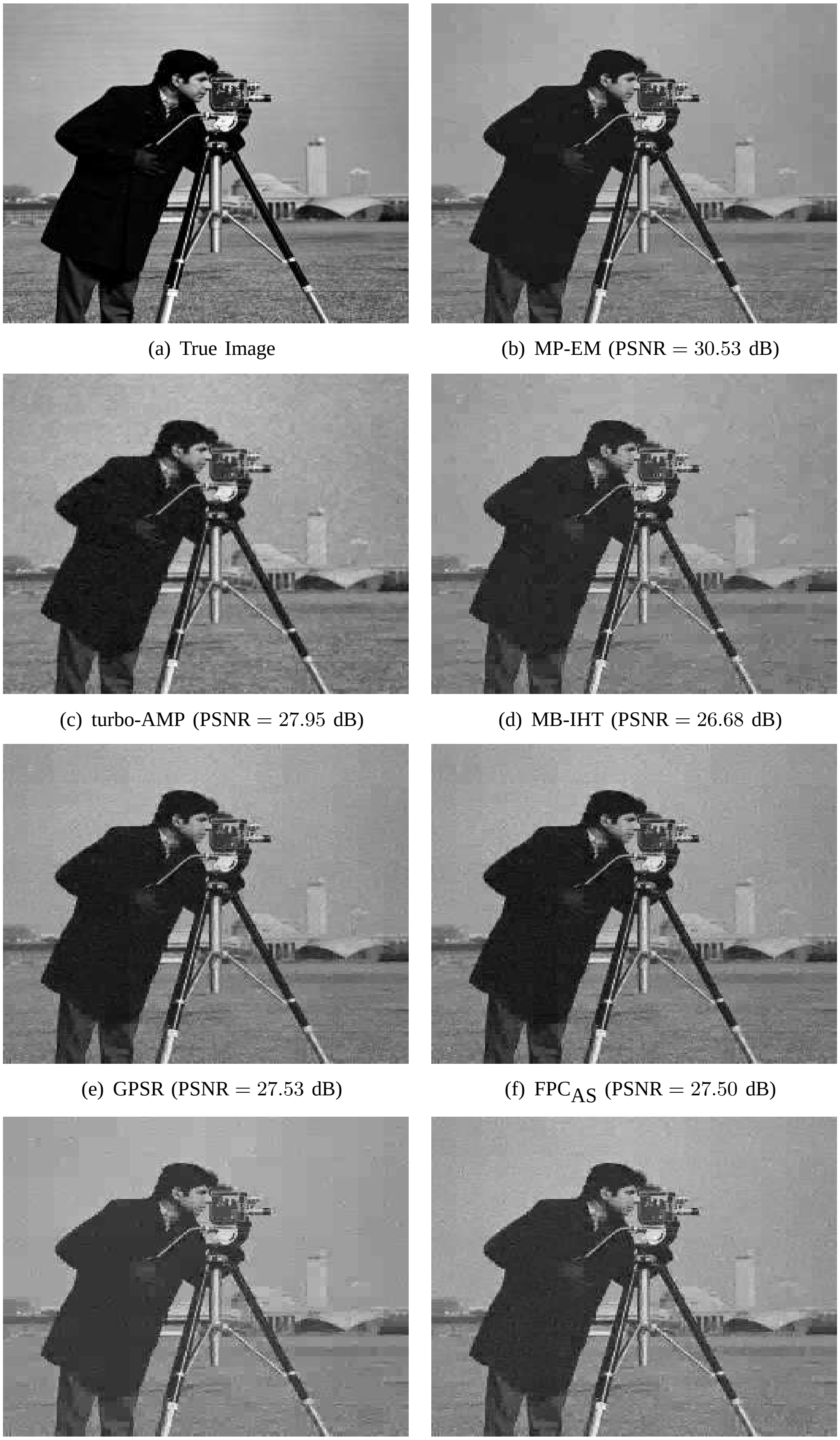
(e) GPSR (PSNR $= 26.01$ dB)

(f) FPC_{AS} (PSNR $= 25.86$ dB)

(g) NIHT (PSNR $= 24.98$ dB)

(h) DORE (PSNR $= 25.36$ dB)

Fig. 11. The ‘Lena’ image reconstructed by various methods for $N/p = 0.375$.



(a) True Image

(b) MP-EM (PSNR $= 30.53$ dB)

(c) turbo-AMP (PSNR $= 27.95$ dB)

(d) MB-IHT (PSNR $= 26.68$ dB)

(e) GPSR (PSNR $= 27.53$ dB)

(f) FPC_{AS} (PSNR $= 27.50$ dB)

(g) NIHT (PSNR $= 26.57$ dB)

(h) DORE (PSNR $= 26.82$ dB)

Fig. 12. The ‘Cameraman’ image reconstructed by various methods for $N/p = 0.375$.

learning the tree parameters from the measurements.

APPENDIX

APPENDIX A
DERIVATION OF THE EM ALGORITHM AND PROOFS OF ITS MONOTONICITY AND (15)

Consider the hierarchical two-stage model in (11). The complete-data posterior distribution for known σ^2 is

$$\begin{aligned} p_{\boldsymbol{\theta},\boldsymbol{z}|\sigma^2,\boldsymbol{y}}(\boldsymbol{\theta},\boldsymbol{z}|\sigma^2,\boldsymbol{y}) &\propto p_{\boldsymbol{y}|\boldsymbol{z},\sigma^2}(\boldsymbol{y}|\boldsymbol{z},\sigma^2)\, p_{\boldsymbol{z}|\boldsymbol{s}}(\boldsymbol{z}|\boldsymbol{s})\, p_{\boldsymbol{s}|\boldsymbol{q},\sigma^2}(\boldsymbol{s}|\boldsymbol{q},\sigma^2)\, p_{\boldsymbol{q}}(\boldsymbol{q})\, (\sigma^2)^{-1} \\ &\propto \frac{\exp\{-\frac{1}{2}(\boldsymbol{y}-H\boldsymbol{z})^T[C(\sigma^2)]^{-1}(\boldsymbol{y}-H\boldsymbol{z})\}}{\sqrt{\det[C(\sigma^2)]}}\,(\epsilon^2/\gamma^2)^{0.5\sum_{i=1}^{p} q_i}\, p_{\boldsymbol{q}}(\boldsymbol{q}) \\ &\quad \cdot \exp[-0.5\,\|\boldsymbol{z}-\boldsymbol{s}\|_2^2/\sigma^2 - 0.5\,\boldsymbol{s}^T D^{-1}(\boldsymbol{q})\,\boldsymbol{s}/\sigma^2] \end{aligned} \tag{A1a}$$

where

$$C(\sigma^2) = \sigma^2(I_N - HH^T) \tag{A1b}$$

and

$$p_{\boldsymbol{z}|\sigma^2,\boldsymbol{y},\boldsymbol{\theta}}(\boldsymbol{z}|\sigma^2,\boldsymbol{y},\boldsymbol{\theta}) = p_{\boldsymbol{z}|\sigma^2,\boldsymbol{y},\boldsymbol{s}}(\boldsymbol{z}|\sigma^2,\boldsymbol{y},\boldsymbol{s}) = \mathcal{N}(\boldsymbol{z}|\mathrm{E}_{\,\boldsymbol{z}|\sigma^2,\boldsymbol{y},\boldsymbol{s}}(\boldsymbol{z}|\sigma^2,\boldsymbol{y},\boldsymbol{s}), \mathrm{cov}_{\boldsymbol{z}|\sigma^2,\boldsymbol{y},\boldsymbol{s}}(\boldsymbol{z}|\sigma^2,\boldsymbol{y},\boldsymbol{s})) \tag{A1c}$$

where

$$\mathrm{E}_{\,\boldsymbol{z}|\sigma^2,\boldsymbol{y},\boldsymbol{s}}(\boldsymbol{z}|\sigma^2,\boldsymbol{y},\boldsymbol{s}) = \{H^T[C(\sigma^2)]^{-1}H + I_p/\sigma^2\}^{-1}\{H^T[C(\sigma^2)]^{-1}\boldsymbol{y} + \boldsymbol{s}/\sigma^2\} \tag{A1d}$$

$$\mathrm{cov}_{\boldsymbol{z}|\sigma^2,\boldsymbol{y},\boldsymbol{s}}(\boldsymbol{z}|\sigma^2,\boldsymbol{y},\boldsymbol{s}) = \{H^T[C(\sigma^2)]^{-1}H + I_p/\sigma^2\}^{-1} \tag{A1e}$$

By using the matrix inversion lemma [23, eq. (2.22), p. 424]:

$$(R+STU)^{-1} = R^{-1} - R^{-1}S(T^{-1}+UR^{-1}S)^{-1}UR^{-1} \tag{A2a}$$

and the following identity [23, p. 425]:

$$(R+STU)^{-1}ST = R^{-1}S(T^{-1}+UR^{-1}S)^{-1} \tag{A2b}$$

we obtain

$$\mathrm{E}_{\,\boldsymbol{z}|\sigma^2,\boldsymbol{y},\boldsymbol{s}}(\boldsymbol{z}|\sigma^2,\boldsymbol{y},\boldsymbol{s}) = \boldsymbol{s} + H^T(\boldsymbol{y} - H\boldsymbol{s}) \tag{A3}$$

which leads to (12).

The objective function $\ln p_{\boldsymbol{\theta}\,|\,\sigma^2,\boldsymbol{y}}(\boldsymbol{\theta}\,|\,\sigma^2,\boldsymbol{y})$ that we aim to maximize in Section III satisfies the following property in the EM iteration:

$$\ln p_{\boldsymbol{\theta}\,|\,\sigma^2,\boldsymbol{y}}(\boldsymbol{\theta}\,|\,\sigma^2,\boldsymbol{y}) = \mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(j)}) - \mathcal{H}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(j)}) \tag{A4a}$$

where

$$\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(j)}) \triangleq \mathrm{E}_{\,\boldsymbol{z}|\sigma^2,\boldsymbol{y},\boldsymbol{\theta}}\left[\ln p_{\boldsymbol{\theta},\boldsymbol{z}|\sigma^2,\boldsymbol{y}}(\boldsymbol{\theta},\boldsymbol{z}|\sigma^2,\boldsymbol{y})|\sigma^2,\boldsymbol{y},\boldsymbol{\theta}^{(j)}\right] \tag{A4b}$$

$$\mathcal{H}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(j)}) \triangleq \mathrm{E}_{\,\boldsymbol{z}|\sigma^2,\boldsymbol{y},\boldsymbol{\theta}}\left[\ln p_{\boldsymbol{z}|\sigma^2,\boldsymbol{y},\boldsymbol{\theta}}(\boldsymbol{z}|\sigma^2,\boldsymbol{y},\boldsymbol{\theta})|\sigma^2,\boldsymbol{y},\boldsymbol{\theta}^{(j)}\right] \tag{A4c}$$

From (A1a) and (A3), $\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(j)})$ could be computed as

$$\begin{aligned}\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(j)}) &= \text{const} + \mathrm{E}_{\boldsymbol{z}|\sigma^2,\boldsymbol{y},\boldsymbol{\theta}}\Big\{ -0.5(\boldsymbol{y}-H\boldsymbol{z})^T[C(\sigma^2)]^{-1}(\boldsymbol{y}-H\boldsymbol{z}) - 0.5\,\|\boldsymbol{z}-\boldsymbol{s}\|_2^2/\sigma^2 \\ &\qquad -0.5\,\boldsymbol{s}^T D^{-1}(\boldsymbol{q})\,\boldsymbol{s}/\sigma^2 + \ln[p_{\boldsymbol{q}}(\boldsymbol{q})] + 0.5\,\ln(\epsilon^2/\gamma^2)\sum_{i=1}^{p} q_i \,\Big|\, \sigma^2, \boldsymbol{y}, \boldsymbol{\theta}^{(j)}\Big\} \\ &= \text{const} - 0.5\,\frac{\|\boldsymbol{z}^{(j)}-\boldsymbol{s}\|_2^2 + \boldsymbol{s}^T D^{-1}(\boldsymbol{q})\,\boldsymbol{s}}{\sigma^2} + \ln[p_{\boldsymbol{q}}(\boldsymbol{q})] + 0.5\,\ln\Big(\frac{\epsilon^2}{\gamma^2}\Big)\sum_{i=1}^{p} q_i \qquad \text{(A5)}\end{aligned}$$

where const denotes the terms that are not functions of $\boldsymbol{\theta}$ and (13a) follows. Since $\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(j)})$ is maximized at $\boldsymbol{\theta}^{(j+1)}$, we have

$$\mathcal{Q}(\boldsymbol{\theta}^{(j+1)}|\boldsymbol{\theta}^{(j)}) \geq \mathcal{Q}(\boldsymbol{\theta}^{(j)}|\boldsymbol{\theta}^{(j)}) \qquad \text{(A6)}$$

and (14) follows from (A4a) by using the inequalities (A6) and

$$\mathcal{H}(\boldsymbol{\theta}^{(j+1)}|\boldsymbol{\theta}^{(j)}) \leq \mathcal{H}(\boldsymbol{\theta}^{(j)}|\boldsymbol{\theta}^{(j)}) \qquad \text{(A7)}$$

where (A7) is a consequence of the fact that $\mathcal{H}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(j)})$ is maximized with respect to $\boldsymbol{\theta}$ at $\boldsymbol{\theta} = \boldsymbol{\theta}^{(j)}$.

Proof of (15)*:* For a given $\boldsymbol{q}$, (A5) is a quadratic function of $\boldsymbol{s}$ that is easy to maximize with respect to $\boldsymbol{s}$:

$$\arg\max_{\boldsymbol{s}} \mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(j)}) = \big[D^{-1}(\boldsymbol{q}) + I_p\big]^{-1}\boldsymbol{z}^{(j)}. \qquad \text{(A8)}$$

Therefore, the estimates of $\boldsymbol{s}$ and $\boldsymbol{q}$ obtained upon convergence of the EM iteration in Section III to its fixed point satisfy:

$$\begin{aligned}\boldsymbol{s}^{(+\infty)} &= \big[D^{-1}(\boldsymbol{q}^{(+\infty)}) + I_p\big]^{-1}\boldsymbol{z}^{(+\infty)} \\ &= \big[D^{-1}(\boldsymbol{q}^{(+\infty)}) + I_p\big]^{-1}\big[\boldsymbol{s}^{(+\infty)} + H^T(\boldsymbol{y} - H\,\boldsymbol{s}^{(+\infty)})\big] \qquad \text{(A9)}\end{aligned}$$

where the second equality follows by using (12). Solving (A9) for $\boldsymbol{s}^{(+\infty)}$ yields

$$\boldsymbol{s}^{(+\infty)} = \big[D^{-1}(\boldsymbol{q}^{(+\infty)}) + H^TH\big]^{-1}H^T\boldsymbol{y} \qquad \text{(A10)}$$

and (15) follows. □

APPENDIX B
DERIVATION OF THE MESSAGES AND BELIEFS IN SECTION III-A

Before we proceed, note the following useful identities:

$$\arg\max_{s_i} \mathcal{N}(z_i\,;\,s_i,\sigma^2)\,\mathcal{N}(s_i\,;\,0,\tau^2) = \frac{\tau^2\,z_i}{\sigma^2+\tau^2} \qquad \text{(B1a)}$$

$$\max_{s_i} \mathcal{N}(z_i\,;\,s_i,\sigma^2)\,\mathcal{N}(s_i\,;\,0,\tau^2) = \frac{1}{\sqrt{2\,\pi\,\sigma^2}\,\sqrt{2\,\pi\,\tau^2}}\,\exp\Big(-0.5\,\frac{z_i^2}{\sigma^2+\tau^2}\Big). \qquad \text{(B1b)}$$

I Upward Messages

1) Upward Messages from Leaf Nodes: When passing upward messages from the leaf nodes $i \in \mathcal{T}_{\text{leaf}}$, we set the multiplicative term $\prod_{k\in\text{ch}(i)} m_{k\to i}(q_i)$ to one, yielding [see (22)]

$$\begin{aligned} m_{i\to\pi(i)}(q_{\pi(i)}) &= \alpha \max_{\boldsymbol{\theta}_i}\{\psi_i(\boldsymbol{\theta}_i)\,\psi_{i,\pi(i)}(q_i,q_{\pi(i)})\} \\ &= \alpha \max_{\boldsymbol{\theta}_i}\big\{\mathcal{N}(z_i\,;\,s_i,\sigma^2)\,[\mathcal{N}(s_i\,;\,0,\gamma^2\sigma^2)]^{q_i}\,[\mathcal{N}(s_i\,;\,0,\epsilon^2\sigma^2)]^{1-q_i} \\ &\quad \cdot[P_{\text{H}}^{q_i}\,(1-P_{\text{H}})^{1-q_i}]^{q_{\pi(i)}}\,[P_{\text{L}}^{q_i}\,(1-P_{\text{L}})^{1-q_i}]^{1-q_{\pi(i)}}\big\}. \end{aligned} \tag{B2}$$

For $q_{\pi(i)} = 0$, we have

$$\begin{aligned} m_{i\to\pi(i)}(0) = \mu_i^{\text{u}}(0) &= \alpha \max_{\boldsymbol{\theta}_i}\big\{\mathcal{N}(z_i\,;\,s_i,\sigma^2)\,[\mathcal{N}(s_i\,;\,0,\gamma^2\sigma^2)]^{q_i}\,[\mathcal{N}(s_i\,;\,0,\epsilon^2\sigma^2)]^{1-q_i}\,P_{\text{L}}^{q_i}\,(1-P_{\text{L}})^{1-q_i}\big\} \\ &= \alpha_1 \max\Big\{(1-P_{\text{L}})\,\exp\big(-0.5\,\frac{z_i^2}{\sigma^2+\sigma^2\epsilon^2}\big)/\epsilon,\; P_{\text{L}}\,\exp\big(-0.5\,\frac{z_i^2}{\sigma^2+\sigma^2\gamma^2}\big)/\gamma\Big\} \end{aligned} \tag{B3a}$$

and, for $q_{\pi(i)} = 1$, we have

$$\begin{aligned} m_{i\to\pi(i)}(1) = \mu_i^{\text{u}}(1) &= \alpha \max_{\boldsymbol{\theta}_i}\big\{\mathcal{N}(z_i\,;\,s_i,\sigma^2)\,[\mathcal{N}(s_i\,;\,0,\gamma^2\sigma^2)]^{q_i}\,[\mathcal{N}(s_i\,;\,0,\epsilon^2\sigma^2)]^{1-q_i}\,P_{\text{H}}^{q_i}\,(1-P_{\text{H}})^{1-q_i}\big\} \\ &= \alpha_1 \max\Big\{(1-P_{\text{H}})\,\exp\Big(-0.5\,\frac{z_i^2}{\sigma^2+\sigma^2\epsilon^2}\Big)/\epsilon,\; P_{\text{H}}\,\exp\Big(-0.5\,\frac{z_i^2}{\sigma^2+\sigma^2\gamma^2}\Big)/\gamma\Big\} \end{aligned} \tag{B3b}$$

where we have used (B1b) with $\tau^2 = \sigma^2\epsilon^2$ and $\tau^2 = \sigma^2\gamma^2$ and $\alpha > 0$ and $\alpha_1 > 0$ are appropriate normalizing constants. It follows from (B1a) that the only two candidates for $\boldsymbol{\theta}_i$ in the maximization of (B2) are $[0, \widehat{s}_i(0)]^T$ and $[1, \widehat{s}_i(1)]^T$.

In summary,

$$m_{i\to\pi(i)}(q_{\pi(i)}) = [\mu_i^{\text{u}}(0)]^{1-q_{\pi(i)}}\,[\mu_i^{\text{u}}(1)]^{q_{\pi(i)}} \tag{B4a}$$

and (B3a) and (B3b) can be rewritten as

$$\mu_i^{\text{u}}(0) = \max\{\boldsymbol{\nu}_{0,i}^{\text{u}}\}/(\max\{\boldsymbol{\nu}_{0,i}^{\text{u}}\} + \max\{\boldsymbol{\nu}_{1,i}^{\text{u}}\}) \tag{B4b}$$

$$\mu_i^{\text{u}}(1) = \max\{\boldsymbol{\nu}_{1,i}^{\text{u}}\}/(\max\{\boldsymbol{\nu}_{0,i}^{\text{u}}\} + \max\{\boldsymbol{\nu}_{1,i}^{\text{u}}\}) \tag{B4c}$$

and $\boldsymbol{\nu}_{0,i}^{\text{u}}, \boldsymbol{\nu}_{1,i}^{\text{u}}$, and $\boldsymbol{\phi}(z)$ were defined in (23c), (23d), and (23f).

2) Upward Messages from Non-Leaf Nodes: For $i \in \mathcal{T}\backslash\mathcal{T}_{\text{leaf}}$, we can use induction to simplify the multiplicative term $\prod_{k\in\text{ch}(i)} m_{k\to i}(q_i)$ in (22) as follows:

$$\prod_{k\in\text{ch}(i)} m_{k\to i}(q_i) = [\prod_{k\in\text{ch}(i)} \mu_k^{\text{u}}(0)]^{1-q_i}\,[\prod_{k\in\text{ch}(i)} \mu_k^{\text{u}}(1)]^{q_i} \tag{B5}$$

see also Fig. 3(a).

Substituting (B5) into (22) yields

$$\begin{aligned} m_{i\to\pi(i)}(q_{\pi(i)}) &= \alpha \max_{\boldsymbol{\theta}_i}\big\{\psi_i(\boldsymbol{\theta}_i)\,\psi_{i,\pi(i)}(q_i,q_{\pi(i)})\prod_{k\in\text{ch}(i)} m_{k\to i}(q_i)\big\} \\ &= \alpha \max_{\boldsymbol{\theta}_i}\Big\{\mathcal{N}(z_i\,;\,s_i,\sigma^2)\,[\mathcal{N}(s_i\,;\,0,\gamma^2\sigma^2)]^{q_i}\,[\mathcal{N}(s_i\,;\,0,\epsilon^2\sigma^2)]^{1-q_i}\,[P_{\text{H}}^{q_i}\,(1-P_{\text{H}})^{1-q_i}]^{q_{\pi(i)}} \\ &\quad \cdot[P_{\text{L}}^{q_i}\,(1-P_{\text{L}})^{1-q_i}]^{1-q_{\pi(i)}}[\prod_{k\in\text{ch}(i)} \mu_k^{\text{u}}(0)]^{1-q_i}\,[\prod_{k\in\text{ch}(i)} \mu_k^{\text{u}}(1)]^{q_i}\Big\}. \end{aligned} \tag{B6}$$

For $q_{\pi(i)} = 0$, we have

$$\begin{aligned} m_{i\to\pi(i)}(0) &= \alpha \max_{\boldsymbol{\theta}_i} \big\{\mathcal{N}(z_i\,;\,s_i,\sigma^2)\,[\mathcal{N}(s_i\,;\,0,\gamma^2\sigma^2)]^{q_i}\,[\mathcal{N}(s_i\,;\,0,\epsilon^2\sigma^2)]^{1-q_i}\,P_{\mathrm{L}}^{q_i}\,(1-P_{\mathrm{L}})^{1-q_i} \\ &\quad \cdot [\prod_{k\in\mathrm{ch}(i)} \mu_k^{\mathrm{u}}(0)]^{1-q_i}\,[\prod_{k\in\mathrm{ch}(i)} \mu_k^{\mathrm{u}}(1)]^{q_i}\big\} \\ &= \alpha_1 \max\Big\{(1-P_{\mathrm{L}})\,[\prod_{k\in\mathrm{ch}(i)} \mu_k^{\mathrm{u}}(0)]\,\exp\big(-0.5\,\frac{z_i^2}{\sigma^2+\sigma^2\epsilon^2}\big)/\epsilon, \\ &\quad P_{\mathrm{L}}\,[\prod_{k\in\mathrm{ch}(i)} \mu_k^{\mathrm{u}}(1)]\,\exp\big(-0.5\,\frac{z_i^2}{\sigma^2+\sigma^2\gamma^2}\big)/\gamma\Big\} \end{aligned} \tag{B7a}$$

and, for $q_{\pi(i)} = 1$, we have

$$\begin{aligned} m_{i\to\pi(i)}(1) &= \alpha \max_{\boldsymbol{\theta}_i} \big\{\mathcal{N}(z_i\,;\,s_i,\sigma^2)\,[\mathcal{N}(s_i\,;\,0,\gamma^2\sigma^2)]^{q_i}\,[\mathcal{N}(s_i\,;\,0,\epsilon^2\sigma^2)]^{1-q_i}\,P_{\mathrm{H}}^{q_i}\,(1-P_{\mathrm{H}})^{1-q_i} \\ &\quad \cdot [\prod_{k\in\mathrm{ch}(i)} \mu_k^{\mathrm{u}}(0)]^{1-q_i}\,[\prod_{k\in\mathrm{ch}(i)} \mu_k^{\mathrm{u}}(1)]^{q_i}\big\} \\ &= \alpha_1 \max\Big\{(1-P_{\mathrm{H}})\,[\prod_{k\in\mathrm{ch}(i)} \mu_k^{\mathrm{u}}(0)]\,\exp\big(-0.5\,\frac{z_i^2}{\sigma^2+\sigma^2\epsilon^2}\big)/\epsilon, \\ &\quad P_{\mathrm{H}}\,[\prod_{k\in\mathrm{ch}(i)} \mu_k^{\mathrm{u}}(1)]\,\exp\big(-0.5\,\frac{z_i^2}{\sigma^2+\sigma^2\gamma^2}\big)/\gamma\Big\} \end{aligned} \tag{B7b}$$

where we have used (B1b) with $\tau^2 = \sigma^2\epsilon^2$ and $\tau^2 = \sigma^2\gamma^2$ and $\alpha > 0$ and $\alpha_1 > 0$ are appropriate normalizing constants.

In summary,

$$m_{i\to\pi(i)}(q_{\pi(i)}) = [\mu_i^{\mathrm{u}}(0)]^{1-q_{\pi(i)}}\,[\mu_i^{\mathrm{u}}(1)]^{q_{\pi(i)}} \tag{B8a}$$

where

$$\mu_i^{\mathrm{u}}(0) = \max\{\boldsymbol{\nu}_{0,i}^{\mathrm{u}} \odot \boldsymbol{\eta}_i^{\mathrm{u}}\}/(\max\{\boldsymbol{\nu}_{0,i}^{\mathrm{u}} \odot \boldsymbol{\eta}_i^{\mathrm{u}}\} + \max\{\boldsymbol{\nu}_{1,i}^{\mathrm{u}} \odot \boldsymbol{\eta}_i^{\mathrm{u}}\}) \tag{B8b}$$

$$\mu_i^{\mathrm{u}}(1) = \max\{\boldsymbol{\nu}_{1,i}^{\mathrm{u}} \odot \boldsymbol{\eta}_i^{\mathrm{u}}\}/(\max\{\boldsymbol{\nu}_{0,i}^{\mathrm{u}} \odot \boldsymbol{\eta}_i^{\mathrm{u}}\} + \max\{\boldsymbol{\nu}_{1,i}^{\mathrm{u}} \odot \boldsymbol{\eta}_i^{\mathrm{u}}\}) \tag{B8c}$$

and

$$\boldsymbol{\eta}_i^{\mathrm{u}} = \bigodot_{k\in\mathrm{ch}(i)} \boldsymbol{\mu}_k^{\mathrm{u}}. \tag{B8d}$$

The general upward message form in (23) follows by combining (B4) and (B8).

II Downward Messages

Based on the results in Section III-A1 and Appendix B-I, we simplify the product of upward messages sent from the siblings of node i in (24) as follows [see (23a)]:

$$\prod_{k\in\mathrm{sib}(i)} m_{k\to\pi(i)}(q_{\pi(i)}) = [\prod_{k\in\mathrm{sib}(i)} \mu_k^{\mathrm{u}}(0)]^{1-q_{\pi(i)}}\,[\prod_{k\in\mathrm{sib}(i)} \mu_k^{\mathrm{u}}(1)]^{q_{\pi(i)}} \tag{B9}$$

see also Fig. 3(b).

1) Downward Messages from Root Nodes: For the node $\pi(i) \in \mathcal{T}_{\text{root}}$, we set the message $m_{\text{gp}(i)\to\pi(i)}(q_{\pi(i)})$ to one, yielding [see (24)]

$$m_{\pi(i)\to i}(q_i) = \alpha \max_{\boldsymbol{\theta}_{\pi(i)}} \Big\{ \psi_{\pi(i)}(\boldsymbol{\theta}_{\pi(i)})\, \psi_{i,\pi(i)}(q_i, q_{\pi(i)}) \prod_{k\in\text{sib}(i)} m_{k\to\pi(i)}(q_{\pi(i)}) \Big\}. \tag{B10}$$

Substituting (B9) into (B10) yields

$$\begin{aligned} m_{\pi(i)\to i}(q_i) &= \alpha \max_{\boldsymbol{\theta}_{\pi(i)}} \Big\{ \psi_{\pi(i)}(\boldsymbol{\theta}_{\pi(i)})\, \psi_{i,\pi(i)}(q_i, q_{\pi(i)}) \prod_{k\in\text{sib}(i)} m_{k\to\pi(i)}(q_{\pi(i)}) \Big\} \\ &= \alpha \max_{\boldsymbol{\theta}_{\pi(i)}} \Big\{ \mathcal{N}(z_{\pi(i)}\,;\, s_{\pi(i)}, \sigma^2)\, [P_{\text{root}} \mathcal{N}(s_{\pi(i)}\,;\, 0, \gamma^2\sigma^2)]^{q_{\pi(i)}}\, [(1-P_{\text{root}})\mathcal{N}(s_{\pi(i)}\,;\, 0, \epsilon^2\sigma^2)]^{1-q_{\pi(i)}} \\ &\cdot [P_{\text{H}}^{q_i}\,(1-P_{\text{H}})^{1-q_i}]^{q_{\pi(i)}}\, [P_{\text{L}}^{q_i}\,(1-P_{\text{L}})^{1-q_i}]^{1-q_{\pi(i)}}\, [\prod_{k\in\text{sib}(i)} \mu_k^{\text{u}}(0)]^{1-q_{\pi(i)}}\, [\prod_{k\in\text{sib}(i)} \mu_k^{\text{u}}(1)]^{q_{\pi(i)}} \Big\}. \end{aligned} \tag{B11}$$

For $q_i = 0$, we have

$$\begin{aligned} m_{\pi(i)\to i}(0) &= \alpha \max_{\boldsymbol{\theta}_{\pi(i)}} \Big\{ \mathcal{N}(z_{\pi(i)}\,;\, s_{\pi(i)}, \sigma^2)\, [\mathcal{N}(s_{\pi(i)}\,;\, 0, \gamma^2\sigma^2)]^{q_{\pi(i)}}\, [\mathcal{N}(s_{\pi(i)}\,;\, 0, \epsilon^2\sigma^2)]^{1-q_{\pi(i)}} \\ &\quad \cdot \{(1-P_{\text{root}})(1-P_{\text{L}})[\prod_{k\in\text{sib}(i)} \mu_k^{\text{u}}(0)]\}^{1-q_{\pi(i)}}\, \{P_{\text{root}}(1-P_{\text{H}})[\prod_{k\in\text{sib}(i)} \mu_k^{\text{u}}(1)]\}^{q_{\pi(i)}} \Big\} \\ &= \alpha_1 \max \Big\{ (1-P_{\text{root}})(1-P_{\text{L}})[\prod_{k\in\text{sib}(i)} \mu_k^{\text{u}}(0)]\, \exp\Big(-0.5\, \frac{z_{\pi(i)}^2}{\sigma^2 + \sigma^2\epsilon^2}\Big)/\epsilon, \\ &\quad P_{\text{root}}(1-P_{\text{H}})[\prod_{k\in\text{sib}(i)} \mu_k^{\text{u}}(1)]\exp\Big(-0.5\, \frac{z_{\pi(i)}^2}{\sigma^2 + \sigma^2\gamma^2}\Big)/\gamma \Big\} \end{aligned} \tag{B12a}$$

and for $q_i = 1$, we have

$$\begin{aligned} m_{\pi(i)\to i}(1) &= \alpha \max_{\boldsymbol{\theta}_{\pi(i)}} \Big\{ \mathcal{N}(z_{\pi(i)}\,;\, s_{\pi(i)}, \sigma^2)\, [\mathcal{N}(s_{\pi(i)}\,;\, 0, \gamma^2\sigma^2)]^{q_{\pi(i)}}\, [\mathcal{N}(s_{\pi(i)}\,;\, 0, \epsilon^2\sigma^2)]^{1-q_{\pi(i)}} \\ &\quad \cdot \{(1-P_{\text{root}})\, P_{\text{L}}\, [\prod_{k\in\text{sib}(i)} \mu_k^{\text{u}}(0)]\}^{1-q_{\pi(i)}}\, \{P_{\text{root}}\, P_{\text{H}}\, [\prod_{k\in\text{sib}(i)} \mu_k^{\text{u}}(1)]\}^{q_{\pi(i)}} \Big\} \\ &= \alpha_1 \max \Big\{ (1-P_{\text{root}})\, P_{\text{L}}\, [\prod_{k\in\text{sib}(i)} \mu_k^{\text{u}}(0)]\, \exp\Big(-0.5\, \frac{z_{\pi(i)}^2}{\sigma^2 + \sigma^2\epsilon^2}\Big)/\epsilon, \\ &\quad P_{\text{root}}\, P_{\text{H}}\, [\prod_{k\in\text{sib}(i)} \mu_k^{\text{u}}(1)]\exp\Big(-0.5\, \frac{z_{\pi(i)}^2}{\sigma^2 + \sigma^2\gamma^2}\Big)/\gamma \Big\} \end{aligned} \tag{B12b}$$

where we have used (B1b) with $\tau^2 = \sigma^2\epsilon^2$ and $\tau^2 = \sigma^2\gamma^2$ and $\alpha > 0$ and $\alpha_1 > 0$ are appropriate normalizing constants. The only two candidates to maximize (B10) are $[0, \widehat{s}_{\pi(i)}(0)]^T$ and $[1, \widehat{s}_{\pi(i)}(1)]^T$.

In summary,

$$m_{\pi(i)\to i}(q_i) = [\mu_i^{\text{d}}(0)]^{1-q_i}\, [\mu_i^{\text{d}}(1)]^{q_i} \tag{B13a}$$

where

$$\mu_i^{\text{d}}(0) = \max\{\boldsymbol{\nu}_{0,i}^{\text{d}} \odot \boldsymbol{\eta}_i^{\text{d}}\}/(\max\{\boldsymbol{\nu}_{0,i}^{\text{d}} \odot \boldsymbol{\eta}_i^{\text{d}}\} + \max\{\boldsymbol{\nu}_{1,i}^{\text{d}} \odot \boldsymbol{\eta}_i^{\text{d}}\}) \tag{B13b}$$

$$\mu_i^{\text{d}}(1) = \max\{\boldsymbol{\nu}_{1,i}^{\text{d}} \odot \boldsymbol{\eta}_i^{\text{d}}\}/(\max\{\boldsymbol{\nu}_{0,i}^{\text{d}} \odot \boldsymbol{\eta}_i^{\text{d}}\} + \max\{\boldsymbol{\nu}_{1,i}^{\text{d}} \odot \boldsymbol{\eta}_i^{\text{d}}\}) \tag{B13c}$$

and

$$\boldsymbol{\nu}^{\rm d}_{0,i} = \big[\ 1-P_{\rm L},\ \ 1-P_{\rm H}\ \big]^T \odot \boldsymbol{\phi}(z_{\pi(i)}) \odot \big[\bigodot_{k\in{\rm sib}(i)} \boldsymbol{\mu}^{\rm u}_k \big] \tag{B13d}$$

$$\boldsymbol{\nu}^{\rm d}_{1,i} = \big[\ P_{\rm L},\ \ P_{\rm H}\ \big]^T \odot \boldsymbol{\phi}(z_{\pi(i)}) \odot \big[\bigodot_{k\in{\rm sib}(i)} \boldsymbol{\mu}^{\rm u}_k \big] \tag{B13e}$$

$$\boldsymbol{\eta}^{\rm d}_i = \big[1-P_{\rm root},\ P_{\rm root}\big]^T. \tag{B13f}$$

2) Downward Messages from Non-Root Nodes: For the node $\pi(i) \in (\mathcal{T}\backslash\mathcal{T}_{\rm root})\backslash\mathcal{T}_{\rm leaf}$, using the same strategy as above, (24) simplifies as

$$\begin{aligned} m_{\pi(i)\to i}(q_i) &= \alpha \max_{\boldsymbol{\theta}_{\pi(i)}} \Big\{ \psi_{\pi(i)}(\boldsymbol{\theta}_{\pi(i)})\, \psi_{i,\pi(i)}(q_i, q_{\pi(i)})\, m_{{\rm gp}(i)\to\pi(i)}(q_{\pi(i)}) \prod_{k\in{\rm sib}(i)} m_{k\to\pi(i)}(q_{\pi(i)}) \Big\} \\ &= \alpha \max_{\boldsymbol{\theta}_{\pi(i)}} \Big\{ \mathcal{N}(z_{\pi(i)}\,;\, s_{\pi(i)}, \sigma^2)\, [\mathcal{N}(s_{\pi(i)}\,;\, 0, \gamma^2\sigma^2)]^{q_{\pi(i)}}\, [\mathcal{N}(s_{\pi(i)}\,;\, 0, \epsilon^2\sigma^2)]^{1-q_{\pi(i)}} \\ &\quad \cdot [P_{\rm H}^{q_i}\,(1-P_{\rm H})^{1-q_i}]^{q_{\pi(i)}}\, [P_{\rm L}^{q_i}\,(1-P_{\rm L})^{1-q_i}]^{1-q_{\pi(i)}}\, [\prod_{k\in{\rm sib}(i)} \mu^{\rm u}_k(0)]^{1-q_{\pi(i)}}\, [\prod_{k\in{\rm sib}(i)} \mu^{\rm u}_k(1)]^{q_{\pi(i)}} \\ &\quad \cdot [\mu^{\rm d}_{\pi(i)}(0)]^{1-q_{\pi(i)}}\, [\mu^{\rm d}_{\pi(i)}(1)]^{q_{\pi(i)}} \Big\} \end{aligned} \tag{B14}$$

For $q_i = 0$, we have

$$\begin{aligned} m_{\pi(i)\to i}(0) &= \alpha \max_{\boldsymbol{\theta}_{\pi(i)}} \Big\{ \mathcal{N}(z_{\pi(i)}\,;\, s_{\pi(i)}, \sigma^2)\, [\mathcal{N}(s_{\pi(i)}\,;\, 0, \gamma^2\sigma^2)]^{q_{\pi(i)}}\, [\mathcal{N}(s_{\pi(i)}\,;\, 0, \epsilon^2\sigma^2)]^{1-q_{\pi(i)}} \\ &\quad \cdot \{\mu^{\rm d}_{\pi(i)}(0)\,(1-P_{\rm L})[\prod_{k\in{\rm sib}(i)} \mu^{\rm u}_k(0)]\}^{1-q_{\pi(i)}}\, \{\mu^{\rm d}_{\pi(i)}(1)\,(1-P_{\rm H})[\prod_{k\in{\rm sib}(i)} \mu^{\rm u}_k(1)]\}^{q_{\pi(i)}} \Big\} \\ &= \alpha_1 \max \Big\{ \mu^{\rm d}_{\pi(i)}(0)\,(1-P_{\rm L})[\prod_{k\in{\rm sib}(i)} \mu^{\rm u}_k(0)]\, \exp\Big(-0.5\,\frac{z^2_{\pi(i)}}{\sigma^2+\sigma^2\epsilon^2}\Big)/\epsilon, \\ &\qquad \mu^{\rm d}_{\pi(i)}(1)\,(1-P_{\rm H})[\prod_{k\in{\rm sib}(i)} \mu^{\rm u}_k(1)]\exp\Big(-0.5\,\frac{z^2_{\pi(i)}}{\sigma^2+\sigma^2\gamma^2}\Big)/\gamma \Big\} \end{aligned} \tag{B15a}$$

and for $q_i = 1$, we have

$$\begin{aligned} m_{\pi(i)\to i}(1) &= \alpha \max_{\boldsymbol{\theta}_{\pi(i)}} \Big\{ \mathcal{N}(z_{\pi(i)}\,;\, s_{\pi(i)}, \sigma^2)\, [\mathcal{N}(s_{\pi(i)}\,;\, 0, \gamma^2\sigma^2)]^{q_{\pi(i)}}\, [\mathcal{N}(s_{\pi(i)}\,;\, 0, \epsilon^2\sigma^2)]^{1-q_{\pi(i)}} \\ &\quad \cdot \{\mu^{\rm d}_{\pi(i)}(0)\,P_{\rm L}\,[\prod_{k\in{\rm sib}(i)} \mu^{\rm u}_k(0)]\}^{1-q_{\pi(i)}}\, \{\mu^{\rm d}_{\pi(i)}(1)\,P_{\rm H}\,[\prod_{k\in{\rm sib}(i)} \mu^{\rm u}_k(1)]\}^{q_{\pi(i)}} \Big\} \\ &= \alpha_1 \max \Big\{ \mu^{\rm d}_{\pi(i)}(0)\,P_{\rm L}\,[\prod_{k\in{\rm sib}(i)} \mu^{\rm u}_k(0)]\, \exp\Big(-0.5\,\frac{z^2_{\pi(i)}}{\sigma^2+\sigma^2\epsilon^2}\Big)/\epsilon, \\ &\qquad \mu^{\rm d}_{\pi(i)}(1)\,P_{\rm H}\,[\prod_{k\in{\rm sib}(i)} \mu^{\rm u}_k(1)]\exp\Big(-0.5\,\frac{z^2_{\pi(i)}}{\sigma^2+\sigma^2\gamma^2}\Big)/\gamma \Big\} \end{aligned} \tag{B15b}$$

where we have used (B1b) with $\tau^2 = \sigma^2\epsilon^2$ and $\tau^2 = \sigma^2\gamma^2$ and $\alpha > 0$ and $\alpha_1 > 0$ are appropriate normalizing constants. The only two candidates to maximize (B14) are $[0, \widehat{s}_{\pi(i)}(0)]^T$ and $[1, \widehat{s}_{\pi(i)}(1)]^T$.

In summary,

$$m_{\pi(i)\to i}(q_i) = [\mu^{\rm d}_i(0)]^{1-q_i}\, [\mu^{\rm d}_i(1)]^{q_i} \tag{B16a}$$

where

$$\mu_i^{\mathrm{d}}(0) = \max\{\boldsymbol{\nu}_{0,i}^{\mathrm{d}} \odot \boldsymbol{\eta}_i^{\mathrm{d}}\}/(\max\{\boldsymbol{\nu}_{0,i}^{\mathrm{d}} \odot \boldsymbol{\eta}_i^{\mathrm{d}}\} + \max\{\boldsymbol{\nu}_{1,i}^{\mathrm{d}} \odot \boldsymbol{\eta}_i^{\mathrm{d}}\}) \tag{B16b}$$

$$\mu_i^{\mathrm{d}}(1) = \max\{\boldsymbol{\nu}_{1,i}^{\mathrm{d}} \odot \boldsymbol{\eta}_i^{\mathrm{d}}\}/(\max\{\boldsymbol{\nu}_{0,i}^{\mathrm{d}} \odot \boldsymbol{\eta}_i^{\mathrm{d}}\} + \max\{\boldsymbol{\nu}_{1,i}^{\mathrm{d}} \odot \boldsymbol{\eta}_i^{\mathrm{d}}\}) \tag{B16c}$$

and

$$\boldsymbol{\eta}_i^{\mathrm{d}} = \boldsymbol{\mu}_{\pi(i)}^{\mathrm{d}} \tag{B16d}$$

The general downward message form in (25) follows by combining (B13) and (B16).

III Beliefs

Define the vector $\boldsymbol{\beta}_i = [\beta_i(0),\, \beta_i(1)]^T$ as

$$\beta_i(0) = \max_{s_i} b([0,\, s_i]^T), \quad \beta_i(1) = \max_{s_i} b([1,\, s_i]^T) \tag{B17}$$

where $b(\boldsymbol{\theta}_i)$ are the beliefs defined in (26).

1) Beliefs for the Root Nodes: For root nodes $i \in \mathcal{T}_{\mathrm{root}}$, the beliefs $b(\boldsymbol{\theta}_i)$ in (26) become

$$\begin{aligned} b(\boldsymbol{\theta}_i) &= \alpha\, \mathcal{N}(z_i\,;\, s_i, \sigma^2)\, [P_{\mathrm{root}}\, \mathcal{N}(s_i\,;\, 0, \gamma^2\sigma^2)]^{q_i}\, [(1-P_{\mathrm{root}})\, \mathcal{N}(s_i\,;\, 0, \epsilon^2\sigma^2)]^{1-q_i} \\ &\quad \cdot \Big[\prod_{k\in \mathrm{ch}(i)} \mu_k^{\mathrm{u}}(0)\Big]^{1-q_i} \Big[\prod_{k\in \mathrm{ch}(i)} \mu_k^{\mathrm{u}}(1)\Big]^{q_i}. \end{aligned} \tag{B18}$$

and (B17) simplify to

$$\beta_i(0) = \alpha\, \frac{1}{\sqrt{2\,\pi\,\sigma^2}\,\sqrt{2\,\pi\,\epsilon^2\sigma^2}}\, \exp\Big(-0.5\, \frac{z_i^2}{\sigma^2 + \sigma^2\epsilon^2}\Big)\, (1 - P_{\mathrm{root}}) \prod_{k\in \mathrm{ch}(i)} \mu_k^{\mathrm{u}}(0) \tag{B19a}$$

$$\beta_i(1) = \alpha\, \frac{1}{\sqrt{2\,\pi\,\sigma^2}\,\sqrt{2\,\pi\,\gamma^2\sigma^2}}\, \exp\Big(-0.5\, \frac{z_i^2}{\sigma^2 + \sigma^2\gamma^2}\Big)\, P_{\mathrm{root}} \prod_{k\in \mathrm{ch}(i)} \mu_k^{\mathrm{u}}(1) \tag{B19b}$$

yielding

$$\boldsymbol{\beta}_i = [\beta_i(0), \beta_i(1)]^T = \alpha_1 [1 - P_{\mathrm{root}}, P_{\mathrm{root}}]^T \odot \boldsymbol{\phi}(z_i) \odot \boldsymbol{\eta}_i^{\mathrm{u}}. \tag{B20}$$

2) Beliefs for the Non-Root Non-Leaf Nodes: For $i \in (\mathcal{T} \setminus \mathcal{T}_{\mathrm{root}}) \setminus \mathcal{T}_{\mathrm{leaf}}$, the beliefs $b(\boldsymbol{\theta}_i)$ in (26) become

$$\begin{aligned} b(\boldsymbol{\theta}_i) &= \alpha\, \mathcal{N}(z_i\,;\, s_i, \sigma^2)\, [\mathcal{N}(s_i\,;\, 0, \gamma^2\sigma^2)]^{q_i}\, [\mathcal{N}(s_i\,;\, 0, \epsilon^2\sigma^2)]^{1-q_i}\, [\mu_i^{\mathrm{d}}(0)]^{1-q_i}\, [\mu_i^{\mathrm{d}}(1)]^{q_i} \\ &\quad \cdot \Big[\prod_{k\in \mathrm{ch}(i)} \mu_k^{\mathrm{u}}(0)\Big]^{1-q_i} \Big[\prod_{k\in \mathrm{ch}(i)} \mu_k^{\mathrm{u}}(1)\Big]^{q_i} \end{aligned} \tag{B21}$$

and (B17) simplify to

$$\beta_i(0) = \alpha\, \frac{1}{\sqrt{2\,\pi\,\sigma^2}\,\sqrt{2\,\pi\,\epsilon^2\sigma^2}}\, \exp\Big(-0.5\, \frac{z_i^2}{\sigma^2 + \sigma^2\epsilon^2}\Big)\, \mu_i^{\mathrm{d}}(0) \prod_{k\in \mathrm{ch}(i)} \mu_k^{\mathrm{u}}(0) \tag{B22a}$$

$$\beta_i(1) = \alpha\, \frac{1}{\sqrt{2\,\pi\,\sigma^2}\,\sqrt{2\,\pi\,\gamma^2\sigma^2}}\, \exp\Big(-0.5\, \frac{z_i^2}{\sigma^2 + \sigma^2\gamma^2}\Big)\, \mu_i^{\mathrm{d}}(1) \prod_{k\in \mathrm{ch}(i)} \mu_k^{\mathrm{u}}(1) \tag{B22b}$$

yielding

$$\boldsymbol{\beta}_i = [\beta_i(0), \beta_i(1)]^T = \alpha_1 \boldsymbol{\phi}(z_i) \odot \boldsymbol{\mu}_i^{\mathrm{d}} \odot \boldsymbol{\eta}_i^{\mathrm{u}}. \tag{B23}$$

3) Beliefs for the Leaf Nodes: For $i \in \mathcal{T}_{\text{leaf}}$, the beliefs $b(\boldsymbol{\theta}_i)$ in (26) become

$$b(\boldsymbol{\theta}_i) = \alpha\, \mathcal{N}(z_i\,;\, s_i, \sigma^2)\, [\mathcal{N}(s_i\,;\, 0, \gamma^2\sigma^2)]^{q_i}\, [\mathcal{N}(s_i\,;\, 0, \epsilon^2\sigma^2)]^{1-q_i}\, [\mu_i^{\text{d}}(0)]^{1-q_i}\, [\mu_i^{\text{d}}(1)]^{q_i} \tag{B24}$$

and (B17) simplify to

$$\beta_i(0) = \alpha\, \frac{1}{\sqrt{2\,\pi\,\sigma^2}\,\sqrt{2\,\pi\,\epsilon^2\sigma^2}}\, \exp\Big(-0.5\, \frac{z_i^2}{\sigma^2 + \sigma^2\epsilon^2}\Big)\, \mu_i^{\text{d}}(0) \tag{B25a}$$

$$\beta_i(1) = \alpha\, \frac{1}{\sqrt{2\,\pi\,\sigma^2}\,\sqrt{2\,\pi\,\gamma^2\sigma^2}}\, \exp\Big(-0.5\, \frac{z_i^2}{\sigma^2 + \sigma^2\gamma^2}\Big)\, \mu_i^{\text{d}}(1) \tag{B25b}$$

yielding

$$\boldsymbol{\beta}_i = [\beta_i(0), \beta_i(1)]^T = \alpha_1 \boldsymbol{\phi}(z_i) \odot \boldsymbol{\mu}_i^{\text{d}}. \tag{B26}$$

In summary,

$$\boldsymbol{\beta}_i = [\beta_i(0), \beta_i(1)]^T = \begin{cases} \alpha_1[1 - P_{\text{root}}, P_{\text{root}}]^T \odot \boldsymbol{\phi}(z_i) \odot \boldsymbol{\eta}_i^{\text{u}}, & i \in \mathcal{T}_{\text{root}} \\ \alpha_1 \boldsymbol{\phi}(z_i) \odot \boldsymbol{\mu}_i^{\text{d}} \odot \boldsymbol{\eta}_i^{\text{u}}, & i \in \mathcal{T} \setminus \mathcal{T}_{\text{root}} \end{cases}.$$

Consequently, the mode $\widehat{\boldsymbol{\theta}}_i$ is computed as

$$\widehat{\boldsymbol{\theta}}_i = (\widehat{q}_i, \widehat{s}_i(\widehat{q}_i)) = \arg\max_{\boldsymbol{\theta}_i} b(\boldsymbol{\theta}_i) = \begin{cases} (1, \widehat{s}_i(1)), & \beta_i(1) \geq \beta_i(0) \\ (0, \widehat{s}_i(0)), & \text{otherwise} \end{cases}. \tag{B27}$$

Note that the normalizing constants α and α_1 in the above upward and downward messages and beliefs have been set so that $m_{i\to\pi(i)}(0) + m_{i\to\pi(i)}(1) = 1$, $m_{\pi(i)\to i}(0) + m_{\pi(i)\to i}(1) = 1$, and $\beta_i(0) + \beta_i(1) = 1$ respectively.

References

[1] M. S. Crouse, R. D. Nowak, and R. G. Baraniuk, "Wavelet-based statistical signal processing using hidden Markov models," *IEEE Trans. Signal Process.*, vol. 46, no. 4, pp. 886–902, Apr. 1998.

[2] V. Cevher, P. Indyk, L. Carin, and R. G. Baraniuk, "Sparse signal recovery and acquisition with graphical models," *IEEE Signal Process. Mag.*, vol. 27, no. 6, pp. 92–103, Nov. 2010.

[3] L. He and L. Carin, "Exploiting structure in wavelet-based Bayesian compressive sensing," *IEEE Trans. Signal Process.*, vol. 57, no. 9, pp. 3488–3497, Sep. 2009.

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

[5] S. Som and P. Schniter, "Compressive imaging using approximate message passing and a Markov-tree prior," *IEEE Trans. Signal Process.*, vol. 60, no. 99, pp. 3439–3448, 2012.

[6] D. Baron, S. Sarvotham, and R. G. Baraniuk, "Bayesian compressive sensing via belief propagation," *IEEE Trans. Signal Process.*, vol. 58, no. 1, pp. 269–280, Jan. 2010.

[7] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proc. Nat. Acad. Sci*, vol. 106, no. 45, pp. 18 914–18 919, 2009.

[8] P. Schniter, "Turbo reconstruction of structured sparse signals," in *Proc. Conf. Inform. Sci. Syst.*, Princeton, NJ, 2010, pp. 1–6.

[9] M. A. T. Figueiredo and R. D. Nowak, "An EM algorithm for wavelet-based image restoration," *IEEE Trans. Image Process.*, vol. 12, pp. 906–916, 2003.

[10] Z. Song and A. Dogandžić, "A Bayesian max-product EM algorithm for reconstructing structured sparse signals," in *Proc. Conf. Inform. Sci. Syst.*, Princeton, NJ, 2012, pp. 1–6.

[11] K. Qiu and A. Dogandžić, "Sparse signal reconstruction via ECME hard thresholding," *IEEE Trans. Signal Process.*, vol. 60, pp. 4551–4569, Sep. 2012.

[12] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*, 2nd ed. Chapman & Hall, 2004.

[13] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[14] G. J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*, 2nd ed. New York: Wiley, 2008.

[15] D. Koller and N. Friedman, *Probabilistic Graphical Models*. Cambridge, MA: MIT Press, 2009.

[16] Y. Weiss and W. Freeman, "On the optimality of solutions of the max-product belief-propagation algorithm in arbitrary graphs," *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 736–744, 2001.

[17] J. Pearl, *Probabilistic Reasoning in Intelligent systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufmann, 1988.

[18] Z. Wen, W. Yin, D. Goldfarb, and Y. Zhang, “A fast algorithm for sparse reconstruction based on shrinkage, subspace optimization, and continuation,” *SIAM J. Sci. Comput.*, vol. 32, no. 4, pp. 1832–1857, 2010.

[19] M. A. T. Figueiredo, R. D. Nowak, and S. J. Wright, “Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems,” *IEEE J. Sel. Topics Signal Process.*, vol. 1, no. 4, pp. 586–597, 2007.

[20] T. Blumensath and M. E. Davies, “Normalized iterative hard thresholding: Guaranteed stability and performance,” *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 2, pp. 298–309, 2010.

[21] R. G. Baraniuk, “Optimal tree approximation with wavelets,” in *Proc. SPIE Wavelet Applicat. Signal Image Processing VII*, Denver, CO, 1999, pp. 196–207.

[22] T. Do, L. Gan, N. Nguyen, and T. Tran, “Fast and efficient compressive sensing using structurally random matrices,” *IEEE Trans. Signal Process.*, vol. 60, no. 1, pp. 139–154, 2012.

[23] D. A. Harville, *Matrix Algebra From a Statistician’s Perspective*. New York: Springer-Verlag, 1997.