**Michigan State University Technical Report**
**MSU-ENGR-004-98**



# Multiscale Hidden Markov Models for Bayesian Image Analysis

**Robert D. Nowak**

ABSTRACT Bayesian multiscale image analysis weds the powerful modeling framework of probabilistic graphs with the intuitively appealing and computationally tractable multiresolution paradigm. In addition to providing a very natural and useful framework for modeling and processing images, Bayesian multiscale analysis is often much less computationally demanding compared to classical Markov random field models. This chapter focuses on a probabilistic graph model called the multiscale hidden Markov model (MHMM), which captures the key inter-scale dependencies present in natural signals and images. A common framework for the MHMM is presented that is capable of analyzing both Gaussian and Poisson processes, and applications to Bayesian image analysis are examined.

## 1 Introduction

The goal of image analysis is to extract some information of interest from image data. The information may simply be the underlying image intensities or the location and/or boundaries of objects, or it may be a high-level description of a scene. The common feature of the vast majority of these challenging problems is that they usually cannot be solved without including prior information or knowledge. Hence, many of the more successful image analysis tools are Bayesian. Arguably, the crucial element of Bayesian techniques is the choice of the prior probability model. The most common tools for Bayesian image analysis are Markov random field (MRF) models, which have been successfully applied in a host of problems including restoration, segmentation, and tomographic reconstruction. For examples, see Cross and Jain (1983), Geman and Geman (1984), and Chellappa and Jain (1993). Since edges and other inhomogeneities are key visual features (as is testified to by the huge amount of research devoted to the problem of edge detection), good image priors should be capable of representing them. Although this is possible within the MRF framework, the resulting inference criteria require, in general, computationally intensive Monte Carlo methods, like the version of the Metropolis algorithm proposed by Geman and Geman (1984).

Multiscale (or multiresolution) techniques have been another popular and successful approach to many image analysis problems. Beginning with the seminal work of Adelson and Burt (1981), multiscale image analysis has found application in a wide range of tasks, from low-level image processing to high-level machine vision, and today, it is the basis for most state-of-the-art image compression schemes. One of the advantages of the multiscale approach is its computational efficiency; multiscale analysis involves efficient schemes for passing intermediate results obtained at one analysis scale to the next scale of analysis. Remarkably, Field (1993) argues that similar processing mechanisms take place in the human visual system.

Recently, attempts have been made to develop multiscale image models, that combine the powerful modeling framework of MRFs with the intuitively appealing and computationally tractable multiscale analysis paradigm. Rather than specifying inter-pixel relationships directly in the spatial domain, multiscale models attempt to represent structural relationships more efficiently through *causal* relationships across scales of analysis. Along this line of thinking, various types of multiscale stochastic image models have recently been proposed by a number of researchers, *e.g.,* Charbonnier et al. (1992), Bouman and Shapiro (1994), Crouse et al. (1998), Luettgen et al. (1993), Malfait and Roose (1997), Simoncelli (1997), Timmermann and Nowak (1997), and Vidakovic (1998). These models have been shown to be useful and adequate for a wide range of problems, and lead to (signal and image) processing methods which are much less computationally demanding than those obtained from the classical MRF framework. The computational efficiency stems from the fact that the joint probability distribution associated with causal multiscale models can be specified in terms of conditional probability functions, instead of MRF clique potential functions which are generally much more difficult to work with. This chapter focuses on one multiscale model in particular, the multiscale hidden Markov model (MHMM), which is a generalization of the wavelet-domain HMM developed by Crouse et al. (1996, 1998) for Gaussian observation models. The MHMM is a probabilistic graph model constructed on a quadtree[1] (or binary tree in the case of one-dimensional signals) associated with multiscale image analysis. MHMMs capture the key inter-scale dependencies present in natural signals and images.

The relationship between classical MRFs and multiscale models has been extensively studied by Gidas (1989), Luettgen et al. (1993), and Pérez and Heitz (1996). It is well known that most multiscale models display *long-range* dependencies, and do not, in general, possess a local Markovian property at all scales as shown by Pérez and Heitz (1996). However,

[1]See Mallat (1998) for general information on the tree structures associated with wavelet and multiscale analysis.

since many natural signals do display long-range dependencies, the non-local behavior of multiscale models may be quite desirable. In fact, certain multiscale models generate $1/f$ random processes[2] which appear to be very well matched to the spectral characteristics of natural imagery as demonstrated in the comprehensive study of van der Schaaf and van Hateren (1996). See the work of Wornell (1996), Nowak (1998), and Timmermann and Nowak (1999) for more information on $1/f$ processes and multiscale analysis.

This chapter is organized as follows. Section 2 reviews the basic multiscale analysis of Gaussian and Poisson processes, which are two of the most commonly encountered data models in image processing. Section 3 considers simple "independent parameter" prior probability models for Bayesian multiscale analysis. Section 4 studies a more sophisticated prior model, the MHMM, that moves beyond the assumption of independence and that better reflects the characteristics of natural signals and images. To keep the presentation as simple as possible, Sections 2, 3, and 4 focus on the one-dimensional (signal) setting. Section 5 discusses some additional issues arising in the two-dimensional (image) setting. Section 6 examines two applications of this framework to image analysis. Conclusions are made in Section 7. In order to deal with both Gaussian and Poisson problems within the same general framework, multiscale analyses and models based on a Haar multiscale analysis are emphasized throughout the chapter. In the Gaussian case, an analogous MHMM framework was developed by Crouse et al. (1998) for multiscale analysis based on any orthogonal wavelet basis.

## 2 Multiscale Data Analysis

In signal and image analysis applications, two of the most common observation models are the Gaussian and Poisson models, see Castleman (1996) for further discussion. For simplicity, let us consider these two models in one dimension; extensions to two dimensions are discussed in Section 5. The Gaussian observation model is:

$$x_k = \mu_k + w_k, \quad k = 0, \ldots, 2^J - 1, \tag{1}$$

where $\mathbf{x} = \{x_k\}$ are observations, $\boldsymbol{\mu} = \{\mu_k\}$ are interpreted as signal samples, and $\{w_k\}$ are realizations of a Gaussian noise process. For convenience in the subsequent multiscale analysis, we adopt the usual convention that the length of the signal is a power of 2, however it is possible to deal with signals of arbitrary length in a multiscale framework. The $\{w_k\}$ are independent, identically distributed samples of a zero-mean Gaussian random

[2] A $1/f$ process is a random process whose power spectrum behaves like $1/|f|^\gamma$, for some power $\gamma > 0$.

variable with *known* variance $\sigma^2$, leading to the likelihood function

$$p(\mathbf{x} \,|\, \boldsymbol{\mu}) = \prod_{k=0}^{2^J-1} \mathcal{N}(x_k \,|\, \mu_k, \sigma^2), \tag{2}$$

where $\mathcal{N}(x \,|\, \mu, \sigma^2)$ denotes a Gaussian density with mean $\mu$ and variance $\sigma^2$ evaluated at the point $x$.

In the Poisson case, the data are (conditionally) independent

$$x_k \sim \mathcal{P}(x_k \,|\, \mu_k), \quad k = 0, \ldots, 2^J - 1, \tag{3}$$

where $\mathcal{P}(x \,|\, \mu)$ denotes the Poisson mass function with intensity $\mu$ evaluated at the point $x$. The likelihood function in this case is simply

$$p(\mathbf{x} \,|\, \boldsymbol{\mu}) = \prod_{k=0}^{2^J-1} \mathcal{P}(x_k \,|\, \mu_k). \tag{4}$$

Now let us consider a multiscale data analysis. In general, multiscale analysis refers to the study of behavior or structure in signals or data at various spatial and/or temporal resolutions. For further background information on multiscale signal analysis see Mallat (1998). Perhaps the simplest technique is the Haar multiscale analysis defined according to:

$$\begin{aligned} x_{J,k} &\equiv x_k, \quad k = 0, \ldots, 2^J - 1 \\ x_{j,k} &= x_{j+1,2k} + x_{j+1,2k+1}, \quad k = 0, \ldots, 2^j - 1, \ 0 \leq j \leq J - 1. \end{aligned}$$

The index $j$ refers to the resolution of the analysis, $2^j$; $j = J$ being the highest resolution (finest scale), and $j = 0$ being the lowest resolution (coarsest scale). This multiscale data analysis is organized into the binary data tree shown in Figure 1.

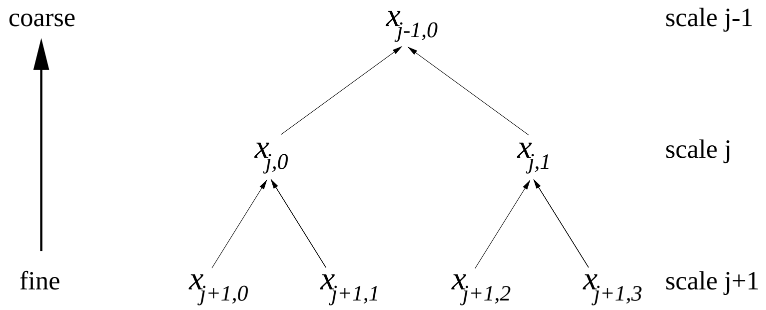



FIGURE 1. *Binary data tree associated with multiscale (fine-to-coarse) analysis. In this figure, analysis begins at fine scale $j+1$ and produces coarser representations of data at scales $j$ and $j-1$.*

The data $\{x_{j,k}\}$ are the (unnormalized) Haar scaling coefficients of $\mathbf{x}$. The relationship between a "parent" (*e.g.*, $x_{j,k}$) and a "child" (*e.g.*, $x_{j+1,2k}$)

is of fundamental interest in multiscale data analysis. Specifically, given an observation model, *e.g.*, Gaussian, we are interested in the conditional distribution of the child given the parent.

The parent-child relationship is expressed by the conditional likelihood $p(x_{j+1,2k}\,|\,x_{j,k},\boldsymbol{\mu})$, which happens to have a very simple form for both the Gaussian and Poisson observation models. Note that it is unnecessary to consider the conditional likelihoods of both children due to the fact that $x_{j+1,2k+1}$ is uniquely determined by $x_{j+1,2k}$ and $x_{j,k}$. Before examining the conditional likelihoods, let us define a multiscale analysis of the parameter $\boldsymbol{\mu}$, analogous to that defined for the data $\mathbf{x}$:

$$\begin{aligned} \mu_{J,k} &\equiv \mu_k, \quad k=0,\ldots,2^J-1 \\ \mu_{j,k} &= \mu_{j+1,2k}+\mu_{j+1,2k+1}, \quad k=0,\ldots,2^j-1,\ 0\le j\le J-1. \end{aligned}$$

The parameters $\{\mu_{j,k}\}$ are the (unnormalized) Haar scaling coefficients of $\boldsymbol{\mu}$. With this definition in hand, and using standard conditional probability relationships between pairs of Gaussian and Poisson random variables, we have the following expressions for the parent-child conditional likelihoods.

- Gaussian model:

$$p(x_{j+1,2k}\,|\,x_{j,k},\boldsymbol{\mu}) \;=\; \mathcal{N}\left(x_{j+1,2k}\,\Big|\,\frac{x_{j,k}}{2}+\frac{\mu_{j+1,2k}-\mu_{j+1,2k+1}}{2},\frac{\sigma_{j+1}^2}{2}\right),$$

where $\sigma_j^2=2^{J-j}\sigma^2$; (5)

- Poisson model:

$$p(x_{j+1,2k}\,|\,x_{j,k},\boldsymbol{\mu}) \;=\; \mathcal{B}\left(x_{j+1,2k}\;\Big|\;x_{j,k},\frac{\mu_{j+1,2k}}{\mu_{j,k}}\right), \qquad (6)$$

where $\mathcal{B}(x\,|\,n,\theta)=\binom{n}{x}\theta^x(1-\theta)^{n-x}$, denotes the binomial distribution with parameters $n$ and $\theta$. From these expressions, we identify the *canonical* multiscale parameters associated with the two models. In the Gaussian case, the canonical parameter is $\theta_{j,k}=\mu_{j+1,2k}-\mu_{j+1,2k+1}$, which is simply the (unnormalized[3]) Haar wavelet coefficient of the mean (signal) $\boldsymbol{\mu}$ at resolution $2^j$ and location $k$. In the Poisson case, the canonical parameter is $\theta_{j,k}=\frac{\mu_{j+1,2k}}{\mu_{j,k}}$, which can be viewed as a "splitting" factor[4] that governs the multiscale refinement of the intensity $\boldsymbol{\mu}$. This type of multiscale intensity analysis was introduced independently by Timmermann and Nowak (1997) and Kolaczyk (1998). Also see Timmermann and Nowak (1999) for further

[3] The normalized Haar wavelet coefficients are obtained by the mapping $\theta_{j,k}\mapsto 2^{(j-J)/2}\theta_{j,k}$.

[4] Note that the splitting factors are also closely related to the Haar wavelet coefficients since $\frac{\mu_{j+1,2k}}{\mu_{j,k}}=\frac{1}{2}\left(1+\frac{\mu_{j+1,2k}-\mu_{j+1,2k+1}}{\mu_{j,k}}\right)$.

details. As discussed in Section 3, the canonical multiscale parameters suggest the use of special prior distributions that complement the observation model leading to tractable and highly efficient processing strategies.

The simplicity of these relationships (well-known parametric distributions) is quite exceptional; in general, under other observation models, the parent-child relationship can be much more complicated, and usually does not admit a standard parametric form. For this reason, one can argue that multiscale analysis is especially well suited to Gaussian and Poisson data. In particular, in either case, one can factorize the likelihood function as follows:

$$p(\mathbf{x} \,|\, \boldsymbol{\mu}) \equiv p(x_{0,0} \,|\, \mu_{0,0}) \prod_{j=0}^{J-1} \prod_{k=0}^{2^j-1} p(x_{j+1,2k} \,|\, x_{j,k}, \theta_{j,k}), \tag{7}$$

where $p(x_{j+1,2k} \,|\, x_{j,k}, \theta_{j,k})$ is given by (5) in the Gaussian case, and (6) in the Poisson case. This factorization is possible because the linear mapping of observations to scaling coefficients, $\mathbf{x} = \{x_k\} \mapsto \{x_{j,2k}\}$, has a unit Jacobian.

Note that in the Gaussian case the likelihood can also be expressed equivalently in terms of the wavelet coefficients of the observation $\mathbf{x}$. That is, the Gaussian likelihood of $\mathbf{x}$ given $\boldsymbol{\mu}$ can be written as a product of univariate Gaussian likelihoods, each involving a single "data" wavelet coefficient $x_{j+1,2k} - x_{j+1,2k+1}$ given the corresponding signal wavelet coefficient $\theta_{j,k} = \mu_{j+1,2k} - \mu_{j+1,2k+1}$. This is a more standard likelihood factorization for the Gaussian case, and it is possible because of the orthogonality of the discrete wavelet transform and the fact that the Gaussian likelihood structure is preserved under orthogonal linear transformations. In fact, this alternative "wavelet-based" factorization can be employed in conjunction with any orthogonal wavelet system, the Haar being one special case.

A similar wavelet-based factorization is not possible in the Poisson case; the difficulty lies in the fact that the Poisson distribution reproduces under straight (unweighted) summation (the sum of Poisson random variables is still Poisson), but not under rescaling. Therefore we use the factorization (7) above in order to treat both Gaussian and Poisson models within a common framework. For more information on this likelihood factorization and a discussion of its fundamental role in multiscale statistical analysis in general, see Kolaczyk (1999) in this volume.

The likelihood factorization also greatly facilitates multiscale analysis and modeling. For example, estimates of the multiscale parameters $\boldsymbol{\theta} = \{\theta_{j,k}\}$ can be used to reconstruct an estimate of the underlying signal or intensity. In Section 6, we will look at two Bayesian image analysis applications based on the multiscale parameters. In general, Bayesian inference based on the multiscale parameters $\boldsymbol{\theta}$ requires; (1) specification of a suitable prior probability model for $\boldsymbol{\theta}$; (2) determination of the posterior probability distribution of $\boldsymbol{\theta}$ resulting from the likelihood and prior. The next two sections examine two types of prior probability models and their respective posterior distributions. Section 3 considers a simple approach in which $\boldsymbol{\theta}$ are modeled as independent random variables with prior probability densities designed to reflect the characteristics of natural signals and images and that lead to simple expressions for the posterior density. Section 4 examines a more sophisticated prior model, the MHMM, that moves beyond the assumption of independence and captures the parent-child dependencies also encountered in practice. To keep the notation and derivations as simple and clear as possible, throughout the remainder of the chapter we assume that the scaling coefficient at the coarsest scale, $\mu_{0,0}$, is known. The Bayesian modeling and analysis methods described next can be easily extended to include prior models and inference schemes that include $\mu_{0,0}$ as well.

# 3 Independent Parameter Models

Let us now consider prior probability models for the (unknown) canonical multiscale parameters $\boldsymbol{\theta}$. *Conjugate priors* are advantageous for computational reasons since the posterior distribution is obtained by simply "updating" the parameters of the prior based on the observations; see (Robert, 1994, pp. 97-111) for further information. Moreover, we will see that conjugate priors can provide very plausible models for the multiscale parameters. For the Gaussian likelihood function, the *natural* conjugate prior is also Gaussian. Hence, a simple approach is to model each multiscale parameter as an independent Gaussian $\theta \sim \mathcal{N}(\theta \,|\, 0, \tau^2)$. In the Poisson case, the natural conjugate prior for the binomial distribution is a beta density. Therefore, we model each multiscale parameter as an independent beta distributed random variable, $\theta \sim \mathcal{B}e(\theta \,|\, \alpha, \beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}$, $0 \leq \theta \leq 1$, where $B(\alpha, \beta)$ denotes the standard beta function. In this chapter, we will only use symmetric beta priors of mean $1/2$, characterized by $\alpha = \beta$. Note that the Gaussian is also a symmetric prior (about its mean, in this case, zero). In both cases, we choose a symmetric prior for $\theta$ since there is no *a priori* support for asymmetry. Also, in general, the variance $\tau^2$ or parameter $\alpha$ may depend on the resolution $2^j$. Here, as in most related approaches, the parameters do not depend on the location $k$, since location dependent signal characteristics are usually not known *a priori*. Thus, a simple model for the unknown parameters $\boldsymbol{\theta}$ is

$$p(\boldsymbol{\theta}) \;=\; \prod_{j=0}^{J-1} \prod_{k=0}^{2^j-1} p(\theta_{j,k}), \qquad (8)$$

with $p(\theta_{j,k})$ equal to $\mathcal{N}(\theta_{j,k} \,|\, 0, \tau_j^2)$ or $\mathcal{B}e(\theta_{j,k} \,|\, \alpha_j, \alpha_j)$ for the Gaussian or Poisson case, respectively.

Although this "independent parameter" prior may appear too simplistic, in certain cases it is quite reasonable because multiscale decompositions tend to decorrelate signals and images. For example, if the signal is a fractional Brownian motion or $1/f$ process, then the correlations between Haar wavelet coefficients decay very rapidly across scale and space as demonstrated by Flandrin (1992) and Wornell (1996). Moreover, it can be shown that with special choices of $\{\tau_j^2\}$ or $\{\alpha_j\}$ the prior above (8) displays $1/f$-like behavior, see Nowak (1998) and Timmermann and Nowak (1999) for more information.

In both the Gaussian and Poisson cases, combining the prior (8) with the likelihood (7) produces a posterior density

$$p(\boldsymbol{\theta}\,|\,\mathbf{x}) \;=\; \prod_{j=0}^{J-1}\prod_{k=0}^{2^j-1} p(\theta_{j,k}\,|\,x_{j,k}, x_{j+1,2k}), \tag{9}$$

where

- Gaussian model:

$$p(\theta_{j,k}\,|\,x_{j,k}, x_{j+1,2k}) \;=\; \mathcal{N}\left(\theta_{j,k}\,|\,\frac{\tau_j^2\,(2x_{j+1,2k}-x_{j,k})}{\tau_j^2+\sigma_j^2}, \frac{\sigma_j^2\tau_j^2}{\tau_j^2+\sigma_j^2}\right);$$

- Poisson model:

$$p(\theta_{j,k}\,|\,x_{j,k}, x_{j+1,2k}) \;=\; \mathcal{B}e\left(\theta_{j,k}\,|\,\alpha_j + x_{j+1,2k}, \alpha_j + x_{j,k} - x_{j+1,2k}\right).$$

The marginal posteriors above can be easily derived; also see (Robert, 1994, p. 104) for general forms of the posterior distributions resulting from these conjugate priors). The factorization of the posterior shows that inferences can be made on each multiscale parameter individually, instead of requiring a complicated high dimensional analysis. For example, it is straightforward to obtain the posterior mean or maximum *a posteriori* (MAP) estimate of each individual parameter, based on the one-dimensional Gaussian or beta posterior densities above. Similarly, other meaningful quantities such as posterior variances and confidence regions can also be easily computed because the posterior factorizes into one-dimensional parametric densities. Some specific examples in image analysis are considered in Section 6. Finally, notice that the analysis presented for the Gaussian case can clearly be generalized to other orthogonal wavelet bases following the work of Crouse et al. (1998).

### *3.1 Mixture Density Priors*

A richer class of priors, more suitable for modeling the multiscale parameters of natural signals and images, is provided by mixture densities. Specifically, one can build larger classes of priors using mixtures of the *elementary* conjugate densities mentioned above; these mixture priors are still conjugate, with all the associated computational convenience; see (Robert, 1994,

p. 108) for more details. For the Gaussian observation model, Gaussian mixtures are often used as in Abramovich et al. (1998), Chipman et al. (1997), and Crouse et al. (1998), while for the Poisson likelihood, beta mixtures are adopted as in Timmermann and Nowak (1999). Formally:

- Gaussian model:

$$p(\theta_{j,k}) = \sum_{m=0}^{M-1} \rho_j(m)\, \mathcal{N}\left(\theta_{j,k} \,|\, 0, \tau_{j,m}^2\right); \tag{10}$$

- Poisson model:

$$p(\theta_{j,k}) = \sum_{m=0}^{M-1} \rho_j(m)\, \mathcal{B}e\left(\theta_{j,k} \,|\, \alpha_{j,m}, \alpha_{j,m}\right), \tag{11}$$

where $\{\rho_j(m)\}_{m=0}^{M-1}$ denote the *a priori* probabilities of each component at scale $j$. To keep the notation simple, we will use $M$ component mixtures at all scales.

The motivation for using mixtures is based on the following reasoning. If we believe that the underlying signal $\boldsymbol{\mu}$ is generally smooth, except for (possibly) a few large singularities, then, for example, a mixture consisting of a highly probable low-variance component (to model the smooth areas of the signal) and a relatively low probability high-variance component (to model the possible singularities) is intuitively reasonable. A state (also called *latent* or *indicator*) variable $s_{j,k}$ is usually associated with each parameter $\theta_{j,k}$. The state takes values that indicate which component of the mixture is in effect; for example, in the Gaussian model, $p(\theta_{j,k} \,|\, s_{j,k} = m) = \mathcal{N}\left(\theta_{j,k} \,|\, 0, \tau_{j,m}^2\right)$. The prior probabilities for each state are the *a priori* mixture weights, *i.e.*, $p(s_{j,k} = m) = \rho_j(m)$. These probabilities, along with the density shape parameters associated with them (i.e., the values of $\tau_{j,m}^2$ and $\alpha_{j,m}$), can be chosen based on prior beliefs about the regularity of the class of signals/images in question following the work of Abramovich et al. (1998), or can be inferred from the observed data through a hierarchical Bayes setting (possibly via an empirical Bayes approach) as in the work of Chipman et al. (1997), Crouse et al. (1998), Timmermann and Nowak (1997, 1999), and Kolaczyk (1998).

The posterior distribution has the same factorized form as (9) with mixture densities in place of the corresponding single component densities. Let $\mathbf{s} = \{s_{j,k}\}$ denote the set of state variables and $\mathbf{m} = \{m_{j,k}\}$ denote a set of state values. We can then write

$$p(\boldsymbol{\theta}, \mathbf{s} = \mathbf{m} \,|\, \mathbf{x}) = p(\boldsymbol{\theta} \,|\, \mathbf{s} = \mathbf{m}, \mathbf{x})\, p(\mathbf{s} = \mathbf{m} \,|\, \mathbf{x}) \tag{12}$$

and

$$p(\boldsymbol{\theta} \,|\, \mathbf{x}) = \sum_{\mathbf{m}} p(\boldsymbol{\theta} \,|\, \mathbf{s} = \mathbf{m}, \mathbf{x})\, p(\mathbf{s} = \mathbf{m} \,|\, \mathbf{x}), \tag{13}$$

where the sum is over all possible sets of state values. The "state-conditional" density $p(\boldsymbol{\theta}\,|\,\mathbf{s}=\mathbf{m},\mathbf{x})$ factorizes just as in (9)

$$p(\boldsymbol{\theta}\,|\,\mathbf{s}=\mathbf{m},\mathbf{x}) \;=\; \prod_{j=0}^{J-1}\prod_{k=0}^{2^j-1} p(\theta_{j,k}\,|\,x_{j,k},x_{j+1,2k},s_{j,k}=m_{j,k}), \tag{14}$$

The posterior state probabilities can be very efficiently computed due to the likelihood factorization (7). Note that

$$\begin{aligned} p(\mathbf{s}=\mathbf{m}\,|\,\mathbf{x}) \;&=\; \int p(\mathbf{s}=\mathbf{m},\boldsymbol{\theta}\,|\,\mathbf{x})\,d\boldsymbol{\theta} \\ &\propto\; \int p(\mathbf{x}\,|\,\mathbf{s}=\mathbf{m},\boldsymbol{\theta})p(\boldsymbol{\theta}\,|\,\mathbf{s}=\mathbf{m})\,p(\mathbf{s}=\mathbf{m})\,d\boldsymbol{\theta} \\ &\propto\; \prod_{j=0}^{J-1}\prod_{k=0}^{2^j-1}\int p(x_{j+1,2k}\,|\,x_{j,k},\theta_{j,k}) \\ &\qquad\qquad \times\, p(\theta_{j,k}\,|\,s_{j,k}=m_{j,k})\,p(s_{j,k}=m_{j,k})\,d\theta_{j,k}. \end{aligned}$$

This shows that

$$p(s_{j,k}=m\,|\,\mathbf{x}) \;=\; \frac{\rho_j(m)\,L_{j,k}(m)}{\sum_{m=0}^{M-1}\rho_j(m)\,L_{j,k}(m)}, \tag{15}$$

where

$$L_{j,k}(m) \;\equiv\; \int p(x_{j+1,2k}\,|\,x_{j,k},\theta_{j,k})\,p(\theta_{j,k}\,|\,s_{j,k}=m)\,d\theta_{j,k}. \tag{16}$$

$L_{j,k}(m)$ is a marginal likelihood, *i.e.*, $L_{j,k}(m)=p(x_{j+1,2k}|x_{j,k},s_{j,k}=m)$, and it has a simple closed-form expression in both the Gaussian and Poisson cases:

- Gaussian model:

$$L_{j,k}(m) \;\propto\; \frac{1}{\left(\sigma_j^2+\tau_{j,m}^2\right)^{1/2}}\exp\left(-\frac{(2x_{j+1,2k}-x_{j,k})^2}{2\left(\sigma_j^2+\tau_{j,m}^2\right)}\right);$$

- Poisson model:

$$L_{j,k}(m) \;\propto\; \frac{B\left(x_{j+1,2k}+\alpha_{j,m},x_{j,k}-x_{j+1,2k}+\alpha_{j,m}\right)}{B\left(\alpha_{j,m},\alpha_{j,m}\right)}.$$

Therefore, the posterior density $p(\boldsymbol{\theta}\,|\,\mathbf{x})$ is given by

$$p(\boldsymbol{\theta}\,|\,\mathbf{x}) \;=\; \prod_{j=0}^{J-1}\prod_{k=0}^{2^j-1}\sum_{m=0}^{M-1} p(s_{j,k}=m\,|\,\mathbf{x})\,p(\theta_{j,k}\,|\,x_{j,k},x_{j+1,2k},s_{j,k}=m), \tag{17}$$

where $p(\theta_{j,k}\,|\,x_{j,k}, x_{j+1,2k}, s_{j,k} = m)$ denotes a Gaussian or Beta density of the forms given in (9), with shape parameter $\tau^2_{j,m}$ or $\alpha_{j,m}$, respectively. Again, the factorization of the posterior and the simple parametric form of each factor shows that inferences can be easily made on each multiscale parameter individually. For example, mixture prior densities typically lead to posterior mean or MAP estimators, as in Abramovich et al. (1998), Chipman et al. (1997), Crouse et al. (1998), Timmermann and Nowak (1997, 1999), that resemble the non-linear shrinkage/thresholding estimators encountered in standard non-Bayesian approaches to wavelet-based noise removal like the now classical denoising methods developed by Donoho and Johnstone (1994).

# 4 Multiscale Hidden Markov Models

The multiscale hidden Markov model (MHMM) is a graphical model based on the binary tree (or quadtree in the case of image data) associated with multiscale signal analysis. One instance of the MHMM is the wavelet domain HMM developed in Crouse et al. (1996, 1998) for Gaussian observation models. Here, we focus on the Haar multiscale analysis discussed above, and develop a more general framework encompassing both the Gaussian and Poisson models. The priors described in Section 3 modeled the multiscale parameters $\boldsymbol{\theta}$ as independent mixture random variables. The MHMM moves beyond this simple prior, by specifying probabilistic dependencies between the states underlying the mixtures of parent and child multiscale parameters.

The MHMM is a *directed acyclic graph* (or *Bayesian network*), as depicted in Figure 2; see Pearl (1988) for more information on graphical models in general. The MHMM has a causal coarse-to-fine[5] scale structure indicated by the direction of the arrows. Specifically, the MHMM is based on the assumption that the value of each state $s_{j,k}$ is *caused* by the value of its parent state $s_{j-1,\lfloor k/2\rfloor}$, where $\lfloor k/2\rfloor$ is the largest integer less than or equal to $k/2$. This means that, given the value of its parent's state, $s_{j,k}$ is independent of all other states at scales $i \leq j$ (at same level and above) in the tree. This enables the following factorization of the joint state probability function

$$p(\mathbf{s}) \quad = \quad \prod_{j=0}^{J-1}\prod_{k=0}^{2^j-1} p\left(s_{j,k} = m_{j,k}\,|\,s_{j-1,\lfloor k/2\rfloor} = m_{j-1,\lfloor k/2\rfloor}\right),$$

with the convention $p(s_{0,0}\,|\,s_{-1,0}) \equiv p(s_{0,0})$. This is a more structured alternative to the independent parameter model previously considered. Another important property of the MHMM is that, given their respective state val-

[5]Alternatively, we could construct a causal fine-to-coarse model.

ues, all parameters $\boldsymbol{\theta}$ are conditionally independent. That is,

$$p(\boldsymbol{\theta} \,|\, \mathbf{s} = \mathbf{m}) \;=\; \prod_{j=0}^{J-1} \prod_{k=0}^{2^j-1} p(\theta_{j,k} \,|\, s_{j,k} = m_{j,k}). \tag{18}$$

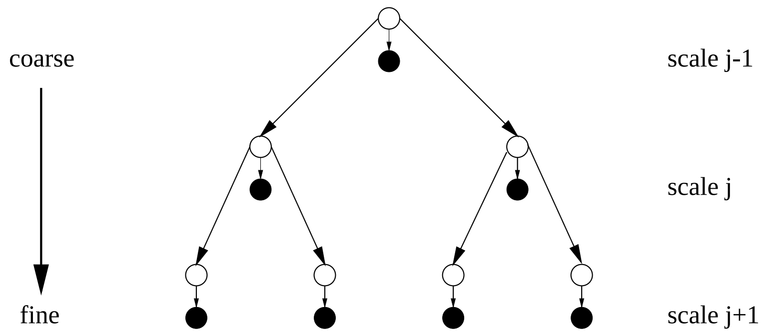
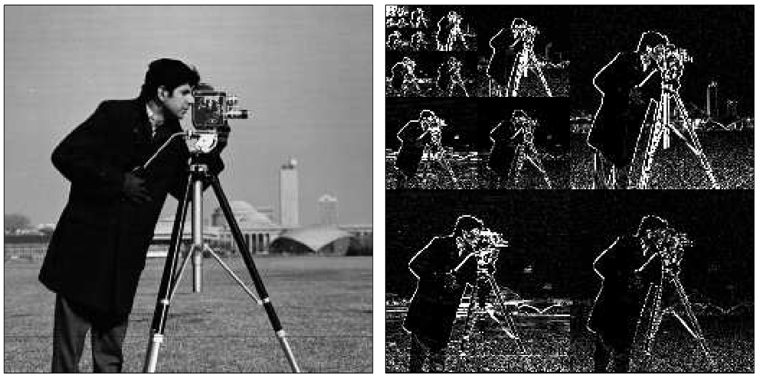



FIGURE 2. *Multiscale HMM graph. Each multiscale parameter of the signal (or intensity) $\boldsymbol{\mu}$ is modeled as a mixture random variable (black node) with an associated hidden state variable (white node). To match the inter-scale dependencies, we link the hidden states across scale. The MHMM is a directed graph that synthesizes a signal in a coarse-to-fine fashion indicated by the direction of the arrows. These connections capture the key inter-scale dependencies present in many natural signals and images.*

This MHMM structure provides a mechanism for sharing (and exploiting) relevant inter-scale information and appears to be well justified by the empirical studies of Shapiro (1993) and Crouse et al. (1998); in fact, it is similar to the principles underlying some of the most successful wavelet based image compression algorithms known today such as the “zerotree” algorithm of Shapiro (1993). The MHMM captures the key inter-scale dependencies present in natural signals and images. These dependencies, termed *clustering* and *persistence-across-scale* by Crouse et al. (1998), refer to the fact that “high energy” multiscale parameters (*e.g.,* Haar wavelet coefficients) tend to cluster near edges in images, and that similar clusters are seen at multiple analysis scales indicating persistence-across-scale, as illustrated in Figure 3.

As a concrete example, consider the Gaussian case and suppose that there are two states associated with each multiscale parameter (Haar wavelet coefficient); state ‘0’ corresponds to a low-variance Gaussian modeling the absence of a singularity, while state ‘1’ is a high-variance Gaussian expressing the possible presence of a singularity (or edge). Because singularities, like edges in images, tend to persist across scales, if the parent state $s_{j-1,\lfloor k/2 \rfloor} = 1$, then it is probable that the child state $s_{j,k} = 1$, as well. Likewise, if $s_{j-1,\lfloor k/2 \rfloor} = 0$, indicating that signal is fairly smooth in the re-



(a) (b)

FIGURE 3. *Persistence and clustering in an image. (a) Test image. (b) Magnitude of 2-d Haar DWT of test image. Each sub-image in (b) depicts the set of wavelet coefficients (white denotes largest magnitudes) at a specific scale and orientation (horizontal, vertical, or diagonal). For example, the three largest sub-images depict the wavelet coefficients at scale $J-1$ and at horizontal (lower-left), vertical (upper-right), and diagonal (lower-right) orientations. The smaller sub-images depict the corresponding sets of wavelet coefficients at progressively coarser scales.*

gion corresponding to $\theta_{j-1,\lfloor k/2 \rfloor}$, then with high probability the sub-region corresponding to $\theta_{j,k}$ is also smooth and $s_{j,k} = 0$.

Now let us determine the posterior density associated with the MHMM. The posterior $p(\boldsymbol{\theta} \,|\, \mathbf{x})$ takes a form similar to (17), except that in this case the posterior state probabilities cannot be determined independently. However, $p(\boldsymbol{\theta} \,|\, \mathbf{x})$ can be efficiently computed as follows. As in (13), given $p(\mathbf{s} = \mathbf{m} \,|\, \mathbf{x})$, we compute

$$p(\boldsymbol{\theta} \,|\, \mathbf{x}) = \sum_{\mathbf{m}} p(\boldsymbol{\theta} \,|\, \mathbf{s} = \mathbf{m}, \mathbf{x})\, p(\mathbf{s} = \mathbf{m} \,|\, \mathbf{x}), \tag{19}$$

where, because of (18), $p(\boldsymbol{\theta} \,|\, \mathbf{s} = \mathbf{m}, \mathbf{x})$ is the state-conditional posterior density given by (14). So all that remains is to compute $p(\mathbf{s} = \mathbf{m} \,|\, \mathbf{x})$:

$$\begin{aligned}
p(\mathbf{s} = \mathbf{m} \,|\, \mathbf{x}) &= \int p(\mathbf{s} = \mathbf{m}, \boldsymbol{\theta} \,|\, \mathbf{x})\, d\boldsymbol{\theta} \\
&\propto \int p(\mathbf{x} \,|\, \mathbf{s} = \mathbf{m}, \boldsymbol{\theta}) p(\boldsymbol{\theta} \,|\, \mathbf{s} = \mathbf{m})\, p(\mathbf{s} = \mathbf{m})\, d\boldsymbol{\theta} \\
&\propto \prod_{j=0}^{J-1} \prod_{k=0}^{2^j-1} \int p(x_{j+1,2k} \,|\, x_{j,k}, \theta_{j,k}, s_{j,k} = m_{j,k}) \\
&\quad \times p(\theta_{j,k} | s_{j,k} = m)\, p(s_{j,k} = m_{j,k} | s_{j-1,\lfloor \frac{k}{2} \rfloor} = m_{j-1,\lfloor \frac{k}{2} \rfloor})\, d\theta_{j,k} \\
&= \prod_{j=0}^{J-1} \prod_{k=0}^{2^j-1} p(s_{j,k} = m_{j,k} | s_{j-1,\lfloor \frac{k}{2} \rfloor} = m_{j-1,\lfloor \frac{k}{2} \rfloor})\, L_{j,k}(m_{j,k}),
\end{aligned}$$

where $L_{j,k}(m_{j,k})$ is defined in (16).

With this expression we can calculate the (marginal) posterior state probabilities using an *upward-downward probability propagation algorithm* (also called a forward-backward algorithm). Then, with posterior state probabilities in hand, we can compute the posterior density of the multiscale parameters according to (17). In the upward-downward algorithm, the **Up Step** recursively marginalizes (sub-tree by sub-tree) the joint posterior state probability beginning at the finest scale $j = J - 1$ (bottom of tree) up to the coarsest scale $j = 0$ (top "root" of tree). This provides us with the posterior state probabilities $\{p(s_{0,0} = m \,|\, \mathbf{x})\}_{m=0}^{M-1}$. The **Down Step** computes the marginal posterior state probabilities for each $s_{j,k}$ in a recursive fashion, making use of partial (sub-tree) marginalizations previously calculated in the **Up Step**. See Frey (1998) for a general overview of upward-downward (forward-backward) algorithms, their extensions, and connections to other inference methods in graphical models. For notational convenience, define $\rho_{j,k}(m|n) \equiv p(s_{j,k} = m \,|\, s_{j-1,\lfloor k/2 \rfloor} = n)$.

### Upward-Downward Propagation Algorithm

**Up Step**

Beginning at $j = J - 1$ compute

$$q_{j,k}(n) \;=\; \sum_{m=0}^{M-1} \rho_{j,k}(m|n) L_{j,k}(m). \tag{20}$$

Then for $j = J - 2, \ldots, 1$

$$q_{j,k}(n) = \sum_{m=0}^{M-1} q_{j+1,2k}(m)\, q_{j+1,2k+1}(m)\, \rho_{j,k}(m|n) L_{j,k}(m) \tag{21}$$

and for $j = 0$

$$q_{0,0}(n) = q_{1,0}(n)\, q_{1,1}(n)\, \rho_{0,0}(n) L_{0,0}(n), \tag{22}$$

where $\rho_{0,0}(n) = p(s_{0,0} = n)$.

Note that $q_{j,k}(n)$ is the partial marginalization at node $j, k$ over all states in the subtree beneath it. Hence, the final quantities $\{q_{0,0}(m)\}$ are the (unnormalized) posterior state probabilities $\{p(s_{0,0} = m \,|\, \mathbf{x})\}_{m=0}^{M-1}$ at the top (root) of the graph.

**Down Step**

Beginning the posterior states probabilities at scale $j = 0$, set $p_{0,0}(m) = q_{0,0}(m)$. Then for $j = 1, \ldots, J - 2$

$$p_{j,k}(m) = \sum_{n=0}^{M-1} \frac{p_{j-1,\lfloor \frac{k}{2} \rfloor}(n)\, \rho_{j,k}(m|n)\, q_{j+1,2k}(m)\, q_{j+1,2k+1}(m)\, L_{j,k}(m)}{q_{j,k}(n)} \tag{23}$$

and for $j = J-1$

$$p_{j,k}(m) \quad = \quad \sum_{n=0}^{M-1} \frac{p_{j-1,\lfloor \frac{k}{2} \rfloor}(n)\, \rho_{j,k}(m|n)\, L_{j,k}(m)}{q_{j,k}(n)}.$$

The final quantities $\{p_{j,k}(m)\}_{m=0}^{M-1}$ are the (unnormalized) posterior state probabilities $\{p(s_{j,k} = m \,|\, \mathbf{x})\}_{m=0}^{M-1}$.

Although the MHMM posterior density is more complicated than that of the independent multiscale parameter models considered in Section 3, the upward-downward algorithm provides us with a very efficient way of calculating the marginal posterior probabilities of the states. With these probabilities in hand, the factorization of the state-conditional posterior density (14) for the multiscale parameters shows that inference can be carried out on each multiscale parameter individually, just as in the independent multiscale parameter cases. The total computational complexity is $O(2^J)$, where $2^J$ is the number of data.

Finally, note that other graph structures (such as those including intrascale dependencies) may be considered. However, computationally efficient algorithms may not be available; it is known that exact inference on general graphs is an NP-hard problem as discussed by Cooper (1990).

# 5 Extensions to Two Dimensions

The multiscale analyses and MHMMs can be easily extended to two dimensions. In two dimensions (2-d) the unnormalized Haar multiscale data analysis is as follows. We begin with data $\{x_{k,l}\}$, $k,l = 0,\ldots,2^J-1$, and define

$$\begin{aligned} x_{J,k,l} &\equiv x_{k,l}, \quad k,l = 0,\ldots,2^J-1 \\ x_{j,k,l} &= x_{j+1,2k,2l} + x_{j+1,2k+1,2l} + x_{j+1,2k,2l+1} + x_{j+1,2k+1,2l+1}, \\ &\qquad k,l = 0,\ldots,2^j-1,\ 0 \le j \le J-1. \end{aligned}$$

Again, the index $j$ refers to the resolution of the analysis, $2^j$; $j = J$ and $j = 0$ are the highest (finest) and lowest (coarsest) resolutions (scales), respectively. This multiscale data analysis is organized into a so-called "quadtree" (the obvious generalization of the binary data tree in Figure 1); see Crouse et al. (1998) for information on quadtree representations.

We can also construct an analogous 2-d multiscale analysis of an image $\boldsymbol{\mu}$. However, there are some additional issues faced in 2-d that distinguish the Gaussian and Poisson case. The standard (unnormalized) 2-d Haar wavelet coefficients are computed as follows. Let $\{\mu_{j,2k+m,2l+n}\}_{m,n=0}^{1}$ denote four neighboring scaling coefficients at scale $j$ in a 2-d Haar analysis of an image

$\boldsymbol{\mu}$. The (unnormalized) Haar scaling coefficient and wavelet coefficients at scale $j-1$ are

$$\begin{aligned}
\mu_{j-1,k,l} &= \mu_{j,2k,2l} + \mu_{j,2k,2l+1} + \mu_{j,2k+1,2l} + \mu_{j,2k+1,2l+1}\,, \\
\theta^1_{j-1,k,l} &= \mu_{j,2k,2l} + \mu_{j,2k,2l+1} - \mu_{j,2k+1,2l} - \mu_{j,2k+1,2l+1}\,, \\
\theta^2_{j-1,k,l} &= \mu_{j,2k,2l} - \mu_{j,2k,2l+1} + \mu_{j,2k+1,2l} - \mu_{j,2k+1,2l+1}\,, \\
\theta^3_{j-1,k,l} &= \mu_{j,2k,2l} - \mu_{j,2k,2l+1} - \mu_{j,2k+1,2l} + \mu_{j,2k+1,2l+1}\,, \qquad (24)
\end{aligned}$$

where the superscripts 1, 2, and 3 refer to the horizontal, vertical, and diagonal differences, respectively. The Haar wavelet coefficients $\{\theta^i_{j-1,k,l}\}_{i=1}^3$ are the *additive* refinements required to split the coarse scaling coefficient $\mu_{j-1,k,l}$ into the four finer scaling coefficients $\{\mu_{j,2k+m,2l+n}\}_{m,n=0}^1$. Due to the orthogonality of the mapping

$$\{\mu_{j,2k+m,2l+n}\}_{m,n=0}^1 \mapsto \mu_{j-1,k,l}, \{\theta^i_{j-1,k,l}\}_{i=1}^3,$$

in the Gaussian case, taking these (standard) Haar wavelet coefficients as multiscale parameters leads to a factorized likelihood.

The Poisson case is more complicated because orthogonality does not imply independence. However, an alternative set of multiscale parameters can be used in the Poisson case, that does lead to a factorized likelihood. Specifically, we take the 2-d multiscale parameters to be the factors corresponding to the *multiplicative* refinement of a coarse scaling coefficient (intensity) into four finer scaling coefficients by first splitting it horizontally (vertically) into two halves, then next vertically (horizontally) splitting each half into two quarters as described by Timmermann and Nowak (1999). That is, take

$$\begin{aligned}
\mu_{j-1,k,l} &= \mu_{j,2k,2l} + \mu_{j,2k,2l+1} + \mu_{j,2k+1,2l} + \mu_{j,2k+1,2l+1}\,, \\
\theta^1_{j-1,k,l} &= \frac{\mu_{j,2k,2l} + \mu_{j,2k,2l+1}}{\mu_{j,2k,2l} + \mu_{j,2k,2l+1} + \mu_{j,2k+1,2l} + \mu_{j,2k+1,2l+1}}, \\
\theta^2_{j-1,k,l} &= \frac{\mu_{j,2k,2l}}{\mu_{j,2k,2l} + \mu_{j,2k,2l+1}}, \\
\theta^3_{j-1,k,l} &= \frac{\mu_{j,2k+1,2l}}{\mu_{j,2k+1,2l} + \mu_{j,2k+1,2l+1}}. \qquad (25)
\end{aligned}$$

Alternatively, it is possible to consider a fully 2-d refinement process in which we simultaneously split a coarse scaling coefficient into four finer coefficients. In this case the conditional parent-child likelihoods would be multinomially instead of binomial, and the natural conjugate prior would be the Dirichlet rather than the beta density, but otherwise the multiscale framework would be essentially the same.

The 2-d multiscale parameters defined in (24) and (25) can be modeled with the same conjugate prior probability density functions proposed for the 1-d case, (10) and (11), respectively. Furthermore, if the states are modeled independently, then the posterior density of the 2-d parameters also factorizes in a fashion similar to the 1-d case.

The 2-d MHMM is also similar to the 1-d development in Section 4, with a quadtree replacing the binary tree structure. To be precise, in the Gaussian case, the standard 2-d Haar wavelet analysis consists of three sets of wavelet coefficients, superscript 1, 2, or 3 in (24), at each scale, associated with horizontal, vertical, and diagonal differences. In the Poisson case, we have three sets of multiplicative splits, superscript 1, 2, or 3 in (25). A single quadtree MHMM (analogous to the binary tree depicted in Figure 2) is associated with each set of multiscale parameters. For example, in the Gaussian case, one quadtree structure is used to specify the parent-child relationships between vertical Haar wavelet coefficients. In the numerical example considered next in Section 6, the three quadtrees are modeled as mutually independent, although it may be possible (and desirable) to introduce dependencies among them. The quadtree upward-downward algorithm is essentially the same as the binary tree upward-downward algorithm, except that each parent has four children instead of two.

# 6 Applications to Image Analysis

## *6.1 Image Denoising*

Suppose that we observe an image $\boldsymbol{\mu}$ with additive Gaussian white noise:

$$\mathbf{x} = \boldsymbol{\mu} + \mathbf{w}, \tag{26}$$

where $\boldsymbol{\mu}$ is an array of image intensities and $\mathbf{w}$ is an array of independent realizations of a zero-mean Gaussian random variable. The goal of the denoising problem is to estimate $\boldsymbol{\mu}$ given the data $\mathbf{x}$. If we specify a prior for multiscale parameters (Haar wavelet coefficients) of $\boldsymbol{\mu}$ and formulate the image estimation problem under squared error or 0/1 loss, it can be shown that an optimal image estimate $\widehat{\boldsymbol{\mu}}$ is obtained from the posterior mean or MAP estimates of the multiscale parameters, respectively, as shown by Figueiredo and Nowak (1998). Later in this section we will consider a numerical example of this problem and compare the posterior mean estimates obtained from an independent parameter prior to that obtained with an MHMM. Examples of 2-d Poisson intensity estimation can be found in the work of Timmermann and Nowak (1999).

## *6.2 Image Edge Detection*

Multiscale methods of edge detection are usually based on finding the local wavelet coefficient maxima, as developed by Mallat (1998). The multiscale models considered in this chapter offer an alternative Bayesian approach to edge detection. Again, let us consider the Gaussian observation model above (26), and recall the simple two-state mixture model. State '0' is associated with a low-variance Gaussian component, indicative of a region of smooth behavior, while state '1', corresponding to a high-variance Gaussian, is a cue for the existence of an edge. This interpretation of the state variables suggests testing for the presence of an edge using the Bayes factors of the states. That is, decide an edge is present at scale $j$, orientation $i = 1, 2$, or 3 (corresponding to horizontal, vertical, and diagonal, respectively), and position $k, l$ if

$$BF(\mathbf{x}) \;=\; \frac{p(s^i_{j,k,l} = 1 \,|\, \mathbf{x})\, p(s^i_{j,k,l} = 0)}{p(s^i_{j,k,l} = 0 \,|\, \mathbf{x})\, p(s^i_{j,k,l} = 1)} \;>\; 1, \tag{27}$$

where $p(s^i_{j,k,l} = 1)$ and $p(s^i_{j,k,l} = 0)$ are the prior probabilities of the state. As we will see in the numerical example considered next, the ability of the MHMM to propagate state information from coarse-to-fine scales results in significantly better edge detection performance compared to that resulting from the independent state model.

### *6.3 Numerical Example*

Here we consider a simple numerical illustration of the ideas presented in this paper. Figure 4 (a) depicts a close-up of the test image shown in Figure 3. Figure 4 (b) shows the same image with additive Gaussian white noise of standard deviation $\sigma = 25$. A two-state MHMM was specified for this problem with the following parameter settings:

$$\begin{aligned}
\tau_0 &= 0, \\
\tau_1 &= 250, \\
\rho_{0,0}(0) &= 0.9, \\
\rho_{j,k}(0|0) &= 0.9, \quad k = 0, \ldots, 2^j - 1, \; j = 1, \ldots, J-1, \\
\rho_{j,k}(0|1) &= 0.25, \quad k = 0, \ldots, 2^j - 1, \; j = 1, \ldots, J-1.
\end{aligned}$$

These parameters were selected with the noise variance and basic parent-child dependencies in mind. One can, however, plug-in maximum likelihood or moment-based estimates of these (hyper) parameters, for a fully automatic procedure. For example, an expectation-maximization algorithm based on the upward-downward algorithm is derived by Crouse et al. (1998) to obtain maximum likelihood estimates of the mixture variances and the transition probabilities. Here, for comparative purposes, we also consider an analogous independent multiscale parameter model (all states, and hence parameters, mutually independent), whose prior state probabilities are the same as the marginal state probabilities of the MHMM specified above.

Posterior mean estimates of the image are obtained by computing inverse Haar wavelet transform of the the posterior means of the Haar wavelet coefficients obtained from the two models.[6] Figure 4 (c) shows the estimate

[6] As is usual in wavelet denoising, the "raw" scaling coefficients obtained directly from the noisy image were used in the inverse transform.

based on the independent coefficient model (average squared pixel error = 171.68) and Figure 4 (d) shows the estimate based on the MHMM (average squared pixel error = 163.79). In comparison, the average square pixel error is 625 in the noisy image shown in Figure 4 (b). The MHMM based estimate also appears to be subjectively better than that obtained from the independent parameter prior; there are less residual noise spikes in the smooth background region of the image, and the edges appear slightly sharper.

Image edges were detected from the noisy data by computing the Bayes factors of the states at finest scale. The "edge maps" obtained from the independent coefficient model are shown in Figure 4 (e) and those from the MHMM are shown in Figure 4 (f). To visualize the complete set of edges, each pixel in these two edge maps was set to "black" if one or more of the three (corresponding to the three possible orientations) Bayes factors at the finest scale ($j = J - 1$) and at the corresponding spatial position tested positive according to (27), and was set to "white" otherwise. The edge map resulting from the MHMM is vastly superior than that resulting from the independent parameter model. There are far fewer false edge detections in the background of the MHMM edge map and the continuity of the true edges is captured to a higher degree.

# 7 Conclusions

The MHMM framework described in this chapter appears to be a promising new approach to Bayesian image analysis. The MHMM captures the key inter-scale dependencies present in natural imagery, and, unlike classical MRF based methods that typically require computationally intensive stochastic optimization, the MHMM allows for simple inference algorithms based on probability propagation. Hence, the computational complexity of the MHMM framework is $O(N)$, where $N$ is the number of data. A common framework for MHMMs, capable of analyzing Gaussian and Poisson processes, was presented; applications to Bayesian image denoising and edge detection were examined.

There are three important features being exploited in the Gaussian and Poisson observations models:

**1.** parametric parent-child conditional probabilities, (5) and (6);

**2.** likelihood factorization, (7);

**3.** conjugate priors for multiscale parameters, (10) and (11).

Without these features, multiscale analysis and modeling would be significantly more complicated. In particular, the likelihood factorization allows us to postulate an alternative multiscale observation (or data generation) model; a single (coarse-scale) Poisson count refined by independent binomial splits in the Poisson case, and in the Gaussian case we have independent Gaussian distributed Haar wavelet coefficients. Again, in the Gaussian case, a similar factorization (multiscale observation model) exists for orthogonal wavelet transforms in general, due to the orthogonality of the transformation; see Crouse et al. (1998) for details. In essence, it is the multiscale observation model that enables the graphical interpretation of the problem, and it is doubtful that a simple inference algorithm exists without such a factorization. Hence, it is natural to seek out other observation models that have a similar factorization property. Some other cases are investigated by Kolaczyk (1999), but it appears that the Gaussian and Poisson cases are quite exceptional and that other common models may not be amenable to the MHMM framework.

The connection between MHMMs and $1/f$ processes also deserves mention. It has been shown in the work of Nowak (1998) and Timmermann and Nowak (1999) that, in certain cases, the independent parameter priors discussed in Section 3 for Gaussian and Poisson models both have $1/f$ spectral characteristics. This is very relevant to image analysis since there is convincing empirical evidence that natural images have similar spectral characteristics; see the comprehensive study by van der Schaaf and van Hateren (1996). MHMM priors can also display this behavior. For example, in the Gaussian case, because the Markov structure of the MHMM is imposed on the variances underlying the zero-mean Gaussian mixtures instead of directly on the Haar wavelet coefficients, the coefficients are uncorrelated (but not independent), and hence the MHMM has the same second order correlation structure as the independent coefficient model. Of course, the higher order correlation behavior is (desirably) different for MHMMs. One can argue that the higher order structure is especially relevant in image analysis. For instance, perhaps more important than the number of edges in an image (roughly speaking, measured by the decay of the second order spectrum) is the arrangement and structure of edges (reflected in higher order correlations). These observations suggest avenues for future investigations of the properties and applications of the MHMM framework.

## Acknowledgments

The author thanks Richard Baraniuk and Mário Figueiredo for carefully reading an early version of this manuscript and making several suggestions that helped to improve the presentation. Special thanks go to Eric Kolaczyk for many helpful discussions and suggestions. The author also thanks Matthew Crouse and Klaus Timmermann for their valuable contributions to many aspects of this work.

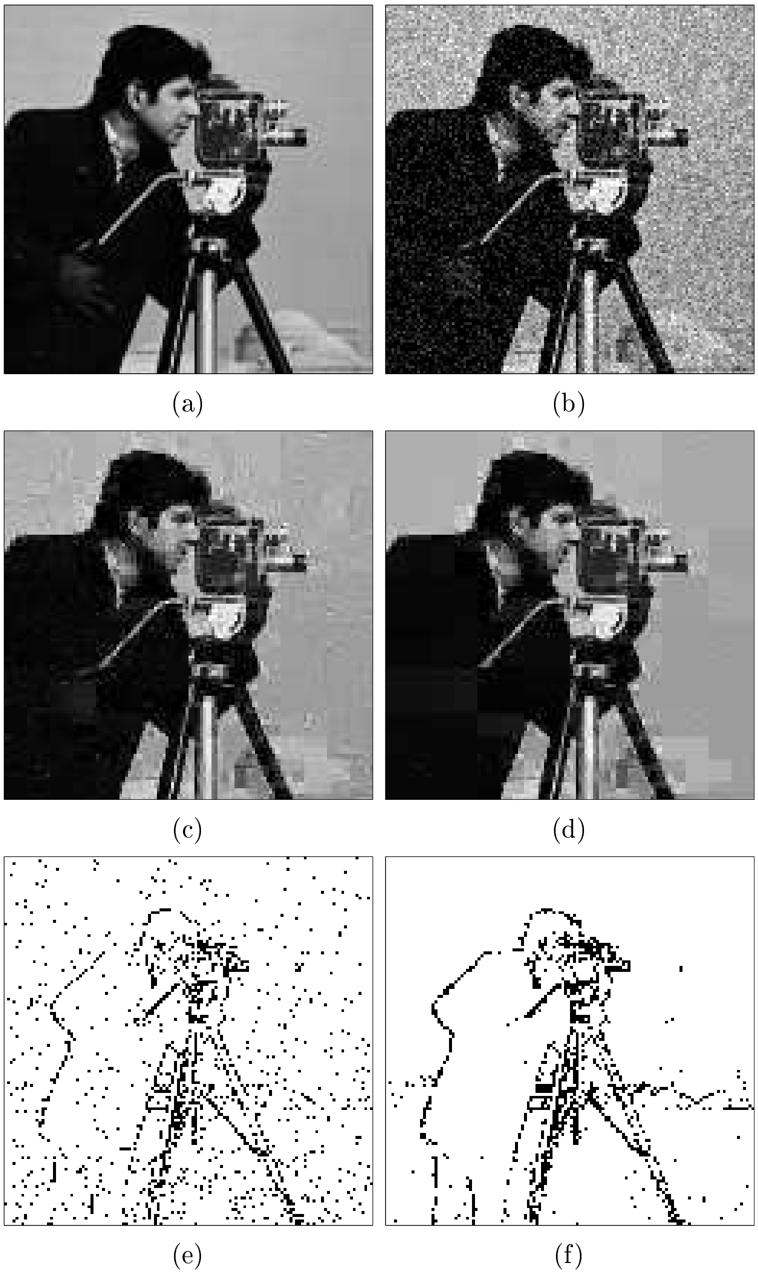



FIGURE 4. *Bayesian image denoising and edge detection. (a) Close-up of original test image (full image shown in Figure 3). (b) Close-up of noisy image. (c) Close-up of posterior mean estimate based on independent multiscale parameter prior model. (d) Close-up of posterior mean estimate based on MHMM. (e) Edges detected from Bayes factors resulting from independent wavelet coefficient prior. (f) Edges detected from Bayes factors resulting from MHMM.*

# Bibliography

Abramovich, F., Sapatinas, T., and Silverman, B. W. (1998). Wavelet thresholding via a Bayesian approach. *J. Roy. Statist. Soc. Ser. B., 60, 725-749*, 60:725–749.

Adelson, E. and Burt, P. (1981). Image data compression with the Laplacian pyramid. In *Proc. Patt. Recog. Info. Proc. Conf.*, pages 218–223, Dallas, TX.

Bouman, C. and Shapiro, M. (1994). A multiscale random field model for Bayesian image segmentation. *IEEE Trans. Image Proc.*, 3(2):162–177.

Castleman, K. (1996). *Digital Image Processing*. Prentice-Hall, Englewood Cliffs, New Jersey.

Charbonnier, P., Blanc-Fèraud, L., and Barlaud, M. (1992). Noisy image restoration using multiresolution Markov random fields. *Journal of Visual Communication and Image Representation*, 3(4):338–346.

Chellappa, R. and Jain, A. (1993). *Markov Random Fields: Theory and Applications*. Academic Press, San Diego, CA.

Chipman, H., Kolaczyk, E., and McCulloch, R. (1997). Adaptive Bayesian wavelet shrinkage. *J. Amer. Statist. Assoc.*, 92:1413–1421.

Cooper, G. F. (1990). The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence*, 42:393–405.

Cross, G. and Jain, A. (1983). Markov random field texture models. *IEEE Trans. Patt. Anal. Mach. Intell.*, 5:25–39.

Crouse, M., Nowak, R., and Baraniuk, R. (1996). Hidden Markov models for wavelet-based signal processing. In *Proc. Thirtieth Asilomar Conf. Signals, Systems, and Comp.,* Pacific Grove, CA, pages 1029–1034. IEEE Computer Society Press.

Crouse, M., Nowak, R., and Baraniuk, R. (1998). Wavelet-based statistical signal processing using hidden Markov models. *IEEE Trans. Signal Processing*, 46:886–902.

Donoho, D. and Johnstone, I. (1994). Ideal adaptation via wavelet shrinkage. *Biometrika*, 81:425–455.

Field, D. (1993). Scale-invariance and self-similar 'wavelet' transforms: an analysis of natural scenes and mammalian visual systems. in *Wavelets, Fractals, and Fourier Transforms*, Claredon Press, Oxford:151–193.

Figueiredo, M. and Nowak, R. (1998). Bayesian wavelet-based signal estimation using non-informative priors. In *Proc. Thirty-Second Asilomar Conf. Signals, Systems, and Comp.*, Pacific Grove, CA. IEEE Computer Society Press.

Flandrin, P. (1992). Wavelet analysis and synthesis of fractional Brownian motion. *IEEE Trans. Inform. Theory*, 38(2):910–916.

Frey, B. (1998). *Graphical Models for Machine Learning and Digital Communication*. MIT Press, Cambridge, Massachusetts.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. *IEEE Trans. Patt. Anal. Mach. Intell.*, 6(6):712–741.

Gidas, B. (1989). A renormalization group approach to image processing problems. *IEEE Trans. Patt. Anal. Mach. Intell.*, 11(2):164–180.

Kolaczyk, E. (1998). Bayesian multi-scale models for Poisson processes. *Technical Report 468, Dept. of Statistics, University of Chicago.*

Kolaczyk, E. (1999). Some observations on the tractability of certain multi-scale models. In *Bayesian Inference in Wavelet Based Models*. Springer-Verlag. Editors B. Vidakovic and P. Müller.

Luettgen, M., Karl, W., Willsky, A., and Tenney, R. (1993). Multiscale representations of Markov random fields. *IEEE Trans. Signal Proc.*, 41(12):3377–3395.

Malfait, M. and Roose, D. (1997). Wavelet based image denoising using Markov random field *a priori* model. *IEEE Transactions on Image Processing*, 6(4):549–565.

Mallat, S. (1998). *A Wavelet Tour of Signal Processing*. Academic Press, San Diego, CA.

Nowak, R. (no. 83, Bryce Canyon, UT, 1998). Shift invariant wavelet-based statistical models and $1/f$ processes. *Proc. IEEE Digital Signal Processing Workshop.*

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Francisco, CA.

Pérez, P. and Heitz, F. (1996). Restriction of a Markov random field on a graph and multiresolution statistical image modeling. *IEEE Trans. Info. Theory*, 42(1):180–190.

Robert, C. (1994). *The Bayesian Choice: A Decision Theoretic Motivation.* Springer-Verlag, New York.

Shapiro, J. (1993). Embedded image coding using zerotrees of wavelet coefficients. *IEEE Trans. Signal Proc.*, 41(12):3445–3462.

Simoncelli, E. (1997). Statistical models for images: Compression, restoration and synthesis. In *Proc. Thirty-First Asilomar Conf. Signals, Systems, and Comp.,* Pacific Grove, CA, pages 673–678. IEEE Computer Society Press.

Timmermann, K. and Nowak, R. (1997). Multiscale Bayesian estimation of Poisson intensities. In *Proc. Thirty-First Asilomar Conf. Signals, Systems, and Comp.,*, pages 85–90. IEEE Computer Society Press.

Timmermann, K. and Nowak, R. (April, 1999). Multiscale modeling and estimation of Poisson processes with application to photon-limited imaging. *IEEE Transactions on Information Theory*, 45(3):846–862.

van der Schaaf, A. and van Hateren, J. (1996). Modelling the power spectra of natural images. *Vision Research*, 36(17):2759–2770.

Vidakovic, B. (ISDS, Duke University, 1998). Honest modeling in the wavelet domain. *Discussion Paper XX-98.*

Wornell, G. (1996). *Signal Processing with Fractals. A Wavelet-Based Approach.* Prentice Hall, Englewood Cliffs, New Jersey.