AALBORG UNIVERSITY

Gaussian-log-Gaussian wavelet trees, frequentist and Bayesian inference, and statistical signal processing applications

by

Jesper Møller and Robert Dahl Jacobsen

R-2014-04 April 2014

Department of Mathematical Sciences
Aalborg University
Fredrik Bajers Vej 7 G ▪ DK - 9220 Aalborg Øst ▪ Denmark
Phone: +45 99 40 99 40 ▪ Telefax: +45 99 40 35 48
URL: http://www.math.aau.dk

ISSN 1399–2503 ▪ On-line version ISSN 1601–7811

Gaussian-log-Gaussian wavelet trees, frequentist and Bayesian inference, and statistical signal processing applications

April 24, 2014

Jesper Møller and Robert Dahl Jacobsen
Department of Mathematical Sciences, Aalborg University

Abstract

We introduce a promising alternative to the usual hidden Markov tree model for Gaussian wavelet coefficients, where their variances are specified by the hidden states and take values in a finite set. In our new model, the hidden states have a similar dependence structure but they are jointly Gaussian, and the wavelet coefficients have log-variances equal to the hidden states. We argue why this provides a flexible model where frequentist and Bayesian inference procedures become tractable for estimation of parameters and hidden states. Our methodology is illustrated for denoising and edge detection problems in two-dimensional images.

Key words: conditional auto-regression; EM algorithm; hidden Markov tree; integrated nested Laplace approximations.

1 Introduction

To model statistical dependencies and non-Gaussianity of wavelet coefficients in signal processing, Crouse, Nowak & Baraniuk (1998) introduced a model where the wavelet coefficients conditional on a hidden Markov tree are independent Gaussian variables, with the hidden states taking values in a finite set (in applications, each hidden variable is often binary) and used for determining the variances of the wavelet coefficient. We refer to this as the *Gaussian-finite-mixture (GFM) wavelet tree model* or just the GFM model. The GFM model and a clever implementation of the EM-algorithm have been widely used in connection to e.g. image segmentation, signal classification, denoising, and image document categorization, see e.g. Crouse et al. (1998), Po & Do (2006), and Choi & Baraniuk (2001). According to Crouse et al. (1998), the "three standard problems" (page 892) are training (i.e. parameter estimation), likelihood determination

(i.e. determining the likelihood given an observed set of wavelet coefficients), and state estimation (i.e. estimation of the hidden states); they focus on the two first problems, but mention that state estimation "is useful for problems such as segmentation" (page 893).

In the present paper, we propose an alternative model—called the *Gaussian-log-Gaussian (GLG) wavelet tree model* or just the GLG model—where the hidden states are jointly Gaussian and described by a similar dependence structure as in the GFM model, and where the wavelet coefficients conditional on the hidden states are still independent Gaussian variables but the log-variance for each wavelet coefficient is given by the corresponding hidden state. In comparison with the GFM model, in many cases the GLG model provides a flexible model and a better fit for wavelet coefficients, it is easy to handle for parameter estimation in a frequetist setting as well as in a Bayesian setting, where state estimation is also possible in the latter case, and it works well for denoising and edge detection problems.

The paper is organized as follows. Section 2 provides further details of the GFM and GLG models. Section 3 studies the moment structure of parametric GLG models and exploits the tractability of the lower-dimensional distributions of the GLG model to develop composite likelihoods so that the EM-algorithm becomes feasible for parameter estimation. Section 4 concerns fast Bayesian procedures for marginal posterior estimation of parameters and hidden states in the GLG model, where we use integrated nested Laplace approximations (Rue, Martino & Chopin 2009). Section 5 demonstrates how our methods in Sections 3 and 4 apply for denoising and edge detection in two-dimensional images. Section 6 contains concluding remarks. Technical details are deferred to Appendix A-D. Matlab and R (R Core Team 2013) codes for our statistical inference procedures are available at `http://www.mathworks.com/matlabcentral/fileexchange/43417`.

2 Wavelet tree models

For both the GFM and the GLG model, we consider wavelet coefficients $\mathbf{w} = (w_1, \ldots, w_n)$, where the units $1, \ldots, n$ represent an abstract single index system. The units are identified with the nodes of a tree with root 1 (the coarsest level of the wavelet transform) and edges corresponding to the parent-child relations of the wavelet coefficients at the coarsest to the finest level, see Figure 1. Conditional on hidden states $\mathbf{s} = (s_1, \ldots, s_n)$, the wavelet coefficients are independent Gaussian distributed, where each w_i depends only on s_i (in Section 5.2 we modify the GFM and GLG models and consider wavelet coefficients with noise). For simplicity and since it is frequently the case in applications, we assume that each conditional mean $\mathrm{E}[w_i|s_i] = 0$ is centered. However, for the two models, the conditional variance $\mathrm{Var}[w_i|s_i]$ depends on i and s_i in different ways; the details are given in Sections 2.1 and 2.2.

The conditional independence structure for the hidden states is the same for the two models and given by the tree structure, i.e. $\mathbf{s}$ is viewed as a directed graphical model (see e.g. Lauritzen (1996)): For $i = 1, \ldots, n$, denote $c(i) \subset \{1, \ldots, n\}$ the children of i, where each child $j \in c(i)$ is at one level lower than i (see Figure 1); if i is at the finest wavelet

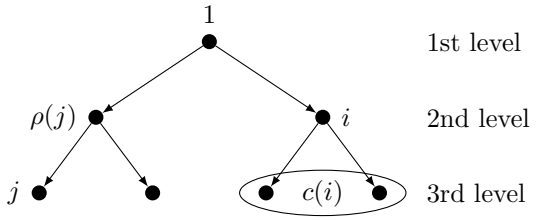

Figure 1: Illustration of a binary tree structure corresponding to a one-dimensional signal with $l = 3$ levels of wavelet coefficients. Node j has one parent $\rho(j)$ and node i has two children $c(i)$.

level, i has no children ($c(i) = \emptyset$); else $c(i) \neq \emptyset$. Typically in applications, if i is not at the finest wavelet level, i has 2^d children, where d is the dimension of the signal/image. Now, the joint density for hidden states and wavelet coefficients factorizes as

$$p(\mathbf{s}, \mathbf{w}) = p_0(s_1) \prod_{i=1}^{n} \Big[q_i(w_i|s_i) \prod_{j \in c(i)} p_i(s_j|s_i) \Big] \tag{1}$$

where $p_0(\cdot)$, $p_i(\cdot|\cdot)$, and $q_i(\cdot|\cdot)$ are either probability mass functions or probability density functions, with the true nature being obvious from the context, and where we set $\prod_{j \in c(i)} p_i(s_j|s_i) = 1$ if $c(i) = \emptyset$.

We refer to (1) as a *wavelet tree model*. We shall consider parametric models, using θ as generic notation for the unknown parameters, and to stress the dependence of θ we write e.g. $p(\mathbf{s}, \mathbf{w}|\theta)$ for the density $p(\mathbf{s}, \mathbf{w})$. When we later discuss parameter estimation (Sections 3 and 4), we consider k independent pairs $(\mathbf{s}^{(1)}, \mathbf{w}^{(1)}), \ldots, (\mathbf{s}^{(k)}, \mathbf{w}^{(k)})$ with density $p(\mathbf{s}, \mathbf{w}|\theta)$, where we suppose only the wavelet coefficients $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(k)}$ are observed.

2.1 The GFM model

For the GFM model, it is assumed in Crouse et al. (1998) that

- the state space of each s_i is a finite set $\{1, \ldots, m\}$ (often $m = 2$),
- $q_i(\cdot|s_i)$ is a Gaussian density where both the mean μ_{i,s_i} and the variance σ^2_{i,s_i} depend on the index i and the argument s_i.

Crouse et al. (1998) remark that instead of a single Gaussian distribution, the m-state Gaussian mixture distribution for each wavelet coefficient is needed because of "the compressing property... resulting in a large number of small coefficients and a small number of large coefficients" (page 887); and the conditional dependence structure is used to "characterize the key dependencies between the wavelet coefficients" (page 887), i.e. it "matches both the clustering and persistence properties of the wavelet transform"

(page 891) so that "if one coefficient is in a high-variance (low-variance) state, then its neighbor is very likely to also be in a a high-variance (low-variance) state" (page 891).

The parameters of the GFM model are

- the variance σ^2_{i,s_i} of $q_i(\cdot|s_i) \sim N(0, \sigma^2_{i,s_i})$, $s_i = 1, \ldots, m$, $i = 1, \ldots, n$,
- the initial probabilities $p_0(\cdot)$ and the unknown transition probabilities $p_i(\cdot|\cdot)$ of the hidden state variables with $c(i) \neq \emptyset$ and $i = 0, \ldots, n$ (setting $c(0) = \{1\}$).

Usually the variances and transition probabilities are assumed only to depend on the level of the nodes. Then, denoting l the number of levels in the tree, the number of parameters is

$$r_{\text{GFM}} = ml + m - 1 + m(m-1)(l-1). \tag{2}$$

2.2 The GLG model

In the GLG model,

- each wavelet coefficient w_i conditional on s_i is zero-mean Gaussian with variance $\exp(s_i)$, i.e. $q_i(\cdot|s_i) = q(\cdot|s_i)$ does not depend on i, and

$$q(w_i|s_i) \sim N(0, \exp(s_i)); \tag{3}$$

- the hidden states are jointly Gaussian, i.e. $p_0(s_1) = p(s_1|\mu_0, \sigma_0^2)$ and $p_i(s_j|s_i) = p(s_j|s_i, \alpha_i, \beta_i, \kappa_i^2)$ for $j \in c(i)$, where

$$p(s_1|\mu_0, \sigma_0^2) \sim N(\mu_0, \sigma_0^2), \tag{4}$$
$$p(s_j|s_i, \alpha_i, \beta_i, \kappa_i^2) \sim N(\alpha_i + \beta_i s_i, \kappa_i^2), \tag{5}$$

  where μ_0, α_i, and β_i are real parameters and σ_0 and κ_i are positive parameters.

The density (3) is completely determined by the variance $\exp(s_i)$, and it appears to be a more flexible model for wavelet coefficients than the m-state Gaussian mixture model used in Crouse et al. (1998): In the GFM model, wavelet coefficients from all trees and associated to the same parent (or to the root) are sharing the same set of m possible variances, while in the GLG model, each wavelet coefficient $w_i^{(t)}$ for each tree t is having its 'own' log-Gaussian hidden state $s_i^{(t)}$.

In Section 3.1.2 and further on we assume—as in the GFM model—'tying within levels', that is the parameters on each level are equal (detailed later in (15)). Then the number of parameters in the GLG model for a tree with l levels is $r_{\text{GLG}} = 3l - 1$. In comparison the GFM model with $m = 2$ is specified by $r_{\text{GFM}} = 4l - 1$ parameters, while the difference will be even larger as m grows, cf. (2).

3 Parameter estimation using composite likelihoods and the EM-algorithm

Section 3.1 describes the first and higher order moment structure of the hidden states and the wavelets coefficients under the GLG model. In particular we clarify the meaning of tying within levels, which is assumed when we in Section 3.2 discuss parameter estimation using composite likelihoods and the EM-algorithm.

3.1 Mean and variance-covariance structure

3.1.1 Full parametrization

This section considers a full parametrization of the GLG model (4), i.e. when

$$\mu_0 \in \mathbb{R}, \quad \sigma_0 > 0, \quad (\alpha_i, \beta_i, \kappa_i) \in \mathbb{R} \times \mathbb{R} \times (0, \infty) \tag{6}$$

for $i \geq 0$ and $c(i) \neq \emptyset$. For each node $j \neq 1$ in the tree structure, let $\rho(j)$ denote the parent to j (see Figure 1), and set $\rho(1) = 0$. By (4) and (5), each hidden state s_j is Gaussian distributed with a mean and variance which are determined by the means and the variances of its ancestors:

$$p_j(s_j) = p(s_j|\mu_{\rho(j)}, \sigma^2_{\rho(j)}) \sim N(\mu_{\rho(j)}, \sigma^2_{\rho(j)}) \tag{7}$$

for $j = 1, \ldots, n$, where the mean and the variance are determined recursively from the coarsest level to the second finest level by

$$\mu_i = \alpha_i + \beta_i \mu_{\rho(i)}, \quad \sigma_i^2 = \kappa_i^2 + \beta_i^2 \sigma^2_{\rho(i)} \tag{8}$$

for $i \geq 1$ and $c(i) \neq \emptyset$. Conversely, the GLG model is parametrized by $(\mu_0, \sigma_0) \in (-\infty, \infty) \times (0, \infty)$ and $(\mu_i, \sigma_i, \beta_i) \in (-\infty, \infty) \times (0, \infty) \times (-\infty, \infty)$ for all $i \geq 1$ with $c(i) \neq \emptyset$, since

$$\alpha_i = \mu_i - \beta_i \mu_{\rho(i)} \quad \text{and} \quad \kappa_i^2 = \sigma_i^2 - \beta_i^2 \sigma^2_{\rho(i)} \tag{9}$$

whenever $i \geq 1$ and $c(i) \neq \emptyset$.

Set $\kappa_0 = \sigma_0$ and denote $\sigma_{i,j} = \mathrm{Cov}(s_i, s_j)$, the covariance of s_i and s_j. Note that $\sigma_{i,i} = \sigma^2_{\rho(i)}$; a general expression for $\sigma_{i,j}$ is given by (34) in Appendix A. In particular,

$$\sigma_{h,j} = \kappa^2_{\rho(h)} \beta_h \quad \text{if } j \in c(h), \quad \sigma_{i,j} = \kappa^2_{\rho(h)} \beta_h^2 \tag{10}$$

if $i, j \in c(h)$ and $i \neq j$.

Moments of the form $\mathrm{E}\left[w_i^a w_j^b\right]$ for $a = 0, 1, \ldots$ and $b = 0, 1, \ldots$ can be derived by conditioning on the hidden states and exploiting well-known moment results for the log-Gaussian distribution, see e.g. (30)-(32) in Appendix A. In particular, letting $c(0) = \{1\}$,

then for any $h \geq 0$ and $j \in c(h)$,

$$\eta_h^{(2)} := \mathrm{E}\big[w_j^2\big] = \exp\big(\mu_h + \sigma_h^2/2\big), \tag{11}$$

$$\eta_h^{(4)} := \mathrm{E}\big[w_j^4\big] = 3\exp\big(2\mu_h + 2\sigma_h^2\big), \tag{12}$$

$$\eta_h^{(2,2)} := \mathrm{E}\big[w_i^2 w_j^2\big] = \exp\Big(2\mu_h + \kappa_{\rho(h)}^2\beta_h^2 + \sigma_h^2\Big) \quad \text{if } i \in c(h) \text{ and } i \neq j, \tag{13}$$

$$\xi_{h,j}^{(2,2)} := \mathrm{E}\big[w_j^2 w_h^2\big] = \exp\big(\mu_h + \mu_{\rho(h)} + \kappa_{\rho(h)}^2\beta_h + \sigma_h^2/2 + \sigma_{\rho(h)}^2/2\big). \tag{14}$$

3.1.2 Tying within levels

For each node i in the tree structure, denote $\ell(i)$ the level of i, i.e. $\ell(i)$ is the number of nodes in the path from the root to i, and let l be the number of levels (see Figure 1). For convenience, define $\ell(0) = 0$. Henceforth we assume tying within levels of the parameters in (4) and (5), that is

$$\alpha_i = \alpha(\ell(i)), \quad \beta_i = \beta(\ell(i)), \quad \kappa_i = \kappa(\ell(i)), \quad 1 \leq \ell(i) < l. \tag{15}$$

Thus the unknown parameters are

$$\mu_0 \in \mathbb{R}, \quad \sigma_0 > 0, \quad (\alpha(1), \ldots, \alpha(l)) \in \mathbb{R}^l,$$
$$(\beta(1), \ldots, \beta(l)) \in \mathbb{R}^l, \quad (\kappa(1), \ldots, \kappa(l)) \in (0, \infty)^l.$$

Note that for $1 \leq \ell(i) < l$, by (8), (15), and induction, (μ_i, σ_i^2) depends on the node i only through its corresponding level $\ell(i)$, i.e.

$$\mu_i = \mu(\ell(i)), \quad \sigma_i^2 = \sigma^2(\ell(i)). \tag{16}$$

Furthermore, for $0 \leq \ell(h) < l$ and $j \in c(h)$, we obtain from (11)-(15) that $(\eta_h^{(2)}, \eta_h^{(4)}, \eta_h^{(2,2)}, \xi_{h,j}^{(2,2)})$ depends on h only through $\ell(h)$, i.e.

$$\eta_h^{(2)} = \eta^{(2)}(\ell(h)), \quad \eta_h^{(4)} = \eta^{(4)}(\ell(h)), \quad \eta_h^{(2,2)} = \eta^{(2,2)}(\ell(h)), \quad \xi_{h,j}^{(2,2)} = \xi^{(2,2)}(\ell(h)). \tag{17}$$

3.2 Parameter estimation

For parameter estimation in the GFM model, Crouse et al. (1998) propose to use the EM-algorithm. Here the main difficulty is the calculation of $p_i(s_i, s_j|\mathbf{w})$ for $j \in c(i)$, the two-dimensional marginal probabilities of any s_i and its child s_j conditional on the wavelet coefficients, where the calculation has to be done for each E-step of the EM-algorithm and each wavelet tree $\mathbf{w} = \mathbf{w}^{(t)}$, $t = 1, \ldots, k$. Crouse et al. (1998) solve this problem using an upward-downward algorithm which is equivalent to the forward-backward algorithm for hidden Markov chains. Durand, Gonçalvès & Guédon (2004) improve on the numerical limitations on this algorithm.

Modifying the upward-downward algorithm in Crouse et al. (1998) to the GLG model is not leading to a computationally feasible algorithm mainly because, for each s_i, we have replaced its finite state space $\{1, \ldots, m\}$ under the GFM model by the real line

under the GLG model, and numerical integration would be repeatably needed at the various (many) steps of the algorithm. As noticed in Appendix C, the Gauss-Hermite quadrature rule provides good approximations with few quadrature nodes when considering trees with no more than two levels. However, since the integrants involved in the transition between levels of the upward-downward algorithm are not sufficiently smooth, we propose instead an EM algorithm for estimating θ based on composite likelihoods for the joint distribution of wavelets corresponding to each parent and its children. These joint distributions are relatively easy to handle. Further details are given in the sequel.

3.2.1 Marginal likelihoods

When defining composite likelihoods in connection to the EM-algorithm in Section 3.2.2, we use the following marginal likelihoods given in terms of the full parametrization (4) and (5).

Combining (3) and (4), we obtain the density of (s_1, w_1),

$$p(s_1, w_1|\mu_0, \sigma_0^2) = \frac{\exp\left(-\frac{1}{2}\left[\frac{w_1^2}{\exp(s_1)} + s_1 + \frac{(s_1-\mu_0)^2}{\sigma_0^2}\right]\right)}{2\pi\sigma_0} \tag{18}$$

and hence the marginal density of the root wavelet,

$$q(w_1|\mu_0, \sigma_0^2) = \int_{-\infty}^{\infty} p(s_1, w_1|\mu_0, \sigma_0^2)\,\mathrm{d}s_1. \tag{19}$$

The marginal log-likelihood based on the root wavelet vector $\bar{\mathbf{w}}_1 = (w_1^{(1)}, \ldots, w_1^{(k)})$ for the k trees is given by

$$l_0(\mu_0, \sigma_0^2|\bar{\mathbf{w}}_1) = \sum_{t=1}^{k} \log q(w_1^{(t)}|\mu_0, \sigma_0^2). \tag{20}$$

Consider any $i \in \{1, \ldots, n\}$ with $c(i) \neq \emptyset$. Denote $\mathbf{w}_{i,c(i)}$ the vector consisting of w_i and all w_j with $j \in c(i)$, and $\mathbf{s}_{i,c(i)}$ the vector consisting of s_i and all s_j with $j \in c(i)$. Using (3)-(5) and (7), we obtain the density of $(\mathbf{s}_{i,c(i)}, \mathbf{w}_{i,c(i)})$,

$$p(\mathbf{s}_{i,c(i)}, \mathbf{w}_{i,c(i)}|\mu_{\rho(i)}, \sigma_{\rho(i)}^2, \alpha_i, \beta_i, \kappa_i^2) = \frac{\exp\left(-\frac{1}{2}\left\{\left[\frac{w_i^2}{\exp(s_i)} + s_i + \frac{(s_i-\mu_{\rho(i)})^2}{\sigma_{\rho(i)}^2}\right] + \sum_{j\in c(i)}\left[\frac{w_j^2}{\exp(s_j)} + s_j + \frac{(s_j-\alpha_i-\beta_i s_i)^2}{\kappa_i^2}\right]\right\}\right)}{(2\pi)^{1+|c(i)|}\sigma_{\rho(i)}\kappa_i^{2|c(i)|}} \tag{21}$$

where $|c(i)|$ denotes the number of children to i. Hence the density of $\mathbf{w}_{i,c(i)}$ is given by the integral

$$q(\mathbf{w}_{i,c(i)}|\mu_{\rho(i)}, \sigma_{\rho(i)}^2, \alpha_i, \beta_i, \kappa_i^2) = \int_{-\infty}^{\infty}\prod_{j\in c(i)}\int_{-\infty}^{\infty} p(\mathbf{s}_{i,c(i)}, \mathbf{w}_{i,c(i)}|\mu_{\rho(i)}, \sigma_{\rho(i)}^2, \alpha_i, \beta_i, \kappa_i^2)\,\mathrm{d}s_j\,\mathrm{d}s_i. \tag{22}$$

Finally, denoting $\bar{\mathbf{w}}_{i,c(i)}$ the vector of the ith wavelets $w_i^{(1)}, \ldots, w_i^{(k)}$ and their children $w_j^{(1)}, \ldots, w_j^{(k)}$, $j \in c(i)$, the log-likelihood based on $\bar{\mathbf{w}}_{i,c(i)}$ is

$$l_i(\mu_{\rho(i)}, \sigma^2_{\rho(i)}, \alpha_i, \beta_i, \kappa_i^2 | \bar{\mathbf{w}}_{i,c(i)}) = \sum_{t=1}^{k} \sum_{j \in c(i)} \log q(w_i^{(t)}, w_j^{(t)} | \mu_{\rho(i)}, \sigma^2_{\rho(i)}, \alpha_i, \beta_i, \kappa_i^2). \quad (23)$$

3.2.2 EM-algorithm

This section shows how the EM-algorithm applies on composite likelihoods (Gao & Song 2011) defined from the marginal likelihoods in Section 3.2.1 under the assumption of tying within levels, cf. (15). We proceed from the coarsest to the finest level, where parameter estimates are calculated by the EM-algorithm as detailed in Appendix C

1. Apply the EM-algorithm for the (marginal) log-likelihood (20) to obtain an estimate $(\widehat{\mu}_0, \widehat{\sigma}_0^2)$.

2. For $r = 1, \ldots, l-1$, denoting $\bar{\mathbf{w}}_{(r)}$ the vector of all $\bar{\mathbf{w}}_{i,c(i)}$ with $\ell(i) = r$, the log-composite likelihood given by the sum of the log-likelihoods (23) based on all $\bar{\mathbf{w}}_{i,c(i)}$ with $\ell(i) = r$ is

$$\begin{aligned} &l_{(r)}(\mu(r-1), \sigma^2(r-1), \alpha(r), \beta(r), \kappa(r)^2 | \bar{\mathbf{w}}_{(r)}) \\ &\quad = \sum_{i:\ell(i)=r} l_i(\mu(r-1), \sigma^2(r-1), \alpha(r), \beta(r), \kappa^2(r) | \bar{\mathbf{w}}_{i,c(i)}). \end{aligned}$$

   Now, suppose we have obtained an estimate $(\widehat{\mu}(r-1), \widehat{\sigma}^2(r-1))$. Then we apply the EM-algorithm on $l_{(r)}(\widehat{\mu}(r-1), \widehat{\sigma}^2(r-1), \alpha(r), \beta(r), \kappa(r)^2 | \bar{\mathbf{w}}_{(r)})$ to obtain an estimate $(\widehat{\alpha}(r), \widehat{\beta}(r), \widehat{\kappa}^2(r))$. Thereby, using (8) and (15), an estimate $(\widehat{\mu}(r), \widehat{\sigma}^2(r))$ is also obtained.

These composite likelihoods for our GLG model can be handled mainly because of the conditional independence structure and since the marginal distribution of s_i is Gaussian. In contrast, for the GFM model, unless n is small or all except a few nodes have at most one child, it is not feasible to handle marginal distributions and corresponding composite likelihoods.

For the initial values used in steps 1 and 2, moment-based estimates obtained as described in Appendix B are used. If such an estimate is not meaningful (see Remark 1 in Appendix B), we replace the parameter estimate by a fixed value which makes better sense. Each iteration of step 1 leads to an increase of the marginal log-likelihood (20), so the value returned by the EM algorithm is a local maximum; and each iteration of step 2 leads to an increase of the log-composite likelihood, so the value returned by the EM algorithm is a local maximum when $(\mu(r-1), \sigma^2(r-1)) = (\widehat{\mu}(r-1), \widehat{\sigma}^2(r-1))$ is fixed. In each step, as usual when applying the EM-algorithm, there is no guarantee that the global maximum will be found.

4 Bayesian inference

For the GFM model, Bayesian methods are not feasible: Indeed Crouse et al. (1998) derive recursions for calculating the conditional densities $p(s_i^{(t)}|\mathbf{w}^{(t)},\theta)$ and $p(s_j^{(t)}, s_i^{(t)}|\mathbf{w}^{(t)},\theta)$, $j \in c(i)$, but it is not possible to calculate or satisfactory approximate the marginal posterior densities for (any subparameter of) θ or (any subvector) of $\mathbf{s}^{(t)}$. For instance, if $\bar{\mathbf{w}} = (\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(k)})$ is the vector of wavelets from all the trees, $p(s_i^{(t)}|\bar{\mathbf{w}}) = \int p(s_i^{(t)}|\mathbf{w}^{(t)},\theta)p(\theta|\bar{\mathbf{w}})\,\mathrm{d}\theta$ where under the GFM model we do not know what $p(\theta|\bar{\mathbf{w}})$ is and it seems hopeless to evaluate this high dimensional integral.

Using a Bayesian approach for the GLG model, with a prior imposed on all the r_{GLG} unknown parameters, also leads to a complicated posterior distribution. In principle it could be handled by Markov chain Monte Carlo (MCMC) methods, but "MCMC sampling remains painfully slow from the end user's point of view" (page 322 in Rue et al. (2009)). However, approximate Bayesian methods based on Laplace approximations (Tierney & Kadane 1986, Rue & Martino 2007, Rue et al. 2009) are feasible for GLG submodels when the number of unknown parameters is not high, as in our GLG submodel introduced in Section 4.1. Furthermore, Section 4.2 considers integrated nested Laplace approximations (INLA) to obtain marginal posterior distributions for θ and the hidden states (Rue et al. 2009).

4.1 Conditional auto-regressions

Romberg, Choi & Baraniuk (2001) consider GFM submodels where θ is of dimension nine, and they demonstrate that the submodels are acceptable for denoising images with a high degree of self-similarity, e.g. as found in images of natural scenes. Encouraged by these results and because of the larger flexibility in modelling the variances of single wavelet coefficients in the GLG model, we consider the following GLG submodel.

First, notice that by (4), $\mathbf{s}$ is a Gaussian Markov random field or in fact a conditional auto-regression (CAR; Besag (1974, 1975); Rue & Held (2005, Chapter 1)). The Gaussian distribution of $\mathbf{s}$ is specified by the mean $\mu_{\rho(i)}$ of each s_i and the precision matrix Δ (the inverse of the variance-covariance matrix of $\mathbf{s}$) which has (i,j)th entry

$$\Delta_{i,j} = \begin{cases} \dfrac{1}{\kappa^2_{\rho(i)}} + |c(i)|\dfrac{\beta_i^2}{\kappa_i^2} & \text{if } i = j, c(i) \neq \emptyset, \\ \dfrac{1}{\kappa_i^2} & \text{if } i = j, c(i) = \emptyset, \\ -\dfrac{\beta_{\rho(i)}}{\kappa^2_{\rho(j)}} & \text{if } i = \rho(j), \\ -\dfrac{\beta_{\rho(j)}}{\kappa^2_{\rho(i)}} & \text{if } j = \rho(i), \\ 0 & \text{otherwise,} \end{cases} \tag{24}$$

cf. Appendix A.

Second, consider the *homogeneous GLG model* specified by that

$$\alpha_i = \alpha, \quad \beta_i = \beta, \quad \kappa_i^2 = \kappa^2,$$

whenever $\ell(i) < l$. Then the free parameters are $\theta = (\mu_0, \sigma_0^2, \alpha, \beta, \kappa^2) \in (-\infty, \infty) \times (0, \infty) \times (-\infty, \infty) \times (-\infty, \infty) \times (0, \infty)$. By (8) and (16), we obtain a special mean and variance structure for the hidden states: For level $r = 1, \ldots, l$,

$$\mu(r) = \begin{cases} \alpha \dfrac{\beta^r - 1}{\beta - 1} + \beta^r \mu_0 & \text{if } \beta \neq 1, \\ r\alpha + \mu_0 & \text{if } \beta = 1, \end{cases}$$

and

$$\sigma^2(r) = \begin{cases} \kappa^2 \dfrac{\beta^{2r} - 1}{\beta^2 - 1} + \beta^{2r} \sigma_0^2 & \text{if } \beta \neq 1, \\ r\kappa^2 + \sigma_0^2 & \text{if } \beta = 1. \end{cases}$$

4.2 INLA

Integrated nested Laplace approximations (INLA) is a general framework for performing approximate Bayesian inference in latent Gaussian models where the number of parameters is small (see Rue et al. (2009) and Martins, Simpson, Lindgren & Rue (2013)). Rue et al. (2009) notice that "The main benefit of INLA is computational: where Markov chain Monte Carlo algorithms need hours or days to run, INLA provides more precise estimates in seconds or minutes." This includes estimates of the posterior marginals for θ and for the hidden states.

Parsimonious GLG submodels fit the INLA assumptions. We have implemented the homogeneous GLG model in INLA, where prior specification is largely handled automatically in INLA. Specific calls used in the experiments reported in the sequel can be seen in our released code.

5 Examples of applications

This section compares results using the GLG and GFM models for wavelet coefficients in real images. The GFM model has proven to be useful for modelling different kinds of multiscale transforms (Crouse et al. 1998, Romberg et al. 2001, Po & Do 2006), but our results are only for the standard wavelet transform, where in both the GFM and the GLG model the directions of the wavelet transform are modelled independently. Section 5.1 discusses how well GLG and GFM models describe standard wavelet coefficients, Section 5.2 considers denoising of images, and Section 5.3 concerns edge detection.

5.1 Modelling standard wavelet coefficients in images

For illustrative purposes, in this and the following sections, we use three test images from the USC-SIPI image database available at `http://sipi.usc.edu/database`: 'Lena',

'mandrill', and 'peppers', see Figure 2. These images are 512-by-512 pixels represented as 8 bit grayscale with pixel values in the unit interval, and we have fitted the GFM and GLG models to wavelet transforms using the corresponding EM algorithms. Figure 3 shows four histograms of the wavelet coefficient from a single subband along with the fitted marginal distributions. The figure illustates that no model is fitting better than the other in all cases: For level 1 of the vertical subband of 'Lena' (upper left panel) and for level 2 of the vertical subband of 'mandrill' (upper right panel), the GLG model provides the best fit; for level 3 of the vertical subband of 'mandrill' (lower left panel), the GLG model is too highly peaked at zero and the GFM model provides a better fit; and for level 3 of the diagonal subband of 'mandrill' (lower right panel), the two models fit equally well.

5.2 Denoising

Consider an image corrupted with additive white noise, i.e. we add an independent term to each pixel value from the same zero-mean normal distribution. Recall that when working with orthonormal wavelets, the distribution and the independence properties of the noise are preserved by the wavelet transform, and the procedure for denoising with wavelets works as follows:

$$\text{noisy data} \to \text{noisy wavelets} \to \text{noise-free wavelets} \to \text{noise-free data.}$$

Thus, a wavelet tree $\mathbf{w} = (w_1, \ldots, w_n)$ is also observed with additive white noise:

$$v_i = w_i + \varepsilon_i, \quad i = 1, \ldots, n, \tag{25}$$

where $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$, the ε_i are mutually independent and independent of $(\mathbf{s}, \mathbf{w})$, and we assume that the noise variance σ_ϵ^2 is known. The dependence structure in the tree with noisy observations is illustrated in Figure 4. From this and (25) we obtain

$$p(\mathbf{w}|\mathbf{v}, \mathbf{s}, \theta) = \prod_{i=1}^{n} p(w_i|v_i, s_i, \theta).$$

Below we discuss estimation of $\mathbf{w}$.

In the frequentist setup, we estimate w_i by $\mathrm{E}[w_i|v_i, \theta]$, with θ replaced by its estimate obtained by the appropriate EM-algorithm (see Section 3.2.2 and Crouse et al. (1998)). For an m-state GFM model,

$$\mathrm{E}[w_i|v_i, \theta] = v_i \sum_{j=1}^{m} p(s_i = j|v_i) \frac{\sigma_{i,j}^2}{\sigma_{i,j}^2 + \sigma_\varepsilon^2}$$

(see Crouse et al. (1998)). Under the GLG model, we have

$$\mathrm{E}[w_i|v_i, \theta] = \frac{v_i}{c(v_i|\mu_{\rho(i)}, \sigma_{\rho(i)}^2)} \int \frac{\exp(s_i)}{(\exp(s_i) + \sigma_\varepsilon^2)^{3/2}} \exp\left(-\frac{1}{2} \left[\frac{v_i^2}{\exp(s_i) + \sigma_\varepsilon^2} + \frac{(s_i - \mu_{\rho(i)})^2}{\sigma_{\rho(i)}^2} \right] \right) \mathrm{d}s_i \tag{26}$$

Figure 2: The three test images: 'Lena', 'mandrill', and 'peppers'.

where we use the Gauss-Hermite quadrature rule for approximating the integral. Equation (26) is derived in Appendix D.

In the Bayesian setup for the homogeneous GLG model, we work with the posterior distribution $p(w_i|v_i)$ from which we can calculate various point estimates. We have

$$p(w_i|v_i) = \int p(w_i|v_i, s_i)p(s_i|v_i)\mathrm{d}s_i, \tag{27}$$
$$\mathrm{E}(w_i|v_i) = \int \mathrm{E}(w_i|v_i, s_i)p(s_i|v_i)\mathrm{d}s_i,$$

where $p(s_i|v_i)$ is calculated in INLA. Since $p(w_i|s_i) \sim N(0, \exp(s_i))$ and $p(v_i|w_i) \sim$

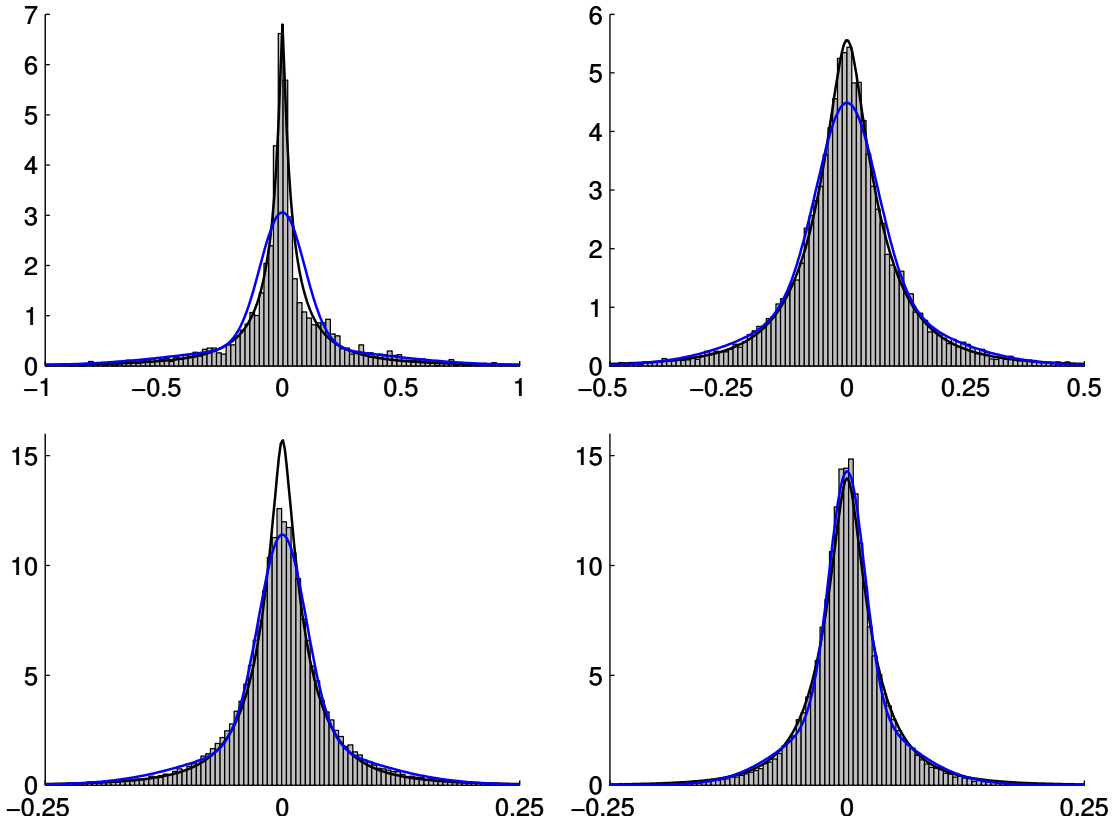

Figure 3: Histograms of wavelet coefficients from one scale of the 3 level wavelet transform with a Daubechies 4 wavelet. The probability density functions of the fitted GLG model (solid line) and the fitted GFM model (gray line) are shown. Upper left panel: Level 1 of the vertical subband of 'Lena'. Upper right panel: Level 2 of the vertical subband of 'mandrill'. Lower left panel: Level 3 of the vertical subband of 'mandrill'. Lower right panel: Level 3 of the diagonal subband of 'mandrill'.

$N(w_i, \sigma_\epsilon^2)$, we obtain

$$p(w_i|v_i, s_i) \propto p(w_i|s_i)p(v_i|w_i) \sim N\left(\frac{v_i \exp(s_i)}{\sigma_\epsilon^2 + \exp(s_i)}, \frac{\sigma_\epsilon^2 \exp(s_i)}{\sigma_\epsilon^2 + \exp(s_i)}\right)$$

and

$$\mathrm{E}(w_i|v_i) = v_i \int \frac{\exp(s_i)}{\sigma_\epsilon^2 + \exp(s_i)} p(s_i|v_i)\mathrm{d}s_i.$$

We apply the two denoising schemes with a three level wavelet transform using the Daubechies 4 filter to noisy versions of the three test images in Figure 2. To estimate the performance of a denoising scheme, we calculate the peak signal-to-noise ratio (PSNR)

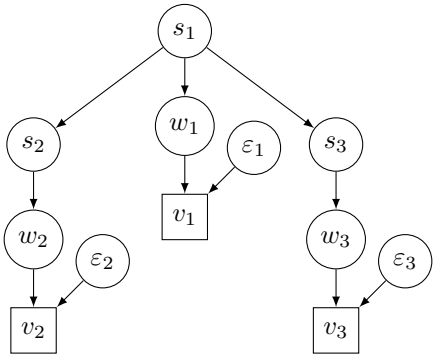

Figure 4: Graphical model of a binary tree with two levels and noisy observations. The rectangular nodes are observed variables and the round nodes are unobserved variables.

in decibels between a test image I and a noisy or cleaned image J. For images of size $N \times N$, the PSNR in decibels is defined as

$$\text{PSNR} = 20 \log_{10} \frac{N(\max\{I(x)\} - \min\{I(x)\})}{\|I - J\|}$$

where the maximum and the minimum are over all pixels x and $\|\cdot\|$ is the Frobenius norm. Table 1 shows for the test images and different noise levels σ_ε, the PSNR between each test image and its noisy or denoised version: For the frequentist results, the images denoised using the GLG model have PSNRs that are consistently higher than those denoised using the GFM model. The Bayesian results yields the lowest PSNR values, but they are also based on a more parsimonious model.

An example of the visual appearance of denoising using frequentist means is seen in Figure 5; again the GLG model performs best, where details around e.g. the stem of the center pepper are more crisp. The median (the 50% quantile) of the posterior distribution is only one possible point estimate of the posterior distribution. However, using other quantiles or the posterior mean are not providing better results, see Figure 6.

5.3 Edge detection

Edge detection in an image is performed by labelling each pixel as being either an edge or a non-edge. Turning to the wavelet transform for this task has the advantage that wavelet coefficients are large near edges and small in the homogeneous parts of an image; the difficulty lies in quantifying "large" and "small". Another advantage is that a multiresolution analysis allows us to search for edges that are present at only selected scales of the image, thereby ignoring edges that are either too coarse or too fine. In this section, for each tree $t = 1, \ldots, k$, we focus on how to label the wavelet coefficient $w_i^{(t)}$ by an indicator variable $f_i^{(t)}$, where $f_i^{(t)} = 1$ means $w_i^{(t)}$ is "large", and $f_i^{(t)} = 0$ means $w_i^{(t)}$ is "small". Labelling of wavelet coefficients using the GFM model is introduced in

Table 1: For the three test images and three noise levels, peak signal-to-noise ratios in dB between the image and its noisy version ot its denoised version obtained using either the GFM model and the EM-algorithm, the GLG model and the EM-algorithm, or the homogeneous GLG model and INLA. In the latter case, the PSNR is calculated using the median of the posterior image. For each image, a three level Daubechies 4 wavelet transform is used.

	noise	PSNR			
test image	level σ_ε	noisy	GFM	GLG	hom. GLG
Lena	0.10	18.76	26.57	**27.93**	23.57
	0.15	15.44	25.18	**26.18**	20.68
	0.20	13.17	24.15	**24.72**	18.77
Mandrill	0.10	19.18	22.68	**23.39**	22.47
	0.15	15.77	21.23	**21.61**	19.70
	0.20	13.49	20.20	**20.52**	18.09
Peppers	0.10	19.18	25.99	**27.96**	24.00
	0.15	15.83	24.68	**25.87**	21.08
	0.20	13.57	23.70	**24.41**	19.18

Sun, Gu, Chen & Zhang (2004); we recap this labelling algorithm and afterwards modify it to work with the GLG model. Finally, we discuss how to transfer these labels to the pixels and show examples.

The labelling in Sun et al. (2004) consists of three steps. First, using the EM algorithm of Crouse et al. (1998), an estimate $\widehat{\theta}$ of the parameter vector θ of a 2-state GFM model is obtained from the data $\{\mathbf{w}^{(t)}\}_{t=1}^{k}$. Second, using an empirical Bayesian approach, the maximum a posteriori (MAP) estimate of the hidden states

$$\widehat{\mathbf{s}^{(t)}} = \operatorname*{argmax}_{\mathbf{s}} p(\mathbf{s}|\mathbf{w}^{(t)}, \widehat{\theta}) = \operatorname*{argmax}_{\mathbf{s}} p(\mathbf{s}, \mathbf{w}^{(t)}|\widehat{\theta}), \tag{28}$$

$t = 1, \ldots, k$, is computed using the Viterbi algorithm (Durand et al. 2004). Third, the MAP estimate is used to define $f_i^{(t)} = (\widehat{\mathbf{s}^{(t)}})_i$.

The idea of labelling wavelet coefficients with the GLG model is overall the same as presented above for the GFM model, with the differences arising from the continuous nature of the hidden states and different algorithms being applied for parameter estimation and state estimation. First, the EM algorithm in Section 3.2.2 is used to provide an estimate $\widehat{\theta}$ of the parameter vector of the GLG model. Second, in analogy with (28) we compute the MAP estimate $\widehat{\mathbf{s}^{(t)}}$. However, the Viterbi algorithm cannot be used here: The Viterbi algorithm computes the MAP estimate by successively maximizing the terms in (1) associated to each level of the wavelet tree. For the GFM model, it is easy to perform these maximization steps due to the fact that the hidden state space is finite. For the GLG model, the MAP estimate can be computed at the finest level,

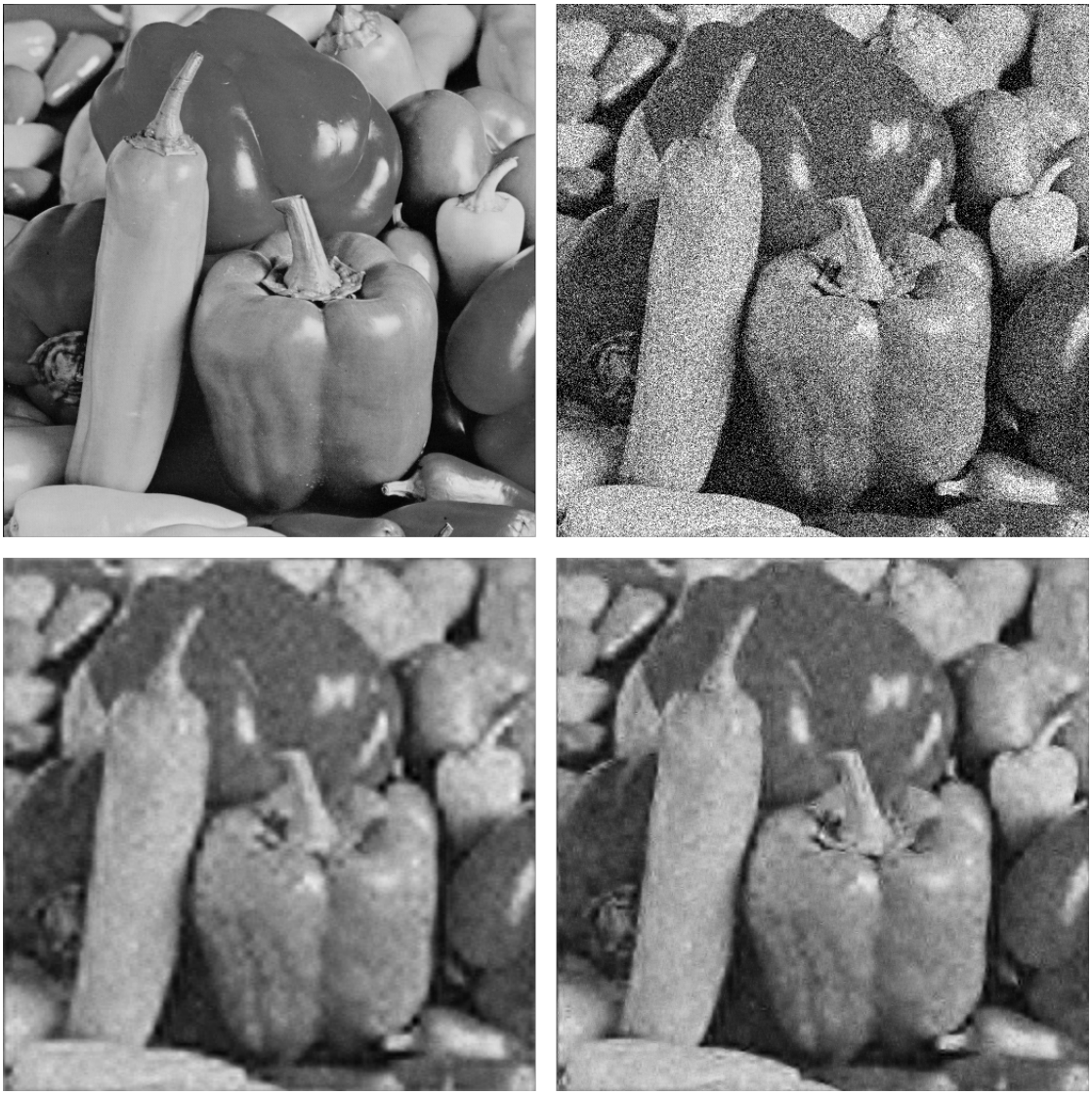

Figure 5: Denoising results for the peppers image from Table 1 when the standard deviation of the noise is 0.20. Top left panel: The original image. Top right panel: The noisy image (PSNR is 13.57). Bottom left panel: The noisy image cleaned using the GFM model and the EM-algorithm (PSNR is 23.70). Bottom right image: The noisy image cleaned using the GLG model and the EM-algorithm (PSNR is 24.41).

but this estimate is a complicated function that cannot easily be used in the remaining

Figure 6: Denoising the 'peppers' image using the posterior distribution (27) and INLA. The original and noisy images are seen in Figure 5. The top left, top right, and bottom left images are based on the 25%, 75%, and 50% quantiles of the posterior distribution, respectively (the PSNRs are 16.38, 16.41, and 19.18, respectively). The bottom right image is based on the mean of the posterior distribution (PSNR is 19.15). The posterior mean and median are almost identical.

maximization steps. Instead, we note that

$$p(\mathbf{s}, \mathbf{w}|\widehat{\theta}) = p(\mathbf{s}|\widehat{\theta}) \prod_{i=1}^{n} p(w_i|s_i) \tag{29}$$

where $p(\mathbf{s}|\widehat{\theta})$ is a multidimensional Gaussian density function with mean vector $\widehat{\boldsymbol{\mu}}$ and precision matrix $\widehat{\Delta}$ given by (24) with $\theta = \widehat{\theta}$. The log of (29) and its gradient vector and Hessian matrix with respect to $\mathbf{s}$ are

$$\log p(\mathbf{s}, \mathbf{w}|\widehat{\theta}) \equiv -\frac{1}{2}\Big\{(\mathbf{s}-\widehat{\boldsymbol{\mu}})^\top \widehat{\Delta}(\mathbf{s}-\widehat{\boldsymbol{\mu}}) + \sum_{i=1}^{n}\big(w_i^2\exp(-s_i)+s_i\big)\Big\},$$

$$\nabla \log p(\mathbf{s}, \mathbf{w}|\widehat{\theta}) = -\widehat{\Delta}(\mathbf{s}-\widehat{\boldsymbol{\mu}}) + \frac{1}{2}\big[w_i^2\exp(-s_i)-1\big]_{1\le i\le n},$$

$$H\big(\log p(\mathbf{s}, \mathbf{w})\big|\widehat{\theta}\big) = -\widehat{\Delta} - \frac{1}{2}\operatorname{diag}\big(w_i^2\exp(-s_i), 1\le i\le n\big),$$

where $\equiv$ means that an additive term which is not depending on $\mathbf{s}$ has been omitted in the right hand side expression. The Hessian matrix is strictly negative definite for all $(\mathbf{s}, \mathbf{w})$ with $\mathbf{w} \neq \mathbf{0}$ and hence $\widehat{\mathbf{s}^{(t)}}$ can be found by solving $\nabla \log p(\mathbf{s}, \mathbf{w}^{(t)}|\widehat{\theta}) = 0$ using standard numerical tools. Third, observe that if the estimate $(\widehat{\mathbf{s}^{(t)}})_i$ is large in the estimated distribution $N(\widehat{\boldsymbol{\mu}}_{\rho(i)}, \widehat{\sigma^2}_{\rho(i)})$ for s_i, then we expect $w_i^{(t)}$ to be "large". Therefore, denoting z_p the p-fractile in $N(\widehat{\boldsymbol{\mu}}_{\rho(i)}, \widehat{\sigma^2}_{\rho(i)})$ (with e.g. $p = 0.9$), we define $f_i^{(t)} = 1$ if $(\widehat{\mathbf{s}^{(t)}})_i \ge z_p$ and zero otherwise.

It remains to specify the transfer of $f_i^{(t)}$ (defined by one of the two methods above) to the pixel domain (this issue is not discussed in Sun et al. (2004)). For specificity, consider a gray scale image $I = \{p_j\}_{j=1}^{kn}$ and $\{\mathbf{w}^{(t)}\}_{t=1}^{k} = W\{p_j\}_{j=1}^{kn}$, where W is the used wavelet transform operator. To each pixel j we associate a binary variable e_j indicating if j is part of an edge or not: Since the wavelet transform does not necessarily map binary values to binary values, we define $\{\widetilde{e}_j\}_{j=1}^{kn} = W^{-1}\{\mathbf{f}^{(t)}\}_{t=1}^{k}$ and set

$$e_j = \begin{cases} 1 & \text{if } \widetilde{e}_j \neq 0, \\ 0 & \text{otherwise.} \end{cases}$$

The e_j's are sensitive to the choice of W, and using the Haar wavelet results in thin edges.

As mentioned, the multiresolution analysis of the wavelet transform allows us to consider edges that are present at only specific scales. To exclude edges at a level l in the wavelet transform, we simply modify $\{\mathbf{f}^{(t)}\}_{t=1}^{k}$ by setting $f_i^{(t)} = 0$ if $\ell(i) = l$. Figure 7 compares the results of the two edge detection algorithms, where we only use the finest scale in the wavelet transform. The method based on the GLG model classifies fewer pixels as edges; in particular the GFM model classifications include many false positives. While the images within Figure 7 are comparable, we notice they are not directly comparable to the images presented in Sun et al. (2004) who use a non-decimated wavelet transform and an extension of the GFM model where the different directions are not modelled independently.

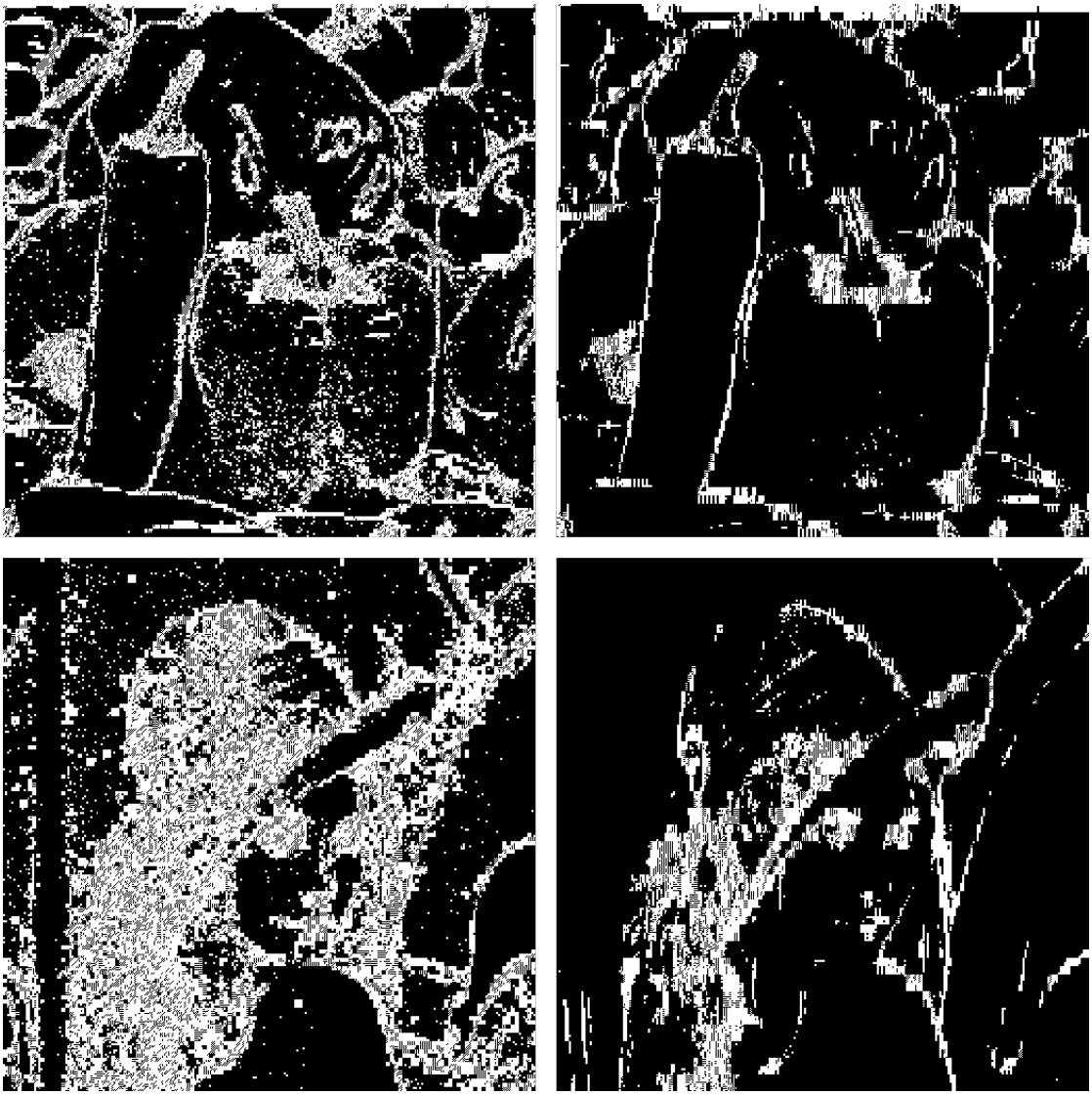

Figure 7: Examples of edge detection of the 'Lena' and 'peppers' images using the method from Sun et al. (2004) (left column) and our variant that uses the GLG model (right column). A three level Haar wavelet transform is used and only the finest level of the wavelet transform is considered. The 90% fractile is used for thresholding with the GLG model.

6 Concluding remarks

We have introduced the GLG model for wavelet trees, developed methods for performing inference, and demonstrated possible applications in signal and image processing, where the GLG model outperforms the GFM model of Crouse et al. (1998). However, there

is still work to be done. We do not have a procedure for likelihood determination of a full wavelet tree given the model parameters in the general GLG model (it is possible to compute the likelihood in INLA, but this is only for submodels). In the GFM model this likelihood is calculated as a by-product of the EM algorithm in Crouse et al. (1998), but as noted we cannot easily modify this EM algorithm to the GLG model.

As an alternative method for inference we have considered a variational EM algorithm (see e.g. Khan (2012)). The parameter estimates obtained with this variational method may be more consistent across the levels of the wavelet transform. We have omitted a further discussion of this variational method, since it cannot be used for making inference with noisy observations.

Acknowledgment

Supported by the Danish Council for Independent Research — Natural Sciences, grant 12-124675, "Mathematical and Statistical Analysis of Spatial Data", and by the Centre for Stochastic Geometry and Advanced Bioimaging, funded by a grant from the Villum Foundation. We are grateful to Håvard Rue for help with INLA. We thank Peter Craigmile, Morten Nielsen, and Mohammad Emtiyaz Khan for helpful discussions.

Appendix A: Moments

Using (1) and (3) and by conditioning on $\mathbf{s}$ and exploiting the conditional independence structure, we obtain

$$\mathrm{E}\left[w_i^2\right] = \exp\left(\mu_{\rho(i)} + \sigma^2_{\rho(i)}/2\right), \tag{30}$$

$$\frac{\mathrm{E}\left[w_i^4\right]}{3\left(\mathrm{E}\left[w_i^2\right]\right)^2} = \exp\left(\sigma^2_{\rho(i)}\right), \tag{31}$$

$$\frac{\mathrm{E}\left[w_i^2 w_j^2\right]}{\mathrm{E}\left[w_i^2\right]\mathrm{E}\left[w_j^2\right]} = \exp(\sigma_{i,j}) \quad \text{if } i \neq j. \tag{32}$$

For $i = 1, \ldots, n$, let $v_i = s_i - \beta_{\rho(i)} s_{\rho(i)}$ where $\beta_0 = 0$. Then

$$s_i = \sum_{j \in \mathcal{P}_{1,i}} v_j \prod_{h \in \mathcal{P}_{j,\rho(i)}} \beta_h \tag{33}$$

where $\mathcal{P}_{1,i}$ is the path of nodes from 1 to i (in the tree, and including 1 and i), $\mathcal{P}_{j,\rho(i)}$ is the path of nodes from j to $\rho(i)$ if $j \in \mathcal{P}_{1,i} \setminus \{i\}$, and we set $\prod_{h \in \mathcal{P}_{j,\rho(i)}} \beta_h = 1$ if $j = i$. Note that $v_1, \ldots, v_n$ are independent Gaussian distributed and $v_i \sim N(\alpha_{\rho(i)}, \kappa^2_{\rho(i)})$ where $\kappa_0 = \sigma_0$. Hence we immediately obtain from (33) that

$$\sigma_{i,j} = \sum_{h_0 \in \mathcal{P}_{1,i} \cap \mathcal{P}_{1,j}} \kappa^2_{\rho(h_0)} \left[\prod_{h_1 \in \mathcal{P}_{h_0,\rho(i)}} \beta_{h_1}\right]\left[\prod_{h_2 \in \mathcal{P}_{h_0,\rho(j)}} \beta_{h_2}\right]. \tag{34}$$

Finally, because of the simple one-to-one linear relationship between $(v_1, \ldots, v_n)$ and $(s_1, \ldots, s_n)$, (24) is straightforwardly derived.

Appendix B: Estimating equations based on moment relations

Assume $|c(i)| \neq 1$, $i = 1, \ldots, n$; this condition is in general satisfied in wavelet applications. Using mean value relations for the full parametrization (4) we describe a simple and fast procedure which provides consistent estimates for the parameters under (15) as the number of wavelet trees tends to infinity. Let n_r denote the number of nodes on level $r \in \{1, \ldots, l\}$.

First, by (11), (12) and (16), for each level $r = 0, \ldots, l-1$, there is a one-to-one correspondence between $(\mu(r), \sigma^2(r))$ and $(\eta^{(2)}(r), \eta^{(4)}(r))$, where

$$\mu(r) = \log\left(\eta^{(2)}(r)\right) - \sigma^2(r)/2,$$
$$\sigma^2(r) = \log\left(\eta^{(4)}(r)/3\right) - 2\log\left(\eta^{(2)}(r)\right).$$

Combining these relations with unbiased estimates given by

$$\widehat{\eta}^{(a)}(r) = \frac{1}{kn_r}\sum_{t=1}^{k}\sum_{i:\ell(i)=r}\sum_{j\in c(i)}\left(w_j^{(t)}\right)^a, \quad a = 2, 4, \quad r < l,$$

we obtain consistent estimates

$$\widehat{\mu}(r) = \log\left(\widehat{\eta}^{(2)}(r)\right) - \widehat{\sigma}^2(r)/2, \tag{35}$$
$$\widehat{\sigma}^2(r) = \log\left(\widehat{\eta}^{(4)}(r)/3\right) - 2\log\left(\widehat{\eta}^{(2)}(r)\right), \quad r < l, \tag{36}$$

Second, by (11)-(14), for each node $h \geq 1$ with $c(h) \neq \emptyset$,

$$\beta_h = \frac{\log\left(\eta_h^{(2,2)}\right) - 2\log\left(\eta_h^{(2)}\right)}{\log\left(\xi_{h,j}^{(2,2)}\right) - \log\left(\eta_h^{(2)}\right) - \log\left(\eta_{\rho(h)}^{(2)}\right)}$$

whenever $j \in c(h)$. This combined with (15) and (17), the unbiased estimates given above, and the consistent estimates

$$\widehat{\xi}^{(2,2)}(r) = \frac{1}{kn_r}\sum_{t=1}^{k}\sum_{i:\ell(i)=r}\sum_{j\in c(i)}\left(w_i^{(t)}\right)^2\left(w_j^{(t)}\right)^2$$

and

$$\widehat{\eta}^{(2,2)}(r) = \frac{1}{kn_r}\sum_{t=1}^{k}\sum_{i:\ell(i)=r}\frac{2}{|c(i)|(|c(i)|-1)}\sum_{\substack{j_1,j_2\in c(i)\\ j_1<j_2}}\left(w_{j_1}^{(t)}\right)^2\left(w_{j_2}^{(t)}\right)^2$$

provide consistent estimates

$$\widehat{\beta}(r) = \frac{\log\big[\widehat{\eta}^{(2,2)}(r)\big] - 2\log\big[\widehat{\eta}^{(2)}(r)\big]}{\log\big[\widehat{\xi}^{(2,2)}(r)\big] - \log\big[\widehat{\eta}^{(2)}(r)\big] - \log\big[\widehat{\eta}^{(2)}(r-1)\big]} \tag{37}$$

for $r = 0, \ldots, l-1$. Finally, using (9) and (35)-(37), we obtain consistent estimates $(\widehat{\alpha}_h, \widehat{\kappa}_h^2) = \big(\widehat{\alpha}(r), \widehat{\kappa}^2(r)\big)$ for $0 \le r = \ell(h) < l$.

Remark 1 The estimating equation (35) does not guarantee that $\widehat{\sigma}^2(r) > 0$; in fact, for small wavelet datasets, we have observed that $\widehat{\sigma}^2(r)$ may be negative. For $\widehat{\sigma}^2(r)$ to be positive is equivalent to require that

$$\widehat{\eta}^{(4)}(r) > 3\big(\widehat{\eta}^{(2)}(r)\big)^2. \tag{38}$$

As $\eta^{(4)}$ is the fourth moment and $\eta^{(2)}$ the second moment of the same random variable, (38) is a much stronger condition than the usual condition for variance estimation, namely with 3 replaced by 1.

Remark 2 The estimation procedure is immediately modified to GLG submodels. In case of the homogeneous GLG model, define

$$\eta^{(2)} = \exp\left(\mu_0 + \sigma_0^2\right),$$
$$\eta^{(4)} = 3\exp\left(2\mu_0 + 2\sigma_0^2\right),$$

and in accordance with (11)-(14) corresponding unbiased estimates

$$\widehat{\eta}^{(a)} = \frac{\sum_{h\ge 1:\, c(h)\neq\emptyset} \widehat{\eta}_h^{(a)}}{c}, \quad a = 2, 4,$$

Thereby

$$\widehat{\mu}_0 = \log\left(\widehat{\eta}^{(2)}\right) - \widehat{\sigma}^2/2, \quad \widehat{\sigma}^2 = \log\left(\widehat{\eta}^{(4)}/3\right) - 2\log\left(\widehat{\eta}^{(2)}\right),$$

provide consistent estimates.

Appendix C: EM-algorithm for the marginal likelihoods

The EM-algorithm (Dempster, Laird & Rubin 1977, Gao & Song 2011) is an iterative estimation procedure which applies for steps 1–2 in Section 3.2.2 as described below.

We start by noticing that the conditional density of s_1 given w_1 is

$$p(s_1|w_1, \mu_0, \sigma_0) = \frac{p(s_1, w_1|\mu_0, \sigma_0^2)}{q(w_1|\mu_0, \sigma_0^2)} \propto \exp\left(-\frac{1}{2}\left[\frac{w_1^2}{\exp(s_1)} + s_1 + \frac{(s_1-\mu_0)^2}{\sigma_0^2}\right]\right) \tag{39}$$

where in the expression on the right hand side we have omitted a factor which does not depend on the argument s_1 of the conditional density. Note also that for $\ell(i) = r < l$, the conditional density of $\mathbf{s}_{i,c(i)}$ given $\mathbf{w}_{i,c(i)}$ is

$$\begin{aligned}
&p(\mathbf{s}_{i,c(i)}|\mathbf{w}_{i,c(i)}, \mu(r-1), \sigma^2(r-1), \alpha(r), \beta(r), \kappa^2(r)) \\
&\quad = \frac{p(\mathbf{s}_{i,c(i)}, \mathbf{w}_{i,c(i)}|\mu(r-1), \sigma^2(r-1), \alpha(r), \beta(r), \kappa^2(r))}{q(\mathbf{w}_{i,c(i)}|\mu(r-1), \sigma^2(r-1), \alpha(r), \beta(r), \kappa^2(r))} \\
&\quad \propto \exp\Bigg(-\frac{1}{2}\Bigg\{ \left[\frac{w_i^2}{\exp(s_i)} + s_i + \frac{(s_i - \mu(r-1))^2}{\sigma^2(r-1)}\right] + \\
&\qquad \sum_{j \in c(i)} \left[\frac{w_j^2}{\exp(s_j)} + s_j + \frac{(s_j - \alpha(r) - \beta(r)s_i)^2}{\kappa^2(r)}\right] \Bigg\}\Bigg). \qquad (40)
\end{aligned}$$

In step 1, suppose $(\tilde{\mu}_0, \tilde{\sigma}_0^2)$ is the current estimate. We consider the conditional expectation with respect to (39) when (μ_0, σ_0^2) is replaced by $(\tilde{\mu}_0, \tilde{\sigma}_0^2)$. Then the next estimate for (μ_0, σ_0^2) is the maximum point for the conditional expectation of the log-likelihood which is based on both $\bar{\mathbf{w}}_1$ and $\bar{\mathbf{s}}_1$; this log-likelihood is given by

$$\sum_{t=1}^{k} \log p(s_1^{(t)}, w_1^{(t)}|\mu_0, \sigma_0^2) \equiv -\frac{1}{2}\sum_{t=1}^{k}\left[\log(\sigma_0^2) + \frac{(s_1^{(t)} - \mu_0)^2}{\sigma_0^2}\right]$$

where $\equiv$ means that an additive term which is not depending on (μ_0, σ_0^2) has been omitted in the right hand side expression, cf. (18). It follows immediately that this maximum point is given by

$$\begin{aligned}
\widehat{\mu}_0 &= \frac{1}{k}\sum_{t=1}^{k} \mathrm{E}\left[s_1^{(t)}|w_1^{(t)}, \tilde{\mu}_0, \tilde{\sigma}_0^2\right], \\
\widehat{\sigma}_0^2 &= \left[\frac{1}{k}\sum_{t=1}^{k} \mathrm{E}\left[\left(s_1^{(t)}\right)^2|w_1^{(t)}, \tilde{\mu}_0, \tilde{\sigma}_0^2\right]\right] - \widehat{\mu}_0^2,
\end{aligned}$$

where the conditional expectation is calculated using (39). We do not have a closed expression for the marginal density nor its moments. Since the joint density is the product of a Gaussian density and a smooth function, the Gauss-Hermite quadrature rule (see e.g. Press, Teukolsky, Vetterling & Flannery (2002)) is well-suited for approximating the integrals using few quadrature nodes. The iteration is repeated with $(\tilde{\mu}_0, \tilde{\sigma}_0^2) = (\widehat{\mu}_0, \widehat{\sigma}_0^2)$ until convergence is effectively obtained, whereby a final estimate $(\widehat{\mu}_0, \widehat{\sigma}_0^2)$ is returned.

In step 2, suppose $(\tilde{\alpha}(r), \tilde{\kappa}^2(r))$ is the current estimate, which we use together with the estimate $(\widehat{\mu}(r-1), \widehat{\sigma}^2(r-1))$ to obtain the next estimate for $(\alpha(r), \kappa(r))$: Replacing $(\mu(r-1), \sigma^2(r-1), \alpha(r), \beta(r), \kappa(r))$ by $(\widehat{\mu}(r-1), \widehat{\sigma}^2(r-1), \tilde{\alpha}(r), \tilde{\beta}(r), \tilde{\kappa}(r))$, this estimate is the maximum point for the conditional expectation with respect to (40) of each term

in the following sum:

$$\sum_{t=1}^{k} \sum_{i:\ell(i)=r} \log p(\mathbf{s}_{i,c(i)}^{(t)}, \mathbf{w}_{i,c(i)}^{(t)} | \widehat{\mu}(r-1), \widehat{\sigma}^2(r-1), \tilde{\alpha}(r), \tilde{\beta}(r), \tilde{\kappa}^2(r))$$

$$\equiv -\frac{1}{2} \sum_{t=1}^{k} \sum_{i:\ell(i)=r} \sum_{j \in c(i)} \left[\log(\kappa^2(r)) + \frac{(s_j - \alpha(r) - \beta(r) s_i)^2}{\kappa^2(r)} \right]$$

where additive terms which do not depend on $(\alpha(r), \kappa(r))$ have been omitted, cf. (21). Now, calculate $s(r)$ defined as the average of the following conditional means:

$$s(r) = \frac{1}{k n_{r-1}} \sum_{t=1}^{k} \sum_{i:\ell(i)=r} \mathrm{E}\big[s_i^{(t)} | \mathbf{w}_{i,c(i)}^{(t)}, \widehat{\mu}(r-1), \widehat{\sigma}^2(r-1), \tilde{\alpha}(r), \tilde{\beta}(r), \tilde{\kappa}^2(r) \big].$$

It is easily seen that the maximum point is given by

$$\widehat{\beta}(r) = \frac{\displaystyle\sum_{t=1}^{k} \sum_{i:\ell(i)=r} \sum_{j \in c(i)} \mathrm{E}\big[s_j^{(t)} (s_i^{(t)} - s(r)) | \mathbf{w}_{i,c(i)}^{(t)}, \widehat{\mu}(r-1), \widehat{\sigma}^2(r-1), \tilde{\alpha}(r), \tilde{\beta}(r), \tilde{\kappa}^2(r) \big]}{\displaystyle\sum_{t=1}^{k} \sum_{i:\ell(i)=r} |c(i)| \, \mathrm{E}\big[(s_i^{(t)} - s(r))^2 | \mathbf{w}_{i,c(i)}^{(t)}, \widehat{\mu}(r-1), \widehat{\sigma}^2(r-1), \tilde{\alpha}(r), \tilde{\beta}(r), \tilde{\kappa}^2(r) \big]},$$

$$\widehat{\alpha}(r) = s(r) - \widehat{\beta}_i s(r),$$

$$\widehat{\kappa}^2(r) = \Bigg[\frac{1}{k n_{r-1}} \sum_{t=1}^{k} \sum_{i:\ell(i)=r} \frac{1}{|c(i)|} \sum_{j \in c(i)} \mathrm{E}\Big[\big(s_j^{(t)} - \widehat{\beta}(r) s_i^{(t)} \big)^2 | \mathbf{w}_{i,c(i)}^{(t)}, \widehat{\mu}(r-1), \widehat{\sigma}^2(r-1), \tilde{\alpha}(r), \tilde{\beta}(r), \tilde{\kappa}^2(r) \Big] \Bigg] - \widehat{\alpha}(r)^2.$$

The iteration is repeated with $(\tilde{\alpha}_(r), \tilde{\kappa}^2(r)) = (\widehat{\alpha}(r), \widehat{\kappa}^2(r))$ until convergence is effectively obtained, whereby a final estimate $(\widehat{\alpha}(r), \widehat{\kappa}^2(r))$ is returned.

Appendix D: Conditional expectation of noisy observations under the GLG model

Let the situation be as in Section 5.2 and consider the GLG model. The joint density of (s_i, v_i) is found just as in the noise-free case in Section 3.2.1,

$$p(s_i, v_i | \mu_{\rho(i)}, \sigma_{\rho(i)}^2) = p(v_i | s_i) p(s_i | \mu_{\rho(i)}, \sigma_{\rho(i)}^2) = \frac{\exp\Big(-\frac{1}{2} \Big[\frac{v_i^2}{\exp(s_i) + \sigma_\varepsilon^2} + \frac{(s_i - \mu_{\rho(i)})^2}{\sigma_{\rho(i)}^2} \Big] \Big)}{2\pi \sigma_{\rho(i)} \sqrt{\exp(s_i) + \sigma_\varepsilon^2}}$$

and the marginal density of the wavelet with noise is

$$q(v_i | \mu_{\rho(i)}, \sigma_{\rho(i)}^2) = \int_{-\infty}^{\infty} p(s_i, v_i | \mu_{\rho(i)}, \sigma_{\rho(i)}^2) \, \mathrm{d}s_i.$$

We do not have a closed form expression for this integral, but due to the form of the integrant we approximate the integral with the Gauss-Hermite quadrature rule, see e.g. Press et al. (2002). The conditional density of s_i given v_i is

$$p(s_i|v_i, \mu_{\rho(i)}, \sigma^2_{\rho(i)}) = \frac{p(s_i, v_i|\mu_{\rho(i)}, \sigma^2_{\rho(i)})}{q(v_i|\mu_{\rho(i)}, \sigma^2_{\rho(i)})} = \frac{\exp\Big(-\frac{1}{2}\Big[\frac{v_i^2}{\exp(s_i)+\sigma_\varepsilon^2} + \frac{(s_i-\mu_{\rho(i)})^2}{\sigma^2_{\rho(i)}}\Big]\Big)}{c(v_i|\mu_{\rho(i)}, \sigma^2_{\rho(i)})\sqrt{\exp(s_i)+\sigma_\varepsilon^2}}$$

where $c(v_i|\mu_{\rho(i)}, \sigma^2_{\rho(i)}) = 2\pi\sigma_{\rho(i)}q(v_i|\mu_{\rho(i)}, \sigma^2_{\rho(i)})$. Furthermore, from well-known results about the bivariate normal distribution we obtain

$$\mathrm{E}[w_i|s_i, v_i, \theta] = \mathrm{Corr}[w_i, v_i|s_i]\sqrt{\frac{\mathrm{Var}[w_i|s_i]}{\mathrm{Var}[v_i|s_i]}}v_i = \frac{\mathrm{Var}[w_i|s_i]}{\mathrm{Var}[v_i|s_i]}v_i = \frac{\exp(s_i)}{\exp(s_i)+\sigma_\varepsilon^2}v_i.$$

Hence

$$\mathrm{E}[w_i|v_i, \theta] = \mathrm{E}\big[\mathrm{E}[w_i|s_i, v_i, \theta]\big|v_i, \theta\big] = v_i\,\mathrm{E}\left[\frac{\exp(s_i)}{\exp(s_i)+\sigma_\varepsilon^2}\middle|v_i, \theta\right]$$

whereby we obtain (26).

References

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion), *Journal of the Royal Statistical Society: Series B* **36**(2): 192–225.

Besag, J. (1975). Statistical analysis of non-lattice data, *Statistician* **24**(3): 179–195.

Choi, H. & Baraniuk, R. (2001). Multiscale image segmentation using wavelet-domain hidden Markov models, *IEEE Transactions on Image Processing* **10**(9): 1309–1321.

Crouse, M. S., Nowak, R. D. & Baraniuk, R. G. (1998). Wavelet-based statistical signal processing using hidden Markov models, *IEEE Transactions on Signal Processing* **46**(4): 886–902.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society: Series B* **39**(1): 1–38.

Durand, J.-B., Gonçalvès, P. & Guédon, Y. (2004). Computational methods for hidden Markov tree models–an application to wavelet trees, *IEEE Transactions on Signal Processing* **52**(9): 2551–2560.

Gao, X. & Song, P. X.-K. (2011). Composite likelihood EM algorithm with applications to multivariate hidden Markov model, *Statistica Sinica* **21**(1): 165–185.

Khan, M. E. (2012). *Variational Learning for Latent Gaussian Models of Discrete Data*, PhD thesis, The University of British Columbia.

Lauritzen, S. L. (1996). *Graphical Models*, Clarendon Press, Oxford.

Martins, T. G., Simpson, D., Lindgren, F. & Rue, H. (2013). Bayesian computing with INLA: new features, *Computational Statistics and Data Analysis* **67**: 68–83.

Po, D. D.-Y. & Do, M. N. (2006). Directional multiscale modeling of images using the contourlet transform, *IEEE Transactions on Image Processing* **15**(6): 1610–1620.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. (2002). *Numerical Recipes in C++*, 2 edn, Cambridge University Press.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
URL: *http://www.R-project.org/*

Romberg, J. K., Choi, H. & Baraniuk, R. G. (2001). Bayesian tree-structured image modeling using wavelet-domain hidden Markov models, *IEEE Transactions on Image Processing* **10**(7): 1056–1068.

Rue, H. & Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*, Chapman and Hall, London. Monographs on Statistics and Applied Probability, vol. 104.

Rue, H. & Martino, S. (2007). Approximate Bayesian inference for hierarchical Gaussian Markov random field models, *Journal of Statistical Planning and Inference* **137**(11): 3177–3192.

Rue, H., Martino, S. & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations, *Journal of the Royal Statistical Society: Series B* **71**(2): 319–392.

Sun, J., Gu, D., Chen, Y. & Zhang, S. (2004). A multiscale edge detection algorithm based on wavelet domain vector hidden Markov tree model, *Pattern Recognition* **37**(7): 1315–1324.

Tierney, L. & Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities, *Journal of the American Statistical Association* **81**(393): 82–86.