To Appear: Journal of Applied Computational and Harmonic Analysis,
Special Issue on Wavelet Applications, Spring 2001.

Random cascades on wavelet trees and their use in analyzing and modeling natural images

Martin J. Wainwright*
Laboratory for Information & Decision Systems
Electrical Engineering & Computer Science
Massachusetts Institute of Technology

Eero P. Simoncelli†
Center for Neural Science
Courant Institute of Mathematical Sciences
New York University

Alan S. Willsky*
Laboratory for Information & Decision Systems
Electrical Engineering & Computer Science
Massachusetts Institute of Technology

November 17, 2000

*MW supported by NSERC 1967 fellowship; AW and MW by AFOSR grant F49620-98-1-0349 and ONR grant N00014-91-J-1004

†ES supported by NSF CAREER grant MIP-9796040 and an Alfred P. Sloan Fellowship.

Corresponding author: Martin Wainwright

Stochastic Systems Group, Building 35-425

Massachusetts Institute of Technology

Cambridge, MA 02139

Telephone: (617)-258-5278

Fax: (617)-258-8364

Email: mjwain@mit.edu

Abstract

We develop a new class of non-Gaussian multiscale stochastic processes defined by random cascades on trees of multiresolution coefficients. These cascades reproduce a semi-parametric class of random variables known as Gaussian scale mixtures, members of which include many of the best-known heavy-tailed distributions. This class of cascade models is rich enough to accurately capture the remarkably regular and non-Gaussian features of natural images, but also sufficiently structured to permit the development of efficient algorithms. In particular, we develop an efficient technique for estimation, and demonstrate in a denoising application that it preserves natural image structure (e.g., edges). Our framework generates global yet structured image models, thereby providing a unified basis for a variety of applications in signal and image processing, including image denoising, coding and super-resolution.

Running head: Random cascades on wavelet trees for modeling natural images

1 Introduction

Stochastic models of natural images underlie a variety of applications in image processing and low-level computer vision, including image coding, denoising and restoration, interpolation and synthesis. Accordingly, the past decade has witnessed an increasing amount of research devoted to developing stochastic models of images [e.g., 19, 38, 45, 48, 55]. Simultaneously, wavelet transforms and other multiresolution representations have profoundly influenced image processing and low-level computer vision [e.g. 34]. Moreover, multiscale theory has proven useful in modeling and synthesizing a variety of stochastic processes [e.g., 60, 12, 33].

The intersection of these three lines of research — statistical image models, multiscale representations, and multiscale modeling of stochastic processes — constitute the focus of this paper. More specifically, our goal is to develop and study a new class of multiscale stochastic processes that are capable of capturing the statistics of natural images. These processes are defined by random coarse-to-fine cascades on trees of wavelet or other multiresolution coefficients. Our cascade models represent a significant variation on linear models defined on multiscale trees [e.g., 8]. Although such models lead to exceptionally efficient algorithms for image processing, their linear nature means that they cannot capture the striking types of non-Gaussian behavior present in wavelet pyramids of natural images. To capture such behavior, we define random cascades that reproduce a rich semi-parametric class of random variables known as Gaussian scale mixtures (GSM). We demonstrate that the structure of our random cascade models not only captures natural image statistics, but also facilitates efficient and optimal processing, which we illustrate by application to image denoising. Preliminary forms of parts of this work have appeared in [57, 56].

1.1 The statistics of natural images

We begin with an overview of previous empirical work on natural image statistics. Typically, the term "natural images" is used in a loose fashion to denote the ensemble of visual images found in the natural environment, as opposed to other image classes (e.g., radar images). The study of image statistics dates back to the pioneering work of television engineers in the 1950s [e.g., 20, 39], who studied the autocovariance function of images. Other work has emphasized the fractal structure of natural images [e.g., 40, 19, 54]. Consistent with fractal behavior, a large body of empirical work has shown that the power spectrum of natural images obeys a $f^{-\gamma}$ law [e.g., 19, 45]. Moreover, natural images exhibit highly non-Gaussian statistical dependencies that can be revealed by examining the statistics of a multiresolution decomposition. Figure 1 contrasts the marginal distributions of wavelet coefficients for Gaussian noise with those for a

typical natural image. Plotted on the vertical axis is log probability, so that the Gaussian curve is an inverted parabola. In contrast, the marginal distribution obtained from the natural image is heavy-tailed and kurtotic. These characteristics, which are are found for a wide range of filters and natural images, have been modeled by a number of researchers [e.g., 34, 48, 55, 21].

Another important feature of natural images is their approximate *scale invariance*, meaning that their statistics are invariant (up to a multiplicative constant) to changes in scale. Intuitively, there should be no preferred scale in an ensemble of natural images, since (disregarding occlusion) the same scene is equally likely to be viewed from a range of distances. One manifestation of the scale invariance of natural images is their $f^{-\gamma}$ spectral characteristic. The marginal distributions of wavelet coefficients provide further support for approximate scale invariance. When they are renormalized by a scale-dependent factor, the resulting histograms tend to coincide, as they should for a scale-invariant process [21, 27].

While a great deal of attention has been devoted to marginal statistics of single coefficients, much less has been paid to joint statistics of groups of wavelet coefficients. Both theoretical [53] and empirical studies [e.g., 48] show that coefficients of orthonormal wavelet decompositions of natural images tend to be roughly decorrelated. More recent work has shown that, nearby wavelet coefficients, despite being roughly uncorrelated, exhibit strong dependencies. The basic form of dependency, which is surprisingly regular over a range of multiscale transforms, choice of coefficient pairs, and natural images [3, 48], is illustrated in Figure 2. Shown are two joint conditional histograms of two wavelet coefficients, which we call the "child" and its coarser scale "parent" at the same spatial position and orientation. Each column of the 2D plots corresponds to a 1D conditional histogram $p(\text{child}\,|\,\text{parent})$ for a fixed value of the parent. Light intensity corresponds to frequency of occurrence, where each column has been independently rescaled to form a conditional histogram. Panel (a) corresponds to Gaussian white noise image. As expected, the two coefficients are independent, because the shape of the cross-section $p(\text{child}\,|\,\text{parent})$ is independent of the value of the parent.

In contrast, panel (b) shows typical behavior for a natural image. Although the two wavelet coefficients are approximately decorrelated, they are highly dependent. In particular, the distribution of the child conditioned on the value of the parent has a standard deviation that scales with the absolute value of the parent. The characteristic "bow tie" shape of this histograms is found for wavelet coefficients at nearby spatial positions, adjacent orientations and spatial scales, and over a wide range of natural images. Thus, wavelet coefficients from natural images exhibit a striking *self-reinforcing* characteristic, in that if one wavelet coefficient is large in absolute value, then "nearby" coefficients (where nearness is measured in scale, position, or orientation) also are more likely to be large in absolute value.

1.2 Overview

The previous section lays out a number of striking empirical characteristics that should be reproduced by a stochastic model for images. The goals of this paper are to develop a mathematical framework for capturing the structure of natural images, and to show that it can be used as the consistent basis for a variety of image processing tasks. As with other work on natural images [e.g., 21, 43, 48], we work in terms of wavelet or other multiresolution coefficients, which can be identified with the nodes of a multiscale tree. The basis of our approach is the decomposition of wavelet coefficients into two underlying stochastic processes defined on the multiscale tree. In particular, we model wavelet coefficients as a product of one white multiscale Gaussian process with a second continuous-valued *multiplier* process. This multiplier process, which is generated as a nonlinear function of a second Gaussian multiscale process (called the premultiplier), serves to control the non-Gaussian dependencies among wavelet coefficients.

The class of marginal distributions generated by this nonlinear mixing is rich, including many of the best known and well-studied heavy-tailed processes. Moreover, the multiscale tree structure allows us to construct global probability distributions on all wavelet coefficients, and hence statistical models for natural images. We show that this framework is powerful enough capture the key characteristics of natural images described above; moreover, it does so in a parsimonious fashion, requiring only a small set of parameters. Both Gaussian processes in the underlying decomposition are modeled by the multiscale framework of [8, 33], which permits efficient and optimal algorithms. As a result, although our models produce highly non-Gaussian statistics, we are able to exploit this embedded linear-Gaussian structure to great advantage. A number of other researchers [e.g., 21, 54, 43, 44, 48] have studied and exploited the properties of natural images on which we focus here, and our approach has both some similarities and important differences with these earlier efforts. Later in the paper, we discuss these links both in image modeling (Section 3.4), and in image denoising and coding (Section 4.2).

In next section, we provide the mathematical preliminaries for our treatment, including an introduction to and some new results concerning so-called Gaussian scale mixtures. We also briefly review the relevant features of the linear multiscale modeling framework in [e.g., 8, 33]. In Section 3, we introduce the class of multiscale wavelet cascade models and illustrate the characteristics that can be captured by such models, including the highly non-Gaussian characteristics of natural images. In Section 4, we develop an algorithm for MAP estimation of the premultiplier process. On the basis of this estimator, we develop a technique for image denoising that preserves the structure of natural images. In addition, we describe an algorithm for estimating model parameters. Section 5 provides illustrative results of applying the wavelet denoising

algorithm to both 1D signals and natural images. Section 6 summarizes our work, and points out directions for future work.

2 Mathematical preliminaries

This section develops mathematical preliminaries necessary for defining random cascades on wavelet trees. We begin by introducing the semi-parametric class of random variables known as Gaussian scale mixtures (GSMs), and providing some analysis of their properties required for our development. The second subsection reviews the relevant aspects of previous work on linear multiscale stochastic processes.

2.1 Gaussian scale mixtures

In this section, we introduce and describe some of the basic properties of GSMs, including several new results whose proofs can be found in Appendix A. To begin, a GSM vector $\mathbf{c}$ is formed by taking the product of two independent random variables, namely a positive scalar random variable z known as the *multiplier* or *mixing variable*, and a Gaussian random vector $\mathbf{u}$ distributed as[1] $\mathcal{N}(0, \Lambda)$. With this notation, we have $\mathbf{c} \stackrel{d}{=} \sqrt{z}\mathbf{u}$, where $\stackrel{d}{=}$ denotes equality in distribution.

The choice of mixing variable specifies the GSM variable $\mathbf{c}$ with associated GSM density $p_{\mathbf{c}}$. In particular, the GSM density can be represented as an integral of a Gaussian kernel function scaled and weighted by the mixing variable:

$$p_{\mathbf{c}}(\mathbf{c}) = \int_0^\infty \frac{1}{(2\pi)^{m/2}|z\Lambda|^{1/2}} \exp\big(-\frac{\mathbf{c}^T\Lambda^{-1}\mathbf{c}}{2z}\big)\, p_z(z)dz \tag{1}$$

where p_z is the density of the mixing variable, and m is the dimension of the random vector $\mathbf{c}$. As a special case, the finite mixture of Gaussians corresponds to choosing p_z to be a (discrete) probability mass function, in which case the integral reduces to a finite sum.

A first question concerns characterizing which random vectors can be represented as GSMs. For simplicity in notation, we focus on the case of a scalar GSM, although the results can be stated more generally. We begin with a few definitions. First of all, recall that the characteristic function of a random variable c is given by $\phi_c(s) = \int_{-\infty}^{\infty} \exp(i\,c\,s)p_c(c)dc$, where p_c is the density function of c. We also need the notion of complete monotonicity: a function f defined on $(0, \infty)$ is *completely monotone* if it has derivatives $f^{(n)}$ of all orders, and $(-1)^n f^{(n)}(y) \geq 0$ for all $y > 0$ and $n = 0, 1, 2, \ldots$. With these definitions, we have the following necessary and sufficient conditions:

[1]The notation $x \sim \mathcal{N}(\mu, \Lambda)$ means that x is distributed as a Gaussian with mean μ and covariance Λ.

Theorem 1. A symmetric random variable c with characteristic function $\phi_c(t)$ is a GSM if and only if $g(s) \triangleq \phi_c(\sqrt{s})$ is completely monotone.

Proof. See Appendix A. □

Andrews and Mallows [1] provide the following necessary and sufficient conditions on the density function:

Theorem 2. Let c have a density function p_c that is symmetric about zero. Then c is a GSM if and only if $f(y) \triangleq p_c(\sqrt{y})$ is completely monotone.

These two theorems provide straightforward criteria for a GSM in the characteristic function and density domains respectively.

The family of Gaussian scale mixtures includes several well-known families of random variables, including those shown in Table 1. The densities of these variables are characterized by a scale parameter λ and a parameter α that controls the heaviness of the tails. Each family typically exhibits a range of tail behavior as α varies, ranging from Gaussian to very heavy-tailed. In fact, although the scale parameter λ is analogous to a variance, the tails of many of these variables are so heavy that variances fail to exist. A classical example is the α-stable family, which has been extensively studied [see 46]. The case $\alpha = 2$ corresponds to the familiar Gaussian, whereas variables with smaller $\alpha > 0$ have increasingly heavy tails. A well-known example with heavy tails is the Cauchy distribution, which corresponds to $\alpha = 1$. The generalized Gaussian family, also known as the generalized Laplacian family, is described by a parameter $\alpha \in (0, 2]$. The choice $\alpha = 2$ again corresponds to a Gaussian, whereas $\alpha = 1$ is a symmetrized Laplacian. The generalized Gaussian family is often used to model the marginals of wavelet coefficients [e.g., 34, 37, 21, 50], where the tail parameter when fit to empirical histograms is typically less than one. The symmetrized gamma family is also important because it (like the α-stable) is infinitely divisible [17], a property emphasized in the context of natural images in [27].

For most of the random variables in Table 1, it is either well-known or straightforward to find the density of the multiplier variable. For the generalized Gaussian family, however, this verification is not entirely straightforward. In order to show that the generalized Gaussian is a GSM, we first need to formally develop a relation apparent in Table 1 (e.g., compare symmetrized gamma and generalized Student variables).

Theorem 3. Let $c \stackrel{d}{=} \sqrt{z}\, u$ be a GSM with characteristic function ϕ_c, and let the mixing variable z have density p_z. Define $f(v) \triangleq p_z(v)/\sqrt{v}$, and suppose that $\int_0^\infty f(v)dv < \infty$, in which case we can consider a random variable v with the density f. Then the GSM $y \stackrel{d}{=} \frac{1}{\sqrt{v}} u$ has density $p_y(y) \propto \phi_c(y)$.

Proof. See Appendix A. □

On the basis of Theorem 3, one would conjecture that the generalized Gaussian family should have a representation $c \stackrel{d}{=} \frac{1}{\sqrt{v}}u$, with the density of v satisfying $f(v) \propto p_{\frac{\alpha}{2}}(v)/\sqrt{v}$, where $p_{\frac{\alpha}{2}}$ is the density of a positive $\frac{\alpha}{2}$-stable random variable. In order to prove this conjecture, it is necessary to verify that f (as defined above) valid density function: i.e. that $\int_0^\infty f(v)\,dv < \infty$. This verification is not entirely straightforward, because with certain exceptions (e.g., $\alpha = \frac{1}{2}$), there are no explicit forms for the positive α-stable densities. Nonetheless, it can be proved by using properties of positive α-stable densities [17], and we summarize the result in the following:

Proposition 1. The generalized Gaussian family has the representation $c \stackrel{d}{=} \frac{1}{\sqrt{v}}u$, where in particular, v has the density proportional to $p_{\frac{\alpha}{2}}(v)/\sqrt{v}$, and $p_{\frac{\alpha}{2}}$ is the density of a positive $\frac{\alpha}{2}$-stable variable.

Proof. See Appendix A. □

In this paper, we will frequently exploit the fact that a large class of non-negative multipliers z can be generated by passing a Gaussian random variable x through the appropriate function $h : \mathbb{R} \to \mathbb{R}^+$. The following result characterizes those GSMs that can be represented in this way:

Proposition 2. Let $c \stackrel{d}{=} \sqrt{z}u$ be a GSM, and suppose that the cumulative distribution function F of the multiplier is invertible. Then c has an equivalent representation $c \stackrel{d}{=} h(x)u$ for an appropriate function $h : \mathbb{R} \to \mathbb{R}^+$, where $x \sim \mathcal{N}(0, 1)$.

Proof. Let F and G be the CDFs of z and x respectively. Since the inverse function $F^{-1} : [0, 1] \to \mathbb{R}^+$ is defined, we have $z \stackrel{d}{=} F^{-1}(G(x))$, and $h(x) \triangleq \left[F^{-1}(G(x))\right]^{1/2}$ is the appropriate function. □

According to this representation, the multiplier z is given by $h^2(x)$. We refer to the Gaussian quantity x as the *premultiplier* since it is the stochastic input to the nonlinearity h that generates the multiplier. The conditions of Proposition 2 (i.e., invertible cumulative distribution function F) will be satisfied under a variety of conditions, including when the density p_z is nowhere zero on $(0, \infty)$. This latter condition includes all random variables listed in Table 1.

In many cases, it is possible to determine explicitly the form of h. For example, choosing $h(x) = |x|$ will generate the square root of gamma variables of index 1/2, which allows us to produce the symmetrized Gamma variable of index 1/2. For the purpose of application, the precise form of GSM may not be critical. In this context, an advantage of the GSM framework is that it does not require an explicit

form of the density of c, but instead focuses attention on the multiplier. Our set-up allows an arbitrary choice of the nonlinearity h, meaning that it permits the use of GSMs which may confer a computational or analytical advantage. For the results in this paper, we will choose h from parameterized families of functions that generate random variables with ranges of behavior. One example is the family of functions $\{(\exp(x/\alpha) \mid \alpha > 0\}$, corresponding to the lognormal family listed in Table 1. Another choice is the family $\{(x^+)^\alpha \mid \alpha > 0\}$, which generates a class of variables with a range of tail behavior that is qualitatively similar to the symmetrized Gamma and generalized Gaussian families.[2]

The GSM class includes many random variables with tails so heavy that variances and lower moments may fail to exist. Such variables are characterized by polynomial decay in the tails of the distribution, where the prototypical example is the α-stable family for $\alpha < 2$. Polynomially-decaying tails are not appropriate for modeling the wavelet coefficients of natural images, for which the tails tend to drop off more quickly. Therefore, for the applications to natural images in this paper, we consider GSMs for which variances exist. Such variables can still exhibit highly non-Gaussian tail behavior, as will be clear in our modeling of wavelet marginal densities.

2.2 Multiscale stochastic processes

In this section, we introduce some of the basic concepts and results concerning linear multiscale models defined on trees. We limit our treatment to those aspects required for subsequent development; the reader is referred to other literature [e.g., 8, 33, 18, 12] for further details of these models, and their application to a variety of 1-D and 2-D statistical inference problems.

The processes of interest to us are defined on a tree $\mathcal{T}$, such as that illustrated in Figure 3. The nodes $s \in \mathcal{T}$ are organized, as depicted in the figure, into a series of scales, which we enumerate $m = 0, 1, \ldots, M$. At the coarsest scale $m = 0$ (the top of the tree) there is a single node $s = 0$, which we designate the *root node*. At the next finest scale $m = 1$ are q nodes, that correspond to the *children* of the root node. We specialize here to regular trees, so that each parent node has the same number of children (q). This procedure of moving from parent to child is then applied recursively, so that a node at scale $m < M$ gives birth to q children at the next scale $(m + 1)$. These children are indexed by $s\alpha_1, \ldots s\alpha_q$. Similarly, each node s at scale $m > 0$ has a unique parent $s\bar{\gamma}$ at scale $(m - 1)$. It should be noted that such trees arise naturally from multiresolution decompositions. For instance, a wavelet decomposition of a 1D signal generates a binary tree ($q = 2$), whereas decomposing an image will generate a quadtree ($q = 4$).

[2]Here the notation x^+ denotes the positive part of x, defined by $x^+ = x$ for $x \geq 0$ and 0 otherwise.

To define a multiscale stochastic process, we assign to each node of the tree a random vector $x(s)$. The processes of interest to us are a particular class that are Markov with respect to the graph structure of the tree. In particular, a multiscale Markov tree process $x(s),\ s \in \mathcal{T}$ has the property that for any two distinct nodes $s,\ t \in \mathcal{T}$, $x(s)$ and $x(t)$ are conditionally independent given $x(\tau)$ at any node τ on the unique path from s to t. For example, if we define $s \wedge t$ as the coarsest-scale node on this path (also the nearest common ancestor of s and t), then $x(s)$ and $x(t)$ are independent given $x(s \wedge t)$.

Multiscale processes in which the random variables $x(s)$ at each node assume a discrete set of values represent a generalization of the usual (discrete) Markov chain to more general tree graphs. A number of researchers have studied and made use of such discrete multiscale processes [e.g., 7, 10]. Of particular relevance here is the work of Baraniuk and colleagues [10, 43], who have used such discrete multiscale stochastic processes as part of their non-Gaussian modeling framework for signal and image processing. In Section 3.4, we briefly discuss this work and its relationship to our framework.

The class of multiscale Markov processes of interest to us are Gaussian processes specified by the distribution $x(0) \sim \mathcal{N}(0, P_x(0))$ at the root node, together with coarse to fine dynamics

$$x(s) = A(s)x(s\bar{\gamma}) + B(s)w(s) \tag{2}$$

where the process noise is white[3] on $\mathcal{T}$. The vector $x(s)$ at each node is distributed as $\mathcal{N}(0, P_x(s))$ where the covariance $P_x(s) \triangleq \mathbb{E}[x(s)x^T(s)]$ evolves according to the discrete-time Lyapunov equation:

$$P_x(s) = A(s)P_x(s\bar{\gamma})A^T(s) + Q(s) \tag{3}$$

where $Q(s) \triangleq B(s)B^T(s)$. In this paper, we will pay particular attention to stationary processes, for which we have $A(s) = A$, $B(s) = B$, and $P_x(s) = P_x$ for all nodes $s \in \mathcal{T}$, where the covariance P_x is the solution of the Lyapunov equation $AP_xA^T + BB^T = P_x$. Processes defined according to the dynamics in equation (2) are called multiscale autoregressive (MAR) processes. It has been shown that the MAR framework can effectively model a wide range of Gaussian stochastic processes, including one-dimensional Markov processes [31, 33], $1/f$-like processes [8, 12, 11, 32, 60], and Markov random fields [33, 32].

An additional benefit of the MAR framework is that it leads to extremely efficient algorithms for estimating the process $x(s)$ on the basis of noisy observations of the form $y(s) = C(s)x(s) + v(s)$ where $v(s)$ is a zero-mean white noise process with covariance $R(s)$. In particular, the optimal estimates of $x(s)$ at every node of the tree based on $\{y(s), s \in \mathcal{T}\}$ can be calculated very efficiently by a direct algorithm [8] that is a generalization of two-pass algorithms for estimation of time series (e.g., the Rauch-Tung-Streibel

[3]Here we assume without loss of generality that means are zero, since it is straightforward to add in non-zero means.

smoother [42]). It consists of an upward pass from the leaf nodes to the root, followed by a downward pass from the root to the leaves. The computational complexity is $\mathcal{O}(d^3 N)$ where d is the maximal dimension of $x(s)$ at any node, and N is the total number of nodes. This same algorithm also computes $P_e(s)$, the covariance of the error $[x(s) - \widehat{x}(s)]$ at each node $s \in \mathcal{T}$.

For notational reasons, it is useful to write down a vectorized form of the solution to the estimation problem. Let $\mathbf{x}$ be a vector formed by stacking the vectors $x(s)$ from each node $s \in \mathcal{T}$ in a fixed order, and define $\mathbf{y}$ analogously so that $\mathbf{y} = C\mathbf{x} + \mathbf{v}$ where C is a block diagonal matrix comprised of the $C(s)$ matrices, and $\mathbf{v} \sim \mathcal{N}(0, R)$ where R is the block diagonal matrix formed using the $R(s)$ matrices. The Bayes least-squares (BLS) and maximum a posteriori (MAP) estimates are identical in this case, and are given by

$$\widehat{\mathbf{x}} = P_{\mathbf{e}} C^T R^{-1} \mathbf{y} \qquad\qquad P_{\mathbf{e}} = \left[P_{\mathbf{x}}^{-1} + C^T R^{-1} C\right] \tag{4}$$

where $P_{\mathbf{e}}$ is the covariance of the error $\mathbf{e} = \mathbf{x} - \widehat{\mathbf{x}}$. It is important to realize that for typical image processing problems (with several hundred thousand nodes), $\widehat{\mathbf{x}}$ and P_e are of extremely high dimension, and thus their computation as suggested by equation (4) is prohibitive. Instead, the fast tree algorithm solves the set of equations $P_{\mathbf{e}}^{-1}\widehat{\mathbf{x}} = C^T R^{-1}\mathbf{y}$ and simultaneously computes the diagonal blocks of $P_{\mathbf{e}}$, with the two pass procedure outlined previously.

3 Random cascades on wavelet trees

In this section, we introduce and develop a new type of multiscale stochastic process defined by random cascades on trees. In particular, each tree node corresponds to a vector of wavelet or multiresolution coefficients, and the cascade process is constructed so as to produce a GSM vector at each node. We show that the GSM variables produced by these cascade processes account well for the statistical properties of wavelet decompositions of natural images, including self-similarity, kurtotic and heavy-tailed marginal histograms, and self-reinforcement among local groups of coefficients.

3.1 Cascades of Gaussian scale mixtures

As noted previously, naturally associated with a multiresolution decomposition like the wavelet transform is a tree of coefficients (a binary tree for 1D signals; a quadtree for images). Lying at each node is a random vector $c(s)$, which will be used to model a vector of d wavelet coefficients at the same scale and position, but different orientations. Using the decomposition of Proposition 2, we model the wavelet vector $c(s)$ as

a GSM of the form

$$c(s) \stackrel{d}{=} h(x(s)) \odot u(s) \tag{5}$$

where $x(s)$ and $u(s)$ are $d-$dimensional independent Gaussian random vectors. Here the nonlinearity h acts element-wise on the vector $x(s)$, and $\odot$ denotes element-wise multiplication of the two d-vectors. We assume that h has been appropriately normalized so that $\mathbb{E}[h^2(x_k(s))] = 1$ for $k = 1, \ldots d$, where $x_k(s)$ denotes the k^{th} element of the vector $x(s)$, so that $u(s)$ controls the variance of $c(s)$.

To specify a multiscale stochastic process, we need to define parent-to-child dynamics on the underlying state variables $x(s)$ and $u(s)$. Recall that for wavelet coefficients of natural images, the parent and child vectors are close to decorrelated. We can express the covariance between $c(s)$ and its parent $c(s\bar{\gamma})$ as $\text{cov}\left[c(s), c(s\bar{\gamma})\right] = \mathbb{E}\left\{ h(x(s))[h(x(s\bar{\gamma}))]^T \right\} \odot \text{cov}\left[u(s), u(s\bar{\gamma})\right]$, where we have used the independence of x and u. This relationship shows that the decorrelation of $c(s)$ and $c(s\bar{\gamma})$ is determined by the u process. Therefore, to model wavelet coefficients of natural images, it is appropriate to choose $u(s)$ as a white noise process on the tree $\mathcal{T}$, uncorrelated from node to node. In contrast, the vector $x(s)$ must depend on its parent $x(s\bar{\gamma})$, in order to capture the strong property of local reinforcement in wavelet coefficients of natural images. Therefore, the GSM representation of equation (5) decomposes the wavelet vector $c(s)$ into two random components, one of which controls the correlation structure, while the other controls reinforcement among wavelet coefficients.

We model the white noise process $u(s)$ as

$$u(s) = D(s)\zeta(s)\,, \qquad\qquad \zeta(s) \sim \mathcal{N}(0, I) \tag{6}$$

so that $D(s)$ controls any scale-to-scale variation (and hence the scaling law) for the process. To capture the dependency in the premultiplier process $x(s)$, we use a MAR model:

$$x(s) \;\; = \;\; Ax(s\bar{\gamma}) + Bw(s) \tag{7}$$

with $x(0) \sim \mathcal{N}(0, P_x(0))$ and $\zeta(s) \sim \mathcal{N}(0, I)$ at the root node. Although we specialize here to the stationary case of a MAR model (i.e., $A(s) \equiv A$ and $B(s) \equiv B$ for all nodes $s \in \mathcal{T}$), it is clear that GSM cascades with non-stationary MAR dynamics are also possible. Figure 4 provides a graphical representation of this model structure for two levels of a binary tree. The premultiplier process $x(s)$ and white noise $u(s)$ both live at the nodes of a multiscale tree, represented by open circles. These processes generate the wavelet coefficient vector $c(s)$, represented by filled squares, via the nonlinearity h.

Equations (5), (6) and (7) together specify the coefficients $c(s)$ of a multiresolution decomposition on a tree. For each node s, let $m(s)$ be its spatial scale, and let $p(s)$ be its spatial location in the image plane.

The quantity $c(s)$ is a random vector of wavelet coefficients for a set of different orientations at the same spatial location. For the 1-D examples shown subsequently, we use an orthonormal wavelet representation, whereas for 2-D applications to images, we use the steerable pyramid [51], an overcomplete representation that divides the image into subbands localized in both scale and orientation. A steerable pyramid can be designed with any number of orientation bands; for the work reported here, we use $d = 4$ orientations. These coefficients then define a random image via the inverse transform

$$\mathcal{I}(p_1, p_2) = \sum_{s \in \mathcal{T}} \sum_{k=1}^{d} c_i(s) \psi_{k;s}(p_1, p_2) \tag{8}$$

where (p_1, p_2) is a point in the 2-D image plane, $c_k(s)$ is the k^{th} element of $c(s)$ (corresponding to the k^{th} orientation), and $\psi_{k;s}$ corresponds to the multiresolution basis element corresponding to orientation k, and centered at scale and position $(m(s), p(s))$.

An advantage of the steerable pyramid for image processing tasks (e.g., denoising) is its translation-invariance [51]. Achieving this invariance requires overcompleteness, implying that there is redundancy in each vector of coefficients $c(s)$. In principle, this can be easily accommodated by taking $\zeta(s)$ in (6) to be a lower-dimensional random vector, so that $D(s)$ is rectangular. For the work reported here, we have taken $\zeta(s)$ to be of the same dimension as $u(s)$ and hence $c(s)$. This is not a strictly accurate model since it suggests that there are more degrees of freedom in the $c(s)$ than there should be; however, we have found this formulation to be adequate in practice.

3.2 Properties of GSM cascades

In this section, we examine the properties of random cascades of Gaussian scale mixtures on trees. We show that they are well-suited to capturing the statistical behavior of multiresolution coefficients from natural images.

3.2.1 Self-similarity

Recall that self-similarity of a process means that its statistics are invariant (up to a multiplicative constant) under any change of scale. Note that GSM tree processes, as defined above, are generated by a discrete multiresolution transform as in equation (8). Such processes can never be strictly self-similar. However, by appropriate choice of parameters, we can ensure that they satisfy a weaker form of self-similarity, known as dyadic self similarity. In particular, dyadic self-similarity of the random image $\mathcal{I}(p_1, p_2)$ means that $\mathcal{I}(p_1, p_2) \stackrel{d}{=} 2^{-k\gamma} \mathcal{I}\big(2^k (p_1, p_2)\big)$ for all integers k, where γ is a parameter. From equation (8), it can be shown

that the synthesized process $\mathcal{I}(t)$ will be dyadically self-similar if and only if the basis coefficients satisfy $c(s) \stackrel{d}{=} 2^{\gamma\left[m(t)-m(s)\right]}\, c(t)$ for all nodes $s,\ t \in \mathcal{T}$. We guarantee this condition by choosing $D(s) = 2^{-\gamma m(s)}$ in equation (6), and taking the state process $x(s)$ to be stationary, so that $x(s) \stackrel{d}{=} x(t)$ for all nodes $s,\ t \in \mathcal{T}$. The parameter $\gamma > 0$ controls the drop-off in the power spectrum of the synthesized process [e.g., 12].

3.2.2 Marginal distributions

That the marginal densities of wavelet coefficients are well-fit by at least one GSM family — namely, the generalized Gaussian with tail exponent α used as a fitting parameter — is widely known [e.g., 34, 37, 21, 50]. In previous work [55], we have demonstrated that other GSM families also provide good fits to wavelet marginals. For example, Figure 5 shows fits of the symmetrized gamma family to the histograms of marginal distributions from various natural images. Fitting was performed by numerically minimizing the Kullback-Leibler divergence between empirical and theoretical histograms. The fits are typically quite good; for instance, panel (d) shows one the worst fits that we obtained from a range of natural images.

Thus, the GSM class provides a flexible framework for choosing probabilistic models that capture real image statistics. As a result, it permits the use of GSM families that may have analytical or computational advantages — that is, families for which the multiplier distribution is easily expressed and manipulated for state and parameter estimation.

3.2.3 Self-reinforcing property

Recall that the tree-structured nature of the dynamics in equations (7) and (6) imposes a powerful Markov property on the wavelet coefficients $c(s)$. In particular, any two vectors of wavelet coefficients $c(s)$ and $c(t)$ are conditionally independent given $x(s \wedge t)$, where $s \wedge t$ denotes the nearest common ancestor in scale of nodes s and t. In this section, we exploit this property to show that the tree structure accounts for the drop-off in dependence between a pair of coefficients as the spatial separation is increased.

The contours of joint distributions of wavelet coefficients from natural images show a wide range of shapes, ranging from circular to a concave star-shape (see top row of Figure 6). Huang and Mumford [21] suggested that these joint contours might be modeled with a 2D generalized Gaussian. Here we show that the dependency structure of a random tree cascade accounts remarkably well for this range of behavior. In particular, we consider a random cascade on a multiresolution tree with $A(s) \equiv \mu I$ and $B(s) \equiv \sqrt{1-\mu^2}\ \mathrm{I}$; and $h(x) \triangleq |x|$ which generates symmetrized gamma variables of index 0.5 (see Section 2.1). The tree structure specifies the joint distribution of any pair of wavelet coefficients $c(s)$ and $c(t)$.

Plotted along the second row of Figure 6 are joint contours of log probability for pairs of steerable pyramid wavelet coefficients [51] taken from the "Mountain" image shown at the top. In this example, we used a complex-valued transform, which incorporates both even and odd phase coefficients [see 41]. Coefficient pairs are at the same spatial scale and orientation, but with a varying spatial separation of Δ pixels. The third row shows the same plots for coefficients of the simulated GSM random cascade. The shapes of the joint contours of image data and simulated model are strikingly similar. First of all, consider the pair of coefficients in quadrature phase (i.e., even and odd phase coefficients at the same spatial location, corresponding to $\Delta = 0$). The joint contours for this quadrature pair are very close to circular for natural images, as has been noted previously [61]. Likewise, the model with $\Delta = 0$ generates a pair of coefficients with circular joint contours. For a pair of nearby coefficients ($\Delta = 8$), the contours are diamond-shaped, whereas they become a concave star-shape for widely separated coefficients ($\Delta = 128$). Plotted in the last two rows are joint conditional histograms that more explicitly illustrate the dependence between the coefficient pairs. While all pairs are decorrelated, they exhibit a range of statistical dependencies. The pair in quadrature phase at the same spatial location are highly dependent, as revealed by the familiar "bow tie" shape of the joint conditional histogram. As the spatial separation Δ increases, the dependence between coefficient pairs drops off with increasing spatially separation, until the widely separated pair (third column) are extremely close to independent. This near independence is clear because the joint conditional histogram has almost constant cross-section regardless of the value of the abscissa. Thus, a GSM cascade on a tree accounts well for pairwise joint dependencies of coefficients over a full range of spatial separations.

3.3 Parameters of GSM cascades

An attractive feature of the wavelet cascade models developed here is that they are specified by a rather small set of parameters. First of all, the matrices $D(s)$ determine any scale-to-scale variation in the process, and hence the scaling law. Secondly, the choice of the nonlinearity h determines the form of the marginal distributions of wavelet coefficients, including tail behavior and kurtosis. Thirdly, the system matrices A determine the dependency of the underlying premultiplier process $x(s)$ from node to node.

Variations in $D(s)$ control the amount of power at high frequencies relative to low frequencies, and hence the overall smoothness of the process. The effect of such changes is well-understood from studies of $f^{-\gamma}$ type Gaussian processes on multiscale trees [e.g., 60, 12]. Here we investigate the effect of varying the nonlinearity h, as well as the system matrices. In particular, we simulate a one-dimensional cascade

(i.e., the wavelet representation of a 1-D process) with the parameters $D(s) = 2^{-\gamma m(s)}$ and $\gamma = 1.5$; the nonlinearity $h(x) = (x^+)^\alpha$; and system matrices $A = \mu$; and $B = \sqrt{1-\mu^2}$, where the choices of the parameter α, and the scale-to-scale dependence μ were varied.

Figure 7 shows simulated random cascades for four combinations of the parameters (α, μ) using the 'Daub4' wavelet. The first three panels in each subfigure correspond to three scales of the wavelet pyramid, ranging from coarse to fine. The fourth panel in each subfigure corresponds to the synthesized GSM process. First considering the effect of the parameter α, note that the wavelet coefficients in cascades with $\alpha = 2$ (panels (c) and (d)) exhibit sparse behavior, in that a few outlying values tend to dominate. The wavelet coefficients of images also exhibit such sparsity, in that coefficients corresponding to edges and other discontinuities will tend to dominate. Of course, for both natural images and simulated cascades, this sparsity is a reflection of heavy tails in the densities/histograms. In contrast, wavelet coefficients in the cascades corresponding to $\alpha = 0.2$ (panels (a) and (b)) are distributed much more densely. In fact, the histograms of these coefficients, as well as the behavior of the synthesized processes, are both quite close to Gaussian. Varying the parameter μ also has a dramatic effect, particularly for the cascades with $\alpha = 2$. With $\mu = 0.05$ (panels (a) and (c)), coefficients from scale to scale are close to independent, so that high valued coefficients do not tend to cluster in patterns through scale. In contrast, the high scale-to-scale dependence for the cascades with $\mu = 0.95$ manifests itself in trails of large (in absolute value) coefficients through scale. One such trail is especially apparent in panel (d). These trails through the scale space of wavelet coefficients lead to a localized area of discontinuity and sharp variations in the synthesized process. Indeed, such trails are the scale space signature of discontinuities and other structures of interest. In this respect, our GSM tree models constitute a precise analytical model for the cascade behavior exploited by successful image coders such as embedded zero-trees [e.g., 47].

3.4 Relation to previous work on image modeling

In this section, we discuss relations between GSM cascades on wavelet trees, and other approaches to image modeling. Simoncelli and colleagues [3, 48, 49] modeled the dependency between wavelet coefficients with a conditionally Gaussian model, where the variance of one wavelet coefficient depends on the absolute value of its neighbors. This local model has proven useful in a variety of applications, including image coding, denoising, and texture synthesis. Our GSM cascades capture these same dependencies, but using an auxiliary multiplier variable that controls dependencies between coefficients. The multiplier variable is defined on a multiscale tree, thereby inducing a global probability distribution on the space of images.

Huang and Mumford [21] analyzed a variety of image statistics, documenting approximate scale invariance and a range of shapes in the joint contours of empirical histograms of wavelet coefficients. Building on earlier work of Ruderman [44], Lee and Mumford [27] developed a random collage model that exhibits both translational invariance and approximate scale invariance. As discussed in Section 3.2.1, our GSM tree models satisfy an approximate form of scale invariance. Moreover, the marginal distributions of GSMs are highly kurtotic for many choices of multiplier variables, and particular choices ensure that the statistics will be infinitely-divisible (e.g., symmetrized Gamma, α-stable.) As shown in Figure 6, our GSM tree models generate a range of behaviors in the joint contours of pairs of wavelet coefficients. Thus, our GSM cascades capture many of the properties emphasized by Mumford and colleagues in a parsimonious manner.

Our work is also related to the framework for non-Gaussian signal processing developed by Baraniuk and colleagues [10], and applied to image denoising [43]. Their framework uses a hidden discrete-state process defined on a tree to capture dependencies between wavelet coefficients, which themselves are modeled as finite scale mixtures of Gaussians. Accurately modeling the heavy tails and high kurtosis of wavelet marginal distributions will typically require a large number of discrete states. The corresponding increase in the number of parameters leads to models that may not provide a parsimonious description. In contrast, we have emphasized the use of infinite parametric mixtures, which as we have shown, accurately capture both the heavy tails and high kurtosis of wavelet marginal distributions with a small number of parameters.

4 Estimation

We now turn to problems of estimation in GSM cascades on wavelet trees. Such problems involve using data or observations to make inferences about either the state (i.e., $x(s)$ and $u(s)$) of the GSM, or about unknown model parameters. Of particular interest are estimates of the premultiplier process $x(s)$, which determines the multiplier $h(x(s))$. A significant benefit of the GSM framework is that conditioned on knowledge of the premultiplier, a GSM model reduces to a linear-Gaussian system, which can be analyzed by standard techniques. In the context of image processing, estimates of the premultiplier are of potential use for a variety of applications (e.g., coding, denoising).

In this section, we develop a Newton-like algorithm for *maximum a posteriori* (MAP) estimation of the premultiplier $x(s)$ based on noisy observations. The cost of computing intermediate quantities within each iteration scales linearly in problem size, because very fast algorithms (see Section 2.2) can be applied to the underlying Gaussian-tree structure. Furthermore, under suitable regularity conditions, this algorithm has a number of desirable properties, including guaranteed convergence to a local optimum at a quadratic rate.

We then show this algorithm can be used as the basis of a method for wavelet domain denoising. Next we turn to the problem of estimating parameters that specify a GSM model, and develop a technique in which state estimates are exploited in intermediate computations. The resultant technique is an approximate form of expectation-maximization algorithm [14], where intermediate computation is again efficient due to the tree structure.

4.1 State estimation

Here we consider the problem of estimating the premultiplier $x(s)$ given noisy observations

$$y(s) \quad = \quad h(x(s)) \odot u(s) + v(s) \tag{9}$$

where $v(s) \sim \mathcal{N}(0, R(s))$ is observation noise. An interesting feature of this problem is that unlike the standard linear observation problem (see Section 2.2)), the task of estimating $x(s)$ given noiseless observations (i.e., $R(s) \equiv 0$) is *not* trivial. Indeed, even in the absence of $v(s)$, the state $u(s)$ effectively acts as a multiplicative form of noise. With the noise $v(s)$ present, we have an estimation problem that is nonlinear, and includes both additive and multiplicative noise terms.

Given that we have a dynamical system defined on a tree, optimal estimation can, in principle, be performed by a two-pass algorithm, sweeping up and down the tree. For the linear-Gaussian case described in Section 2.2, computation of the optimal estimate (which is simultaneously the Bayes' least-squares (BLS) and maximum a posteriori (MAP) estimate) is particularly simple, involving the passing of conditional means and covariances only. In general, for nonlinear/non-Gaussian problems, however, not only are the BLS and MAP estimates different, but neither is easy to compute. However, the GSM models developed here have structure that can be exploited to produce an efficient and conceptually interesting algorithm for MAP estimation.

To set up the estimation problem, let $\mathbf{x}$ denote a vector formed by concatenating the state vectors $x(s)$ at each node, and define the vector $\mathbf{y}$ similarly. Recall that the computation of the MAP estimate involves the solution of the optimization problem $\widehat{\mathbf{x}}_{MAP} \triangleq \arg\min_x \big[-\log p(\mathbf{x}|\mathbf{y}) \big]$. Herein we simply write $\widehat{\mathbf{x}}$ to mean this MAP estimate. At a global level, our algorithm is a Newton-type method applied to the objective function $f(\mathbf{x}) \triangleq -\log p(\mathbf{x}|\mathbf{y})$. That is, it entails generating a sequence $\{\mathbf{x}^n\}$ via the recursion $\mathbf{x}^{n+1} = \mathbf{x}^n + \alpha^n S^{-1}(\mathbf{x}^n) \nabla f(\mathbf{x}^n)$, where the matrix $S(\mathbf{x}^n)$ is the Hessian of f, or some suitable approximation to it; and α^n is a step size parameter. This class of methods is attractive [see 2], because under suitable regularity conditions, not only is convergence to a local minimum guaranteed, but in addition the convergence rate is quadratic. The disadvantage of such methods, in general, is that the computation

of the descent direction $\mathbf{d}^n \triangleq -S^{-1}(\mathbf{x}^n)\nabla f(\mathbf{x}^n)$ may be extremely costly. This concern is especially valid in image processing applications, where the dimension of the matrix $S(\mathbf{x}^n)$ will be of the order 10^5 or higher.

One of the most important features of our model set-up is that the computation required for each step of the Newton recursion can indeed be performed efficiently. More precisely, the computation of the descent direction is equivalent to the solution of a *linear* MAR estimation problem, allowing the efficient algorithm of [8] described in Section 2.2 to be used for its computation. In order to demonstrate this equivalence, we rewrite the objective function as $f(\mathbf{x}) = -\log p(\mathbf{y}|\mathbf{x}) - \log p(\mathbf{x}) + C$ using Bayes' rule, where C is a constant that absorbs terms not depending on $\mathbf{x}$. The vector $\mathbf{x}$ is distributed as $\mathcal{N}(0, P_\mathbf{x})$, where the large covariance matrix $P_\mathbf{x}$ is defined by the system matrices A and B in equation (7). As a result, the log prior term can be written as $\frac{1}{2}\mathbf{x}^T P_\mathbf{x}^{-1}\mathbf{x} + C$. Finally, since the data $y(s)$ at each node is conditionally independent of all other data given the state vector $\mathbf{x}$, we can write:

$$f(\mathbf{x}) = -\sum_{s=1}^{N} \log p(y(s)|x(s)) + \frac{1}{2}\mathbf{x}^T P_\mathbf{x}^{-1}\mathbf{x} + C$$

From this representation of f, it can be seen that the Hessian of f will have the form $\nabla^2 f(\mathbf{x}) = P_\mathbf{x}^{-1} + D(\mathbf{x})$, where $D(\mathbf{x})$ is a block diagonal matrix, with each block corresponding to a node s. With this form of the Hessian, the descent direction $\mathbf{d}^n$ is given by $\mathbf{d}^n = -\left[P_\mathbf{x}^{-1} + D(\mathbf{x}^n)\right]^{-1}\nabla f(\mathbf{x}^n)$. Comparing this form of the descent direction to the linear-Gaussian problem given in equation (4), it is clear that the two problems are equivalent with appropriate identification of data terms, observation matrix, and noise covariance. Further details of these identifications, as well as calculation of the Hessian, the gradient $\nabla f(\mathbf{x})$ and $D(\mathbf{x})$, can be found in Appendix B.

Note that the overall structure of this MAP estimation algorithm is of a hybrid form. The Newton-like component involves an approximation of the objective function f that is performed *globally* on the entire graph at once. Local graphical structure is exploited within each iteration where the descent direction is computed by extremely efficient and direct algorithms for linear multiscale tree problems [8]. Thus, the complexity per iteration scales as $\mathcal{O}(d^3 N)$, where N is the number of nodes, and d is the number of orientations. As a Newton method, quadratic convergence is guaranteed for suitably smooth choices of the nonlinearity. This method is distinct from extended Kalman filtering [e.g. 24], a technique for approximate estimation of nonlinear dynamic systems, because the objective function is approximated globally on the entire state trajectory at once.

Another important characteristic of the GSM framework is that conditioning on the premultiplier $x(s)$ reduces the model to the linear-Gaussian case. That is, when the multiplier is known, the observa-

tions (9) are of the standard linear-Gaussian form. If, indeed $x(s)$ were known exactly, we would have that $P_c(s) = H[x(s)]P_u(s)H[x(s)]$ where $P_u(s) = D(s)D^T(s)$ is the covariance of $u(s)$, and the matrix $H[x(s)] \triangleq \text{diag}\{h(x(s))\}$. This suggests a suboptimal estimate in which we replace $x(s)$ by $\widehat{x}(s)$ — namely:

$$\widehat{c}(s) = \widehat{P}_c(s)\big[\widehat{P}_c(s) + R(s)\big]^{-1} y(s) \tag{10}$$

where $\widehat{P}_c(s) = H[\widehat{x}(s)]P_u(s)H[\widehat{x}(s)]$. It is this form of wavelet estimator that we use in our application to image denoising in Section 5.

4.2 Relation to other estimators

There are a number of interesting links between the GSM tree estimator developed here, and previous approaches to wavelet denoising. In particular, there is a large class of *pointwise* approaches to denoising, so-called because they operate independently on each wavelet coefficient. The link to the GSM framework comes from the Bayesian perspective, in which many of these methods can be shown to be equivalent to MAP or BLS estimation under a particular kind of GSM prior for the marginal distribution. For example, soft shrinkage [15], a widely studied form of pointwise estimate, is equivalent to a MAP estimate with a certain GSM prior — namely, a Laplacian or generalized Gaussian distribution with tail exponent $\alpha = 1$ [see 4]. Specifically, suppose that the prior on x has the form $p_x(x) \propto \exp\left(-\frac{\lambda}{2}|x|\right)$, and that y is an observation of x contaminated by Gaussian noise of variance σ^2. Under these assumptions, it is straightforward to verify that the MAP estimate is given by

$$\widehat{x}_{MAP} = [y - \text{sign}(y - \tau)\,\tau]^+ \tag{11}$$

where $\tau \triangleq \frac{\lambda\sigma^2}{2}$. For the purposes of comparison, we apply this type of soft thresholding to image denoising in Section 5. Additional relations between thresholding and MAP estimators are discussed in [37]. It is shown in [50] that by varying the tail parameter α of a generalized Gaussian prior, it is possible to derive a full family of pointwise Bayes least-squares (BLS) estimators.

The GSM framework can also be related to the James-Stein estimator (JSE), a technique with an interesting and often controversial history. The JSE applies to the problem of estimating the fixed mean c of a multivariate normal distribution from noisy observations $y = c + v$, where $v \sim \mathcal{N}(0, \sigma^2 I)$, and the length of the vector quantities is p. The maximum likelihood estimate of c, which is simply the data y itself, was long thought to be best in the sense that no other estimator could achieve a lower mean-squared error (MSE) for all values of c. However, in 1961 James and Stein [23] introduced an estimator of the mean for dimension $p \geq 3$ that achieves a uniformly lower MSE for all values of c. The *empirical Bayesian*

derivation for the JSE [see, e.g. 16] provides the link to GSMs. In the empirical Bayes formulation, c is modeled as a random quantity, distributed according to $\mathcal{N}(0, \tau^2)$. If the quantity τ were known, then the Bayes least-squares (BLS) estimate of c given y would be given by $\widehat{c}_{BLS} = \left[(\tau^2)/(\tau^2 + \sigma^2)\right] y$. For τ unknown, we can imagine trying to mimic the BLSE by estimating τ^2, and then substituting this estimate into the formula for the BLSE. In fact, the JSE proceeds more directly by estimating the quantity $\frac{\sigma^2}{\tau^2+\sigma^2}$ as $\frac{(p-2)\sigma^2}{\|y\|^2}$, which can be shown [26] to be an unbiased estimate. Substituting this estimate into the BLSE formula yields the positive-part JSE, defined as $\widehat{c} = \left([\|y\|^2 - (p-2)\sigma^2]^+\right]/\|y\|^2\right)\, y$.

The link to Gaussian scale mixtures is clear. Under the empirical Bayesian interpretation, the JSE decomposes the unknown mean c into two parts $c = \tau u$ where $u \sim \mathcal{N}(0, I)$, and τ is an unknown but fixed quantity. That is, the JSE decomposes the mean into a type of Gaussian scale mixture, involving a Gaussian component u and an unknown multiplier τ. For the Gaussian scale mixtures discussed in this paper, we typically viewed τ as a random variable, and assigned it a prior under which we computed the MAP estimate. The JSE is very similar, except that it does not assign a prior to τ, and performs an operation that is very close to ML estimation of $\frac{\sigma^2}{\tau^2+\sigma^2}$. Finally, both the JSE and the GSM tree method replace the variance in standard linear-Gaussian equations (e.g., in equation (10)) by an estimated variance.

Although not always explicitly stated, many other approaches to image denoising and image coding rely on a GSM type decomposition. The roots of this approach lie in the image coding literature, where researchers in the 1970s proposed dividing DCT coefficients into groups according to their variance [6]. Similarly, Lee [28] proposed an enhancement technique that used local variances in the pixel domain, which is now implemented in the MATLAB *wiener2* routine. More recent approaches also involve modeling wavelet coefficients as a scale mixture distributions [e.g., 30, 9, 5, 29, 35, 36, 52]. Another approach is to model dependency between the variance of a subband coefficient and its neighbors directly, using a conditionally Gaussian model [3, 48, 49]. Some models permit the variance parameter to assume only a discrete set of values [e.g., 29], whereas others allows a continuum of values. The latter models effectively correspond to infinite mixture models, similar to those emphasized in the current paper.

A step common to all these techniques, whether for denoising or coding, is to estimate the multiplier or variance. Conditioned on the variance estimate, coefficients can be denoised by the standard LLS estimator in equation (10). Many approaches use a ML-like estimate for the variance parameter, based on a local neighborhood of coefficients. In such a ML framework, the variance parameter is viewed as an unknown but fixed quantity, without a prior distribution. These forms of estimator are thus very close to the James-Stein estimator discussed previously. More recently, Mihcak et al. [36] assumed an exponential distribution on the variance parameter, and performed a local and approximate form of MAP estimation.

This corresponds to a local GSM model using a symmetrized Gamma distribution with parameter $\alpha = 1$. Overall, the GSM tree framework presented in this paper represents an extension from local to global models. Our models allow an arbitrary choice of the prior on the multiplier, which is controlled by the choice of the nonlinearity h. Moreover, the GSM tree algorithm computes the MAP estimate based on a *global prior model* on the full multiresolution representation. This global prior, which incorporates the strong self-reinforcing properties among wavelet coefficients, is induced by the multiscale tree structure.

In the context of the underlying tree, our GSM cascade models are closely related to the non-Gaussian modeling framework of Baraniuk and colleagues [10]. In their models, a multiscale discrete-state multiplier process defined on a tree controls the dependency among wavelet coefficients, which are modeled as finite scale mixtures of Gaussians. Such models have proven useful in various applications, including image denoising [43]. For finite mixtures in which the multiplier variable takes on discrete values, there exist direct recursive algorithms for computing the marginal distributions of the discrete multiplier states conditioned on the data. The BLS estimate of wavelet coefficients given noisy observations can be obtained by taking expectations over these marginal distributions [see 10]. However, the computational complexity of computing marginal distributions scales exponentially as $\sim M^d$, where M is the number of multiplier states and d is the dimension of the multiplier. In practice, therefore, both the number of states and dimension of the multiplier may be limited; for example, the denoising algorithm of [43] uses a low and high variance state ($M = 2$), and a scalar multiplier at each node ($d = 1$). A small number of multiplier states means that the models may not properly capture the non-Gaussian tail behavior and high kurtosis of wavelet marginals, whereas a low multiplier dimension will restrict the modeling of dependencies between orientations. In contrast, our GSM modeling framework emphasizes infinite scale mixtures of Gaussians. As we have illustrated, these infinite mixtures accurately capture the non-Gaussian tail behavior and high kurtosis of wavelet coefficients. Regardless of the particular GSM used, the complexity of our algorithm scales as $\sim d^3$, where d is the dimension of multiplier vector at each node.

4.3 Parameter estimation

We now address the problem of estimating the parameters of a GSM random cascade model. Recall that a GSM model is specified by a small set of quantities — namely, the matrices $D(s)$ that control the scaling law; the pointwise nonlinearity h; and the system matrices A and B that control the MAR dynamics. Determining the matrices $D(s)$ amounts to estimating the variance, and hence can be done with standard methods. The nonlinearity h controls the marginal distributions, so that estimating h is similar to fitting

a parameterized distribution to the marginal histograms of wavelet coefficients, again a fairly standard procedure. The novel aspect of our GSM models are the system matrices A and B that control the scale-to-scale dependence of the underlying premultiplier process, and it is on the estimation of these quantities that we focus here. In particular, let θ be a vector of parameters that specify these system matrices, so that we write the stationary MAR dynamics as:

$$x(s) = A(\theta)x(s\bar{\gamma}) + B(\theta)w(s) \tag{12}$$

The task is to estimate the parameter vector θ on the basis of noisy observations given by equation (9).

We begin by observing that this set-up shares a characteristic common to many parameter estimation problems: namely, the estimation of θ would be relatively straightforward given the premultiplier $\mathbf{x}$. Given this property, the parameter estimation problem lends itself to the use of the expectation-maximization (EM) algorithm [14], a technique frequently used to obtain the maximum likelihood (ML) estimate of θ. Recall that the ML estimate is given by $\widehat{\theta}_{ML} = \arg\max_{\theta\in\Theta} \left[\log p(\mathbf{y}|\theta)\right]$ where Θ is the domain of θ. In accordance with its name, the EM algorithm alternates between taking *expectations* over a set of "hidden" variables $\mathbf{x}$, and then performing *maximization* of the resulting function. In particular, the E-step of iteration n involves taking the expectation of the augmented log likelihood $\log p(\mathbf{x},\, \mathbf{y};\, \theta)$ with respect to the conditional density $p(\mathbf{x}\,|\,\mathbf{y};\theta^{n-1})$, where θ^{n-1} is the parameter estimate from the previous iteration. In the standard version of the EM algorithm, the M-step entails finding the global maximum of the resulting function. However, there exist other versions of EM (often called GEM for generalized EM [14]) in which the M-step consists of taking gradient step.

A disadvantage of EM-type algorithms is that calculating the expectation over the conditional density $p(\mathbf{x}\,|\,\mathbf{y};\theta^{n-1})$ can be difficult. This problem is often encountered for continuous-valued variables, where the integrals are typically intractable. One approach in such cases is to develop an approximation $q(\mathbf{x}|\mathbf{y};\theta^{n-1}) \approx p(\mathbf{x}\,|\,\mathbf{y};\theta^{n-1}))$, and perform an approximate E-step by taking expectations with respect to the distribution q, whose form is chosen to make such expectations comparatively easy to compute. It can be shown that such approximate methods will still converge, although they need not converge to a local maximum of the log likelihood, but rather to a local maximum of a lower bound on the likelihood [25].

We have developed such an approximate EM method for parameter estimation in GSM systems, where the approximation q to the conditional density is obtained from the algorithm described in Section 4.1. It should be noted that even with an approximate form of the density, taking the expectation is not, in general, a straightforward task. Again the problem stems from the high dimensionality of the conditional density — in applications such as image processing, it will be on the order of 10^5 or 10^6. Nonetheless, we

have found that the tree structure of the problem can again be exploited to great advantage. In particular, we make use of highly efficient algorithms for Gaussian likelihood calculation on multiscale trees in order to perform gradient ascent.[4] This approximate EM algorithm itself is developed in Appendix C. Thus, by exploiting the tree structure, we obtain a tractable technique for estimating the parameters specifying the system matrices.

5 Illustrative Results

In this section, we present some illustrative results of the state estimation algorithm developed in the previous section. We focus, in particular, on the problem estimating wavelet coefficients $c(s)$ on the basis of noisy observations $y(s)$. The wavelet coefficients are generated by GSM tree dynamics, and hence lie at the nodes of a multiresolution tree. However, to illustrate the basic properties of our estimator, we first consider its application to the estimation of 1-D sequence of scalar-valued coefficients $c(s)$ from a corresponding sequence of measurements. These sequences can be thought of as the successive values of one of the components of $c(s)$ and $y(s)$ on a single coarse-to-fine path in a tree, such as that in Figure 3. Following this 1-D example, we illustrate the application of our full algorithm to perform image denoising on a multiresolution quadtree of coefficients.

5.1 Examples in 1D

We first consider a scalar GSM process obtained by sampling a GSM tree process along the unique tree path beginning at the root node and moving down the tree (from parent to child), terminating at a specified fine-scale node. Such a sample path reveals the scale-to-scale dependence inherent in a GSM tree process. We generate the process on the tree with dynamics of the form $x(s) = \mu x(s-1) + \sqrt{1-\mu^2}\, w(s)$ and $c(s) = h(x(s))u(s)$, where $u(s)$ and $w(s)$ distributed as $\mathcal{N}(0,1)$ at each node. We estimate $c(s) = h(x(s))u(s)$ on the basis of the noisy observations given in equation (9), with $R(s) = \sigma^2$.

Shown in Figure 8 are sample paths from two different GSM processes, as well as estimates based on noisy observations. The sample paths were generated with $\mu = 0.95$, and the nonlinearities $h(x) = \exp(1.5x)$ for panel (a) and $h(x) = (x^+)^3$ for panel (b). Observe that the sample paths of both GSM processes alternate between regions of low amplitude values, interspersed with regions of high amplitude process values. Changes in the premultiplier $x(s)$ cause the transition from one region to another. In both

[4] Thus, the overall procedure actually exploits tree structure *twice*: once to compute the density $q(\mathbf{x}|\mathbf{y}, \theta^{n-1})$ using the estimation algorithm of Section 4.1 and again in order to calculate the required expectation.

examples, the signal-to-noise ratio (SNR) of the noisy observations was on the order of 2.5 dB, where the SNR of the observations is defined as $SNR_{obs} = 10\log_{10}(\mathrm{var}[c(s)]/\sigma^2)$. For any estimator $\bar{c}(s)$, we can define an SNR for comparison as $SNR_{est} = 10\log_{10}(\mathrm{var}[c(s)]/\mathrm{var}\left[\bar{c}(s) - c(s)\right])$. Recall that our estimator of $c(s)$ consists of two steps: first computing the MAP estimate of $x(s)$, and then computing the mean of $c(s)$ conditioned on the data $y(s)$ and the estimate $\widehat{x}(s)$. As a result, a fair comparison is to see how the SNR enhancement of our estimator compares to that of an "ideal" case in which we know $x(s)$ exactly (so that the corresponding estimate of $c(s)$ is obtained node-by-node via standard linear estimation). For the example in Figure 8(a), our estimator achieves an SNR of 9.71 dB, while the unachievable ideal SNR is less than 0.50 dB higher. In addition to this quantitative comparison, it is also worthwhile to comment on the qualitative properties of the estimator. Note that for both GSM process, the estimator effectively suppresses noise in regions where the multiplier $h(x(s))$ is of low amplitude, while simultaneously preserving peaks in high amplitude regions. Thus, the estimator behaves in a way well-suited for data with the characteristics of natural imagery — i.e., for which it is desirable to smooth low variance regions, while simultaneously preserving edges and other discontinuities of interest.

Figure 9 illustrates statistical properties of the estimator. Plotted in panel (a) are empirical histograms of the original wavelet coefficients c, the noisy observations y, and the estimates $\widehat{c}$. Observe that the histogram of the original values shows the high kurtosis and heavy tails that are typical of a GSM. In contrast, while the noisy histogram of observations retains the heavy tails, the noise contamination removes the high kurtosis and makes it appear roughly Gaussian near the origin. The estimation routine restores the high kurtosis, as shown in the histogram of estimated coefficients.

Note that the estimate $\widehat{c}(s)$ at any node s can be viewed as a random variable given by a function $\widehat{c}(s) = G_s(\mathbf{y})$ of the vector of data $\mathbf{y}$. Plotted in panel (b) is a joint conditional histogram of noisy observations $y(s)$ and estimates $\widehat{c}(s)$ for a given node s. In particular, each column in this figure corresponds to the distribution of $|\widehat{c}(s)|$ conditioned on the corresponding value of $|y(s)|$ represented on the abscissa. Note that we always have $|\widehat{c}(s)| \leq |y(s)|$, since $\widehat{c}(s)$ is obtained multiplying $y(s)$ by an adaptive factor always less than one. Therefore, all parts of the histogram in panel (b) lie below the diagonal. For data $|y(s)|$ near zero, the estimate also tends to cluster near zero. At the other extreme, as the data becomes large in absolute value, then $|\widehat{c}(s)|$ clusters near $|y(s)|$. The overlaid solid line in panel (b) corresponds to the mean of the estimator conditioned on different values of the data. It shows that *in an average sense*, this estimator behaves similarly to a form of shrinkage or soft thresholding [e.g., 15, 50]. That is, the estimator preferentially shrinks smaller observation values while modifying larger ones much less. Based on the discussion in Section 4.2, this is not surprising since many forms of thresholding, when interpreted in

a Bayesian framework, correspond to a pointwise GSM model. Of course, it is important to emphasize that the GSM tree estimator is similar to thresholding only in this average sense. Thresholding is a deterministic operation applied pointwise to each coefficient, whereas our estimate of each coefficient is based on the full vector of data $\mathbf{y}$, using a global prior model that incorporates the strong cascade dependencies among coefficients.

5.2 Image denoising

Here we illustrate the application of the GSM-tree framework to denoising natural images, using the steerable pyramid [51]. This is an overcomplete representation that decomposes the image into subbands localized in both scale and orientation. In all cases, we use a decomposition with four orientations, which corresponds to a state dimension of $d = 4$. Therefore, lying at each node of a quadtree are the two 4-vectors $x(s)$ and $u(s)$, which are used to model the 4-vector of wavelet coefficients $c(s)$. By the notation $c_k(s)$, we mean the coefficient at scale s and orientation[5] k. We refer to a collection of all coefficients at the same scale and orientation (but different spatial positions) as a subband. Noisy observations of the wavelet coefficients are given by equation (9), where $R(s) = \sigma^2 I$.

Recall that the GSM-tree algorithm first computes the MAP estimate of the premultipliers $x(s)$, which it then uses to compute denoised wavelet coefficients via equation (10). We have experimented with different choices of the nonlinearity h, including the previously discussed families $\{\exp(x/\alpha) \,|\, \alpha > 0\}$ and $\{(x^+)^\alpha | \alpha \geq 0\}$. As a Newton-like method, convergence of the algorithm tends to be rapid for sufficiently smooth (i.e., C^2) choices of this nonlinearity. The computational cost per iteration scales linearly in the number of wavelet coefficients. Given the denoised multiresolution coefficients $c(s)$, the clean image is obtained by inverting the multiresolution decomposition.

We compare the denoising behavior of the GSM-tree algorithm to a number of other techniques. With the exception of one algorithm (MATLAB's adaptive filtering), all techniques are applied to the steerable pyramid decomposition, and involve an estimate of the subband variance. This estimate is given by $\sigma_c^2 = [\mathrm{var}(y(s)) - \sigma_n^2]^+$, where σ_n^2 is the variance of the noise in the subband (which can be computed directly from σ). All of the algorithms compared here are semi-blind, in that we assume that the noise variance σ^2 is known. The techniques to which we compare our algorithm here are:

1. *Wiener subband technique:* for each subband, compute denoised coefficients as $\widehat{c}(s) = \sigma_c^2 \left[\sigma_c^2 + \sigma_n^2\right]^{-1} y(s)$. where σ_c^2 is the variance of the subband, and σ_n^2 is the noise variance in that subband.

[5] Here the orientations $k = 1, \ldots, 4$ are ordered from vertical through to the $-45°$ orientation.

2. *Adaptive:* MATLAB's adaptive filtering routine called by *wiener.m*: it performs pixel-wise Wiener filtering with a variance computed from a local 5×5 neighborhood [see 28].

3. *Soft thresholding:* [15] For each subband, compute the soft threshold given in equation (11), where the threshold $t = \lambda\sigma_n^2/2$ is determined by the noise variance σ_n^2 and the scale parameter λ of a Laplacian distribution fit to the subband marginal.

We have applied these algorithms to a variety of natural images. In Figure 10, we depict representative results for the 256×256 "Einstein" image. Shown in Table 2 are the SNR in decibels (dB) of the denoised images for all algorithms, based on original noisy images at four levels of SNR. For all levels of SNR, the GSM tree algorithm is superior to other techniques. Figure 10 depicts cropped denoised images for the "Einstein" image (a), on the basis of the noisy observations (SNR 4.80 dB) shown in (b). Panels (c), (d), (e), and (f) show the results of the Wiener subband denoising, MATLAB adaptive filtering, thresholding, and the tree algorithm respectively.

Although the GSM-tree algorithm is superior to these other techniques, it is important to note that the method presented here is not as good as we ultimately expect to be able to achieve. The reason can be traced directly to one of the well-known limitations of tree models [22], namely that nodes corresponding to nearby spatial positions in the original image may be much farther apart in terms of tree distance (for example, variables $x(4)$ and $x(5)$ in Figure 4). As a result, although tree models are very successful at capturing longer range dependencies, they may improperly model the dependency between certain pairs of nearby variables, which can lead to artifacts. In this context, it is worth noting that Strela et al. [52] have recently obtained excellent denoising results by using a local GSM model that avoids the problems associated with a tree structure.

There are several ways to address the problem of these boundary artifacts while retaining a global probability model. One approach is the so-called overlapping tree framework of [22], which retains the tree structure but uses nodes that overlap spatially. Another is to relax the requirement of a tree structure by introducing graphical connections between wavelet coefficients that are spatially close. The addition of extra connections between spatially adjacent nodes should increase modeling power significantly. However, it also presents difficult algorithmic issues for estimation, since we can no longer exploit extremely fast tree-based algorithms. Nonetheless, there exist a number of alternative and emerging approaches, including techniques from numerical linear algebra [13], as well as our recent work on estimation in graphs with cycles [58]. Other directions for future work, including exploiting the phase information provided by complex-valued transforms, are discussed briefly in the following section.

6 Conclusion

In this paper, we have developed a semi-parametric class of non-Gaussian multiscale stochastic processes defined by random cascades on trees of multiresolution coefficients. As we have pointed out, although our methodology has strong intellectual ties to a variety of different image models and methods for image analysis, it also differs in fundamental and important ways. First of all, the power of our modeling framework is demonstrated by its ability to accurately capture both the approximate decorrelation and dramatic non-Gaussian dependencies of wavelet coefficients of natural images. This is achieved by decomposing wavelet coefficients into two underlying stochastic processes: a Gaussian white noise process is mixed with a non-Gaussian multiscale process that captures self-reinforcing dependencies. A second significant feature of our modeling framework is its parsimony: only a very small set of parameters are needed to specify a GSM wavelet cascade. This suggests that fitting such models from data is a far better-posed problem than other approaches which require many more degrees of freedom to be specified. Thirdly, our modeling framework is sufficiently structured to permit efficient application to image processing. In particular, we showed how very fast tree algorithms can be used to perform estimation, and established their effectiveness in application to image denoising.

A number of extensions to the modeling framework presented here are possible. First, previous empirical work [55] shows that a small set of multipliers is sufficient to describe a local neighborhood of wavelet coefficients. In contrast, models described in this paper use a number of multipliers equal to the number of wavelet coefficients. Estimating the order of the underlying multiplier process, though a challenging problem, is an important one in order to develop models of even more power. Second, in the current application, we have considered only fixed parametric forms of nonlinearity. Using a nonparametric form of this nonlinearity would allow the model to further adapt to the image under consideration, with no loss of efficiency. Third, using the information about phase provided by a complex-valued multiresolution decomposition [see, e.g., 41] should lead to even better image models. Finally, in order to overcome the well-known limitations of tree-structured models, we are investigating GSM processes defined on graphs with cycles (i.e., non-trees). The addition of extra edges to the graph leads to more powerful models, but also presents new challenges in performing estimation.

A Proofs on Gaussian scale mixtures

We collect here proofs of various results stated about Gaussian scale mixtures.

A.1 Proof of Theorem 1

Combining the following lemmas give us the proof of Theorem 1.

Lemma 1. Consider a GSM variable with representation $x \stackrel{d}{=} \sqrt{z}u$, and let $\phi_c(t)$ and $\psi_z(t)$ be the characteristic function and Laplace transform of c and z respectively. Then $\phi_c(t) = \psi_z(t^2/2)$.

Proof. Apply iterated expectation to the representation of $\phi_c(t) = E[\exp(jct)]$, and use the fact that the characteristic function of a $\mathcal{N}(0,1)$ variable is $\exp(-t^2/2)$. □

Lemma 2. A function g on $(0,\infty)$ is the Laplace transform of a probability distribution $F \iff g$ is completely monotone and $g(0) = 1$.

Proof. See Section (XIII; 4) of Feller [17]. □

A.2 Proof of Theorem 3

Theorem 3. Let $x \stackrel{d}{=} \sqrt{z}u$ be a GSM with characteristic function ϕ_c, and let the mixing variable z have density $p_z(u)$. Define $f(v) \triangleq p_z(v)/\sqrt{v}$, and suppose that $\int_0^\infty f(v)dv < \infty$, in which case we can consider a random variable v with the density f. Then the GSM $y \stackrel{d}{=} \frac{1}{\sqrt{v}}u$ has density $p_y(y) \propto \phi_c(y)$.

Proof. We write

$$\begin{aligned}
\phi_c(t) &= \int_{-\infty}^{\infty} \left\{ \int_0^\infty \frac{1}{\sqrt{2\pi z}} \exp\left(-\frac{u^2}{2z}\right) p_z(z)dz \right\} \exp(jut)du \\
&= \int_0^\infty \left\{ \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi z}} \exp\left(-\frac{u^2}{2z}\right) \exp(jut)du \right\} p_z(z)dz \\
&= \int_0^\infty \sqrt{z} \exp\left(-z\frac{t^2}{2}\right) \frac{p_z(z)}{\sqrt{z}}\, dz \qquad (13)
\end{aligned}$$

where we have used Fubini's theorem, and the fact that the characteristic function of a $\mathcal{N}(0,z)$ variable is $\exp(-zt^2/2)$. From the final equation, it is clear that if v has density $f(v) \triangleq p_z(v)/\sqrt{v}$, then $y \triangleq \frac{1}{\sqrt{v}}u$ is a GSM with density $p_y(y) \propto \phi_c(y)$. □

A.3 Proof of Proposition 1

The following classical result is required in the proof:

Lemma 3. For $0 < \alpha < 1$, let G_α be the distribution function of a positive α-stable variable. Then as $x \to 0$, we have $e^{x^{-\alpha}}\, G_\alpha(x) \to 0$.

Proof. See Feller [§XIII; 6 17]. □

Equipped with this result, we can now prove the proposition:

Proposition 1. The generalized Gaussian family has the representation $y \stackrel{d}{=} \frac{1}{\sqrt{v}}u$, where in particular, v has the density proportional to $p_{\frac{\alpha}{2}}(v)/\sqrt{v}$, and $p_{\frac{\alpha}{2}}$ is the density of a positive $\frac{\alpha}{2}$-stable variable.

Proof. We need to establish existence of the integral $\int_0^\infty \frac{p_{\frac{\alpha}{2}}(u)}{\sqrt{u}} du$, where $p_{\frac{\alpha}{2}}(u) = \frac{d}{du} G_{\frac{\alpha}{2}}(u)$. Integrating by parts, we obtain $\int_0^\infty \frac{p_{\frac{\alpha}{2}}(u)}{\sqrt{u}}\, du = \frac{G_{\frac{\alpha}{2}}(u)}{\sqrt{u}} \Big|_0^\infty + \frac{1}{2}\int_0^\infty \frac{G_{\frac{\alpha}{2}}(u)}{u^{\frac{3}{2}}}\, du$. Examining the first term on the right side, clearly $\lim_{u\to\infty} \frac{G_{\frac{\alpha}{2}}(u)}{\sqrt{u}} = 0$ since $G_{\frac{\alpha}{2}}(u) \leq 1$ for all $u \in \mathbb{R}$. Otherwise, we write $\frac{G_{\frac{\alpha}{2}}(u)}{\sqrt{u}} = \left[e^{u^{-\frac{\alpha}{2}}}\, G_\alpha(u)\right]\left[\frac{e^{-u^{-\frac{\alpha}{2}}}}{\sqrt{u}}\right]$. By inspection, the second term in square brackets tends to zero as $u \to 0$; using Lemma 3, the first term in square brackets also tends to zero. By the product theorem for limits, we have $\lim_{u\to 0} \frac{G_{\frac{\alpha}{2}}(u)}{\sqrt{u}} = 0$. As for the second term in the integration by parts, similar arguments show that the integral exists. □

B State estimation

Here we explicitly compute the gradient and Hessian of the objective function $f(\mathbf{x}) \triangleq -\log p(\mathbf{x}\,|\,\mathbf{y})$. To begin, we write $-\log p(\mathbf{y}|\mathbf{x}) = \frac{1}{2}\sum_{s=1}^N \Big\{ \log\det[\mathcal{B}(x(s))] + y^T(s)\mathcal{B}^{-1}(x(s))y(s) \Big\} + C$ where the matrix $\mathcal{B}(x(s)) \triangleq H(x(s))P_u(s)H(x(s)) + R(s)$ is the covariance of $y(s)$ given $x(s)$. Here $P_u(s)$ is the covariance of $u(s)$, and the matrix $H(x(s)) \triangleq \mathrm{diag}\{h(x(s))\}$. Using this expansion, we can write

$$f(\mathbf{x}) = \frac{1}{2}\sum_{s=1}^N \Big\{ \log\det[\mathcal{B}(x(s))] + y^T(s)\mathcal{B}^{-1}(x(s))y(s) \Big\} + \frac{1}{2}\mathbf{x}^T P_{\mathbf{x}}^{-1}\mathbf{x} + C \tag{14}$$

where $P_{\mathbf{x}}$ is the covariance matrix of $\mathbf{x}$, and C absorbs terms not dependent on $\mathbf{x}$. Note that P is defined by the system matrices $A(s)$ and $Q(s)$ at each node s (see equation (7)). We compute the derivative of f with respect to $\mathbf{x}$:

$$(\nabla f(\mathbf{x}))_{(s;i)} \;=\; \frac{1}{2}\,\mathrm{trace}\left[\mathcal{B}^{-1}\frac{\partial \mathcal{B}}{\partial x(s;i)}\right] - y^T(s)\mathcal{B}^{-1}\frac{\partial \mathcal{B}}{\partial x(s;i)}\mathcal{B}^{-1}y(s) + \frac{1}{2}\left[P_{\mathbf{x}}^{-1}\mathbf{x}\right]_{(s;i)}$$

where $\frac{\partial \mathcal{B}}{\partial x(s;i)} = \frac{\partial H(x(s))}{\partial x(s;i)}P_u(s)H(x(s)) + H(x(s))P_u(s)\frac{\partial H(x(s;i))}{\partial x(s;i)}$. Here the notation $(s;i)$ refers to the i^{th} element of the vector $x(s)$ at node s, and $\frac{\partial}{\partial x(s;i)}$ refers to the partial derivative with respect to this element. Similarly, the Hessian can be computed as $\nabla^2 f(\mathbf{x}) = P_{\mathbf{x}}^{-1} + D(\mathbf{x})$ where $D(\mathbf{x})$ is a block diagonal matrix.

We now show that the computation of the descent direction $\mathbf{d}^n \triangleq -\left[P_{\mathbf{x}}^{-1} + D(\mathbf{x}^n)\right]^{-1}\nabla f(\mathbf{x}^n)$ corresponds to the canonical form of a linear-Gaussian problem shown in equation (4). In particular, we let $P_{\mathbf{x}}^{-1}$ be the inverse covariance matrix in both cases; we set the inverse noise covariance $R^{-1} \equiv D(\mathbf{x}^n)$; and the observations matrix $C \equiv I$. Finally, we define a vector of fictitious data as $\mathbf{y} = -D^{-1}(\mathbf{x}^n)\nabla f(\mathbf{x}^n)$. Note that we have assumed here that the blocks of $D(\mathbf{x}^n)$ are positive definite to ensure that is constitutes a valid covariance. Satisfying this condition may require modifying D, in which case the method is not exact Newton but a Newton-like method.

C Details of parameter estimation

C.1 Initial set-up

In this section, we provide the details of estimating the parameter vector θ in the model (12) given the noisy wavelet coefficients $y(s)$ in equation (9). We approximate the conditional density $p(\mathbf{x}|\mathbf{y};\theta)$ by expanding the negative log conditional density in a Taylor series about the MAP estimate $\widehat{\mathbf{x}}$:

$$f(\mathbf{x};\theta) \quad \approx \quad f(\widehat{\mathbf{x}};\theta) + \frac{1}{2}(\mathbf{x}-\widehat{\mathbf{x}})'\left\{P_{\mathbf{x}}^{-1}(\theta) + D(\widehat{\mathbf{x}})\right\}(\mathbf{x}-\widehat{\mathbf{x}}) \tag{15}$$

where we have used the fact that $\nabla f(\widehat{\mathbf{x}};\theta) = 0$ by definition of the MAP estimate. Here the matrix $D(\widehat{\mathbf{x}})$ is the one that appeared earlier in the Hessian of f. This Taylor series expansion yields the approximation $p(\mathbf{x}|\mathbf{y};\theta) \approx q(\mathbf{x}|\mathbf{y};\theta) \triangleq \mathcal{N}(\widehat{\mathbf{x}}, \mathcal{C}(\widehat{\mathbf{x}};\theta))$ where the covariance is given by $\mathcal{C}(\widehat{\mathbf{x}};\theta) \triangleq \left\{P_{\mathbf{x}}^{-1}(\theta) + D(\widehat{\mathbf{x}})\right\}^{-1}$. At iteration n, we use the approximating density $q(\mathbf{x}|\mathbf{y};\theta^{n-1})$ to perform approximate E-step by calculating the expectation of the augmented log likelihood $L(\theta;\theta^n) \triangleq \mathbb{E}_{q(\mathbf{x}|\mathbf{y};\theta^n)}\Big[\log p(\mathbf{x},\mathbf{y};\theta)\Big]$. It is straightforward to show [25] that this function is a lower bound on the log likelihood $p(\mathbf{y};\theta)$. Like many generalized EM methods, instead of performing an exact maximization of L at the M-step, we will simply take a gradient step. This generates a series of parameter estimates $\{\theta^n\}$ via the recursion $\theta^n = \theta^{n-1} + \beta^n S(\theta^{n-1};\theta^{n-1})\nabla L(\theta^{n-1};\theta^{n-1})$ where S is the Hessian of L (or some approximation to it); and β^n is a step size parameter.

To perform these updates, we need to calculate the gradient ∇L. The i^{th} element of this gradient is given by $\frac{\partial L}{\partial \theta_i} = \mathbb{E}_q\left[\frac{\partial}{\partial \theta_i}\Big(\log p(\mathbf{x};\theta)\Big)\right]$, where we have used the dominated convergence theorem to interchange expectation and differentiation, and the fact that $\log p(\mathbf{y}|\mathbf{x};\theta)$ does not depend on θ. Recall that for a Gaussian process $\mathbf{x} \sim \mathcal{N}(0, P_{\mathbf{x}})$, we have $-\log p(\mathbf{x};\theta) = \frac{N}{2}\log(2\pi) + \frac{1}{2}\log\det P(\theta) + \frac{1}{2}\mathbf{x}^T P^{-1}(\theta)\mathbf{x}$, where we write $P \equiv P_{\mathbf{x}}$ for simplicity in notation. The partial derivative with respect to θ_i is given by $-\frac{\partial}{\partial \theta_i}\big[\log p(\mathbf{x};\theta)\big] = \frac{1}{2}\operatorname{trace}\left(P^{-1}\frac{\partial P}{\partial \theta_i}\right) - \frac{1}{2}\mathbf{x}^T P^{-1}\frac{\partial P}{\partial \theta_i}P^{-1}\mathbf{x}$. We calculate the i^{th} element of the gradient

∇L by taking the expectation of of $-\frac{\partial}{\partial\theta_i}\big[\log p(\mathbf{x}|\theta)\big]$ with respect to this approximating normal density $q(\mathbf{x};\theta^n) \equiv \mathcal{N}(\widehat{\mathbf{x}}, \mathcal{C}(\mathbf{x};\theta^n))$, where the covariance $\mathcal{C}$ was defined earlier. Following some elementary calculations, we obtain

$$\frac{\partial L}{\partial \theta_i} = \frac{1}{2}\,\mathrm{trace}\left[P^{-1}\frac{\partial P}{\partial \theta_i}\right] - \frac{1}{2}\widehat{\mathbf{x}}' P^{-1}\frac{\partial P}{\partial \theta_i}P^{-1}\widehat{\mathbf{x}} - \frac{1}{2}\,\mathrm{trace}\left[\mathcal{C}^T P^{-1}\frac{\partial P}{\partial \theta_i}P^{-1}\right] \tag{16}$$

C.2 Gradient evaluation via likelihood calculations

Although equation (16) is analytically straightforward, its actual computation is non-trivial. Recall that the matrices P and $\mathcal{C}$, as well as their inverses and derivatives, are all $N \times N$, where N is very large (say 10^5). This large dimension renders infeasible any brute force approach. However, the tree structure can be exploited to develop a very fast algorithm for likelihood calculation of MAR models [see 59], consisting of a single upward sweep from leaves to root.

This algorithm for computing MAR likelihoods turns out to be useful here. By applying the matrix inversion lemma to equation (16) and simplifying, we have $\frac{\partial L}{\partial \theta_i} = -\frac{1}{2}\widehat{\mathbf{x}}' P^{-1}\frac{\partial P}{\partial \theta_i}P^{-1}\widehat{\mathbf{x}} + \frac{1}{2}\,\mathrm{trace}\left[\left(P + D^{-1}\right)^{-1}\frac{\partial P}{\partial \theta_i}\right]$. For any covariance matrix Γ, let $J(\mathbf{u};\Gamma) \triangleq \frac{1}{2}\,\mathrm{trace}\,\log(\Gamma) + \frac{1}{2}\mathbf{u}^T\Gamma^{-1}\mathbf{u}$ be the corresponding Gaussian likelihood. With this definition, it can be shown be shown that $\frac{\partial L}{\partial \theta_i}(\widehat{\mathbf{x}}) = \frac{\partial J}{\partial \theta_i}(P;\widehat{\mathbf{x}}) - \frac{\partial J}{\partial \theta_i}(P;\mathbf{0}) + \frac{\partial J}{\partial \theta_i}(P + D^{-1};\mathbf{0})$. Thus, the gradient computation can be performed by taking derivatives of standard Gaussian likelihoods on the tree. Similarly, this structure permits efficient computation of elements of the Hessian.

References

[1] D.F. Andrews and C.L. Mallows. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society*, 36:99–102, 1974.

[2] D.P. Bertsekas. *Nonlinear programming.* Athena Scientific, Belmont, MA, 1995.

[3] R. W. Buccigrossi and E. P. Simoncelli. Image compression via joint statistical characterization in the wavelet domain. *IEEE Trans Image Proc*, 8(12):1688–1701, December 1999.

[4] A. Chambolle, R. A. DeVore, N. Lee, and B. J. Lucier. Nonlinear wavelet image processing: Variational problems, compression, and noise removal through wavelet shrinkage. *IEEE Trans. Image Proc.*, 7:319–335, March 1998.

[5] S. G. Chang, B. Yu, and M. Vetterli. Spatially adaptive wavelet thresholding with context modeling for image denoising. In *Proc. IEEE ICIP*, pages 535–539, October 1998.

[6] W. H. Chen and C. H. Smith. Adaptive coding of monochrome and color images. *IEEE Transactions on Communications*, COM-25:1285–1292, November 1977.

[7] H. Cheng and C. A. Bouman. Trainable context model for multiscale segmentation. In *Proc. IEEE ICIP*, volume 1, pages 610–614, 1998.

[8] K. Chou, A. Willsky, and R. Nikoukhah. Multiscale systems, Kalman filters, and Riccati equations. *IEEE Trans. AC*, 39(3):479–492, March 1994.

[9] C. Chrysafis and A. Ortega. Efficient context based entropy coding for lossy wavelet image compression. In *Proc. Data Compression Conference*, pages 241–250, 1997.

[10] M.S. Crouse, R.D. Nowak, and R.G. Baraniuk. Wavelet-based statistical signal processing using hidden Markov models. *IEEE Trans. on Signal Processing*, 46:886–902, April 1998.

[11] M. Daniel and A. Willsky. Modeling and estimation of fractional Brownian motion using multiresolution stochastic processes. In J. L. Vehel, E. Lutton, and C. Tricot, editors, *Fractals in Engineering*, pages 124–137. Springer, 1997.

[12] M. Daniel and A. Willsky. The modeling and estimation of statistically self-similar processes in a multiresolution framework. *IEEE Transactions on Information Theory*, 45(3):955–970, April 1999.

[13] J.W. Demmel. *Applied numerical linear algebra*. SIAM, Philadelphia, 1997.

[14] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Royal Stat. Soc. B*, 39:1–38, 1977.

[15] D. L. Donoho and I.M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90(432):1200–1224, December 1995.

[16] B. Efron and C. N. Morris. Limiting the risk of Bayes and empirical Bayes estimators – part II: The empirical Bayes case. *Journal of the American Statistical Association*, 67:130–139, 1972.

[17] W. Feller. *An Introduction to Probability Theory and its Applications: Volume II*. John Wiley and Sons, New York, 1966.

[18] P. Fieguth and A. Willsky. Fractal estimation using models on multiscale trees. *IEEE Transactions on Signal Processing*, 44(5):1297–1300, May 1996.

[19] D. J. Field. Relations between the statistics of natural images and the response properties of cortical cells. *J. Opt. Soc. Am. A*, 4(12):2379–2394, 1987.

[20] C.W. Harrison. Experiments with linear prediction in television. *Bell Syst. Tech. Journal*, 31:764–783, 1952.

[21] J. Huang and D. Mumford. Statistics of natural images and models. In *CVPR*, 1999.

[22] W. Irving, P. Fieguth, and A. Willsky. An overlapping tree approach to multiscale stochastic modeling and estimation. *IEEE Transactions on Image Processing*, 6(11), November 1997.

[23] W. James and C. Stein. Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 361–379, 1961.

[24] A. H. Jazwinski. *Stochastic processes and filtering theory*. Academic Press, New York, 1970.

[25] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul. *Learning in graphical models*, chapter An introduction to variational methods for graphical models, pages 105–161. MIT Press, 1999.

[26] G. G. Judge and M. E. Bock. *The statistical implications of pre-test and Stein-rule estimators in econometrics*. North-Holland, Amsterdam, 1978.

[27] A. Lee and D. Mumford. An occlusion model generating scale-invariant images. In *CVPR*, 1999.

[28] J S Lee. Digital image enhancement and noise filtering by use of local statistics. *IEEE Pat. Anal. Mach. Intell.*, PAMI-2:165–168, March 1980.

[29] J. Liu and P. Moulin. Image denoising based on scale-space mixture modeling of wavelet coefficients. In *Proc. IEEE ICIP*, volume 1, pages 386–390, October 1999.

[30] S. LoPresto, K. Ramchandran, and M. Orchard. Image coding based on mixture modeling of wavelet coefficents and a fast estimation-quantization framework. In *Proc. Data Compression Conf.*, March 1997.

[31] M. Luettgen. *Image processing with multiscale stochastic models*. PhD thesis, Massachusetts Institute of Technology, May 1993.

[32] M. Luettgen, W. Karl, and A. Willsky. Efficient multiscale regularization with application to optical flow. *IEEE Trans. Im. Proc.*, 3(1):41–64, Jan. 1994.

[33] M. Luettgen, W. Karl, A. Willsky, and R. Tenney. Multiscale representations of Markov random fields. *IEEE Transactions on Signal Processing*, 41(12):3377–3396, December 1993.

[34] S. G. Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Pat. Anal. Mach. Intell*, 11:674–693, July 1989.

[35] M. Mihçak, I. Kozintsev, and K. Ramchandran. Spatially adaptive statistical modeling of wavelet image coefficients, and its application to denoising. In *Proc. IEEE ICASSP*, pages 3253–3256, March 1999.

[36] M. Mihçak, I. Kozintsev, K. Ramchandran, and P. Moulin. Low-complexity image denoising based on statistical modeling of wavelet coefficients. *IEEE Sig. Proc. Let.*, 6:300–303, December 1999.

[37] P. Moulin and J. Liu. Analysis of multiresolution image denoising schemes using a generalized Gaussian and complexity priors. *IEEE Trans. Info. Theory*, 45:909–919, April 1999.

[38] D. Mumford and B. Gidas. Stochastic models for generic images. Preprint available at http://www.dam.brown.edu/people/mumford.

[39] B.M. Oliver. Efficient coding. *Bell Syst. Tech. Journal*, 31:724–750, 1952.

[40] A. Pentland. Fractal based description of natural scenes. *IEEE Trans. PAMI*, 6(6):661–674, 1984.

[41] J Portilla and E P Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *Int'l Journal of Computer Vision*, 40(1):49–70, October 2000.

[42] H. Rauch, F. Tung, and C. Striebel. Maximum likelihood estimates of linear dynamic systems. *AIAA Journal*, 3(8):1445–1450, August 1965.

[43] J. Romberg, H. Choi, and R. Baraniuk. Bayesian wavelet domain image modeling using hidden Markov trees. In *Proc. IEEE ICIP*, Kobe, Japan, October 1999.

[44] D. L. Ruderman. Origins of scaling in natural images. *Vision Research*, 37:3385–3395, 1997.

[45] D. L. Ruderman and W. Bialek. Statistics of natural images: Scaling in the woods. *Phys. Rev. Letters*, 73(6):814–817, 1994.

[46] G. Samorodnitsky and M. Taqqu. *Stable non-Gaussian random processes: stochastic models with infinite variance*. Chapman and Hall, New York, 1994.

[47] J.M. Shapiro. Embedded image coding using zerotrees of wavelet coefficients. *IEEE Transactions on Signal Processing*, 41:3445–3462, December 1993.

[48] E. P. Simoncelli. Statistical models for images: Compression, restoration and synthesis. In *31st Asilomar Conf.*, pages 673–678. IEEE Sig. Proc. Soc., November 1997.

[49] E. P. Simoncelli. Bayesian denoising of visual images in the wavelet domain. In P. Müller and B. Vidakovic, editors, *Bayesian Inference in Wavelet Based Models*, chapter 18, pages 291–308. Springer-Verlag, New York, June 1999. Lecture Notes in Statistics, vol. 141.

[50] E. P. Simoncelli and E. H. Adelson. Noise removal via Bayesian wavelet coring. In *Third Int'l Conf on Image Proc*, volume I, pages 379–382, September 1996.

[51] E. P. Simoncelli and W. T. Freeman. The steerable pyramid: A flexible architecture for multi-scale derivative computation. In *Int'l Conf on Image Proc*, volume III, pages 444–447, Washington, DC, October 1995. IEEE Sig Proc Soc.

[52] V. Strela, J. Portilla, and E. Simoncelli. Image denoising using a local Gaussian scale mixture model in the wavelet domain. In *Proceedings of SPIE*, San Diego, CA, July 2000.

[53] A. H. Tewfik and M. Kim. Correlation structure of the discrete wavelet coefficients of fractional Brownian motion. *IEEE Trans. Info. Theory*, 38:904–909, March 1992.

[54] A. Turiel, G. Mato, N. Parga, and J. P. Nadal. The self-similarity properties of natural images resemble those of turbulent flows. *Phys. Rev. Lett.*, 80:1098–1101, 1998.

[55] M. J. Wainwright and E. P. Simoncelli. Scale mixtures of Gaussians and the statistics of natural images. In *Neural Information Processing Systems 12*, volume 12, pages 855–861, December 1999. Paper available at http://ssg.mit.edu/group/mjwain/mjwain.shtml,.

[56] M. J. Wainwright, E. P. Simoncelli, and A. S. Willsky. Random cascades of Gaussian scale mixtures and their use in modeling natural images with application to denoising. In *IEEE Int. Conf. Image Proc.*, Vancouver,Canada, September 2000.

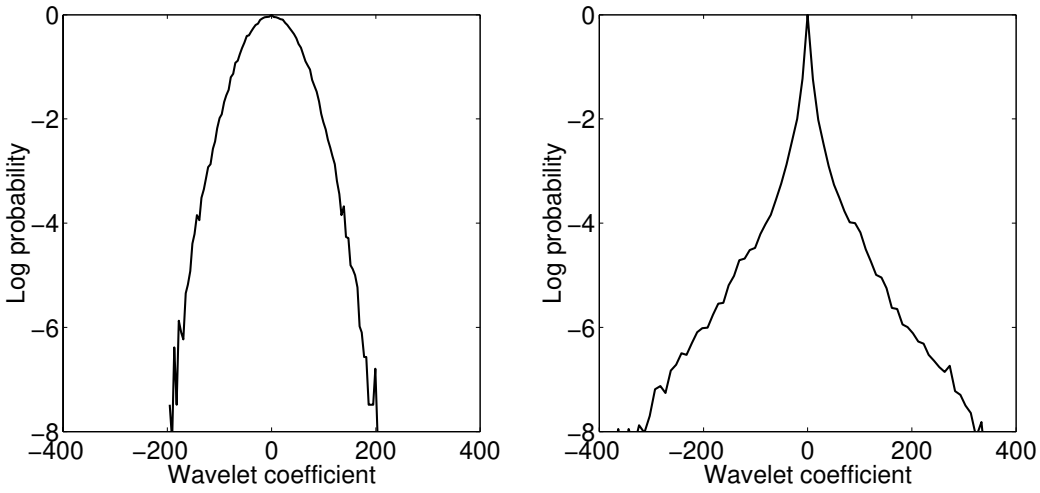
[57] M. J. Wainwright, E. P. Simoncelli, and A. S. Willsky. Random cascades of Gaussian scale mixtures on wavelet trees with application to natural images. In *Proceedings of SPIE*, San Diego, CA, July 2000. Paper available at http://ssg.mit.edu/group/mjwain/mjwain.shtml.

[58] M. J. Wainwright, E. B. Sudderth, and A. S. Willsky. Tree-based modeling and estimation of Gaussian processes on graphs with cycles. In *Neural Information Processing Systems 13*, December 2000. Paper available at http://ssg.mit.edu/group/mjwain/mjwain.shtml.

[59] R.B. Washburn, W.I. Irving, J.K. Johnson, D.S. Artgis, J.W. Wissinger, R.P. Tenney, and A.S. Willsky. Multiresolution image compression and image fusion algorithms. Technical report, Alphatech Company, February 1996.

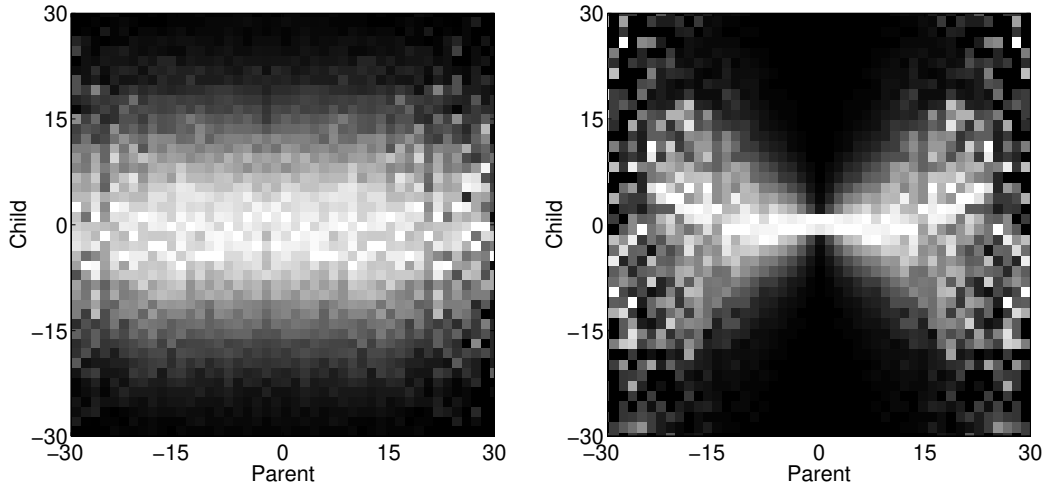
[60] G. Wornell. Wavelet-based representations for the $1/f$ family of fractal processes. *Proc. IEEE*, September 1993. 93-01.

[61] C. Zetzsche, B. Wegmann, and E. Barth. Nonlinear aspects of primary vision: Entropy reduction beyond decorrelation. In *Int'l Symp. Soc. for Info. Display*, volume 24, pages 933–936, 1993.



(a) Gaussian (b) Natural image

Figure 1. Histograms of wavelet marginal distributions for (a) Gaussian noise; and (b) a typical natural image. Vertical axis gives log probability (rescaled).



(a) Gaussian white noise (b) Natural image

Figure 2. Joint conditional histograms for a wavelet coefficient (parent) and its coarser scale child taken from Gaussian white noise (a), contrasted with a natural image (b). Each column of the 2-D plots corresponds to 1-D conditional histogram of $p(\text{child}\,|\,\text{parent})$. Lightness corresponds to frequency of occurrence,where each column has been independently rescaled to form a conditional histogram.

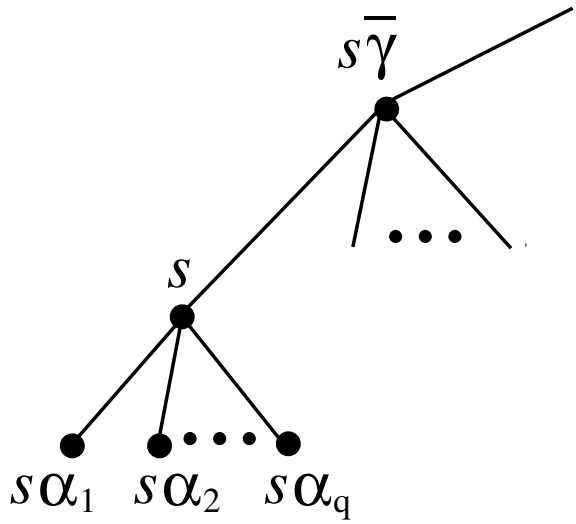


Figure 3. A segment of a q-adic tree, with the unique parent $s\bar{\gamma}$ and children $s\alpha_q, \ldots s\alpha_q$ corresponding to node s.

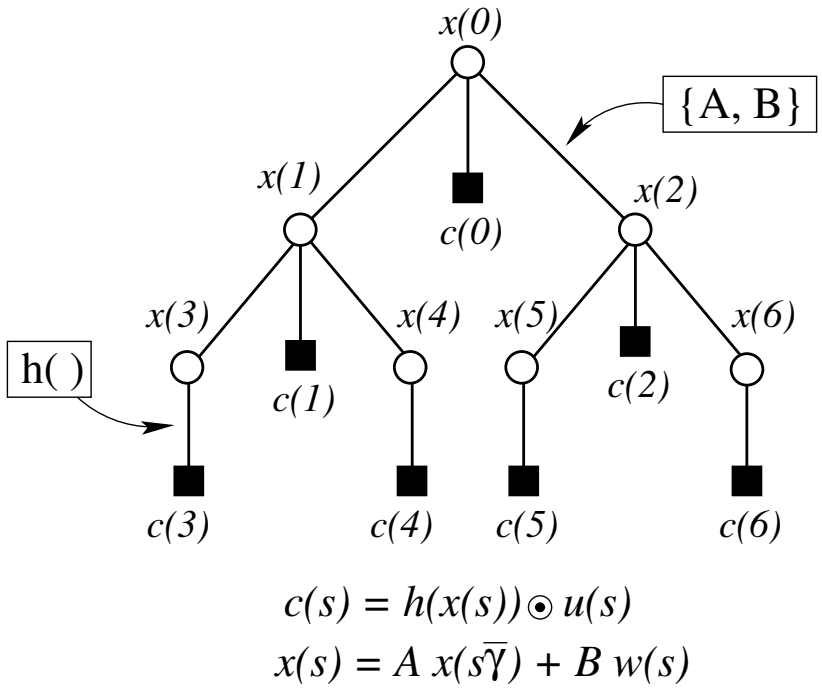


Figure 4. Graphical illustration of model structure. Premultiplier process $x(s)$ and white noise $u(s)$ are defined on nodes (represented by ◯) of the multiscale tree. Wavelet coefficient vectors $c(s)$ (represented by ■) is generated via nonlinearity h.

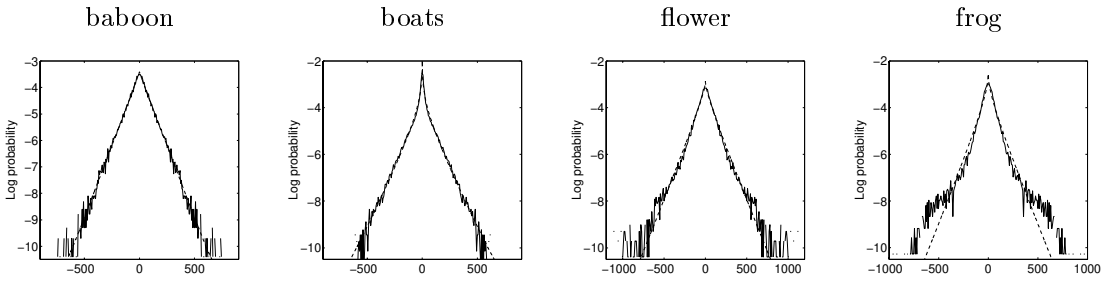


Figure 5. Log histograms of GSM model fits (dotted line) to the log empirical histograms of steerable pyramid coefficients (a single subband) applied to natural images. Parameters are computed by numerical minimization of the Kullback-Leibler divergence.

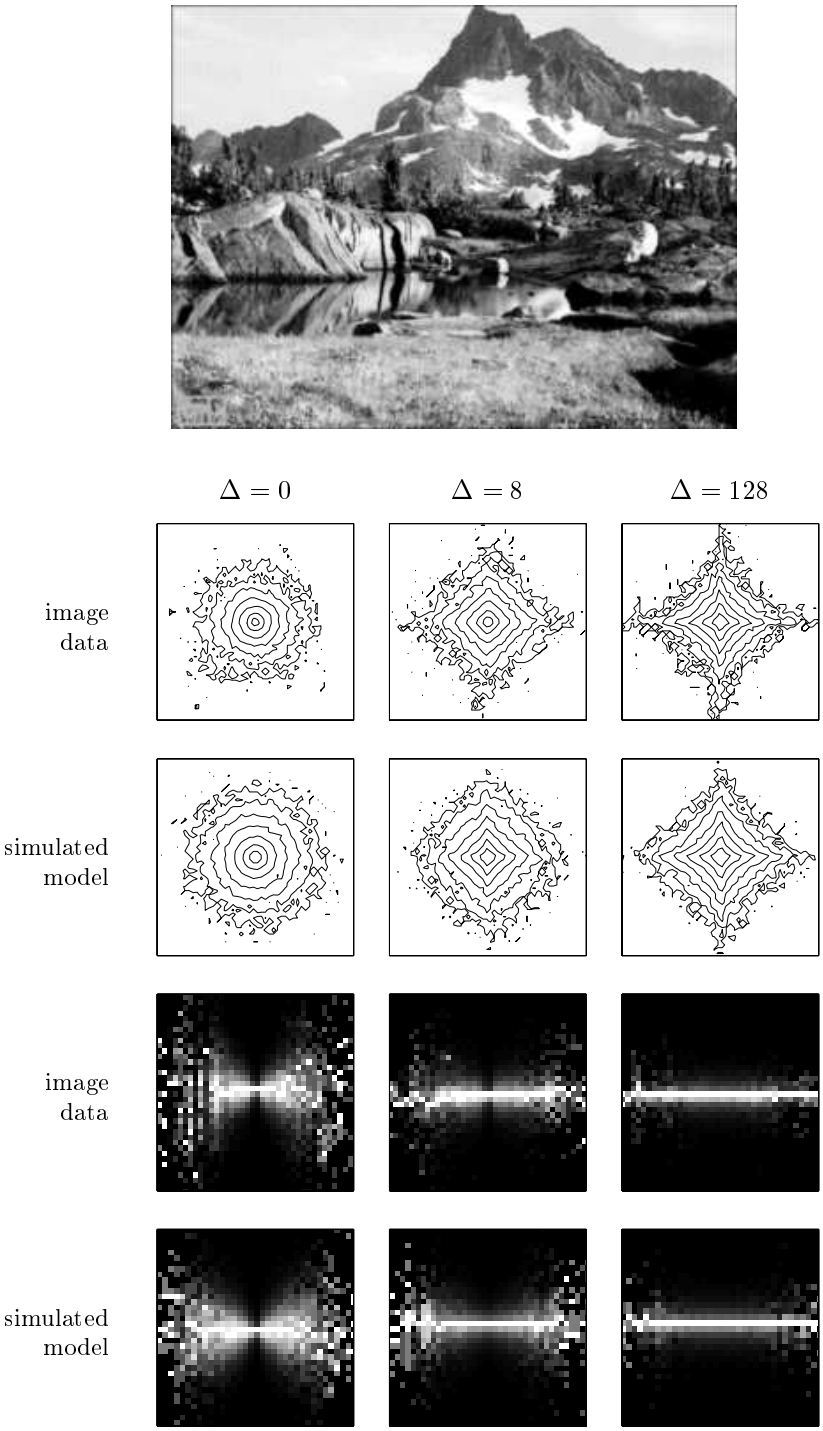
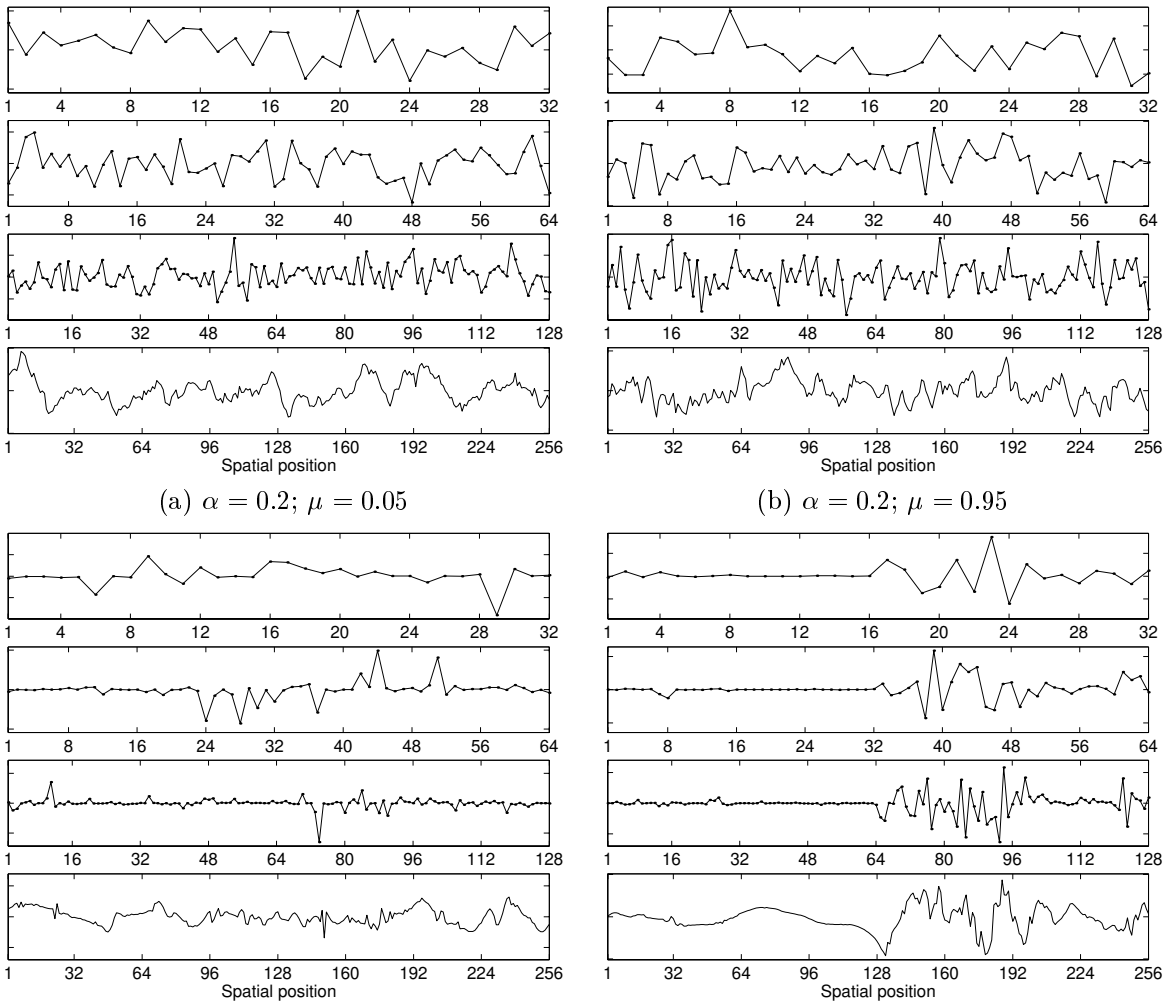


Figure 6. Examples of empirically observed distributions of wavelet coefficients, compared with simulated distributions from the GSM gamma model. Top row: "Mountain" image. Second row: Empirical joint histograms for the "mountain" image, for three pairs of wavelet coefficients, corresponding to basis functions with spatial separations $\Delta = \{0, 8, 128\}$. Third row: Simulated joint distributions for $\mu = 0.92$, $h(x) = |x|$, and the same spatial separations. Contour lines are drawn at equal intervals of log probability. Fourth row: Empirical conditional histograms for the "mountain" image. Fifth row: Simulated conditional histograms for the GSM cascade. For these conditional distributions, intensity corresponds to probability, except that each column has been independently rescaled to fill the full range of intensities.



(a) $\alpha = 0.2$; $\mu = 0.05$

(b) $\alpha = 0.2$; $\mu = 0.95$

(c) $\alpha = 2$; $\mu = 0.05$

(d) $\alpha = 2$; $\mu = 0.95$

Figure 7. Simulated random cascades for various choices of the parameters. Heaviness of tails (and hence impulsiveness of the process) increases with the parameter α, whereas the parameter μ controls the scale-to-scale dependence.

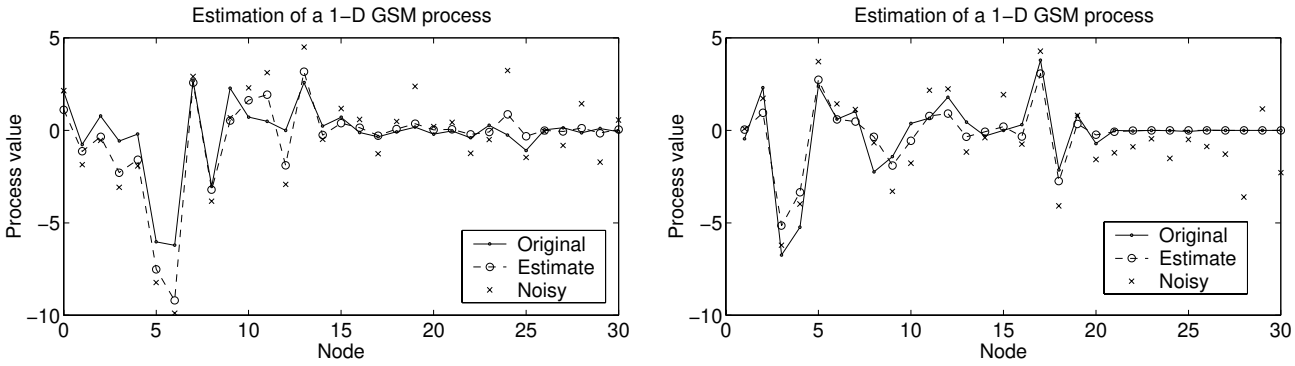
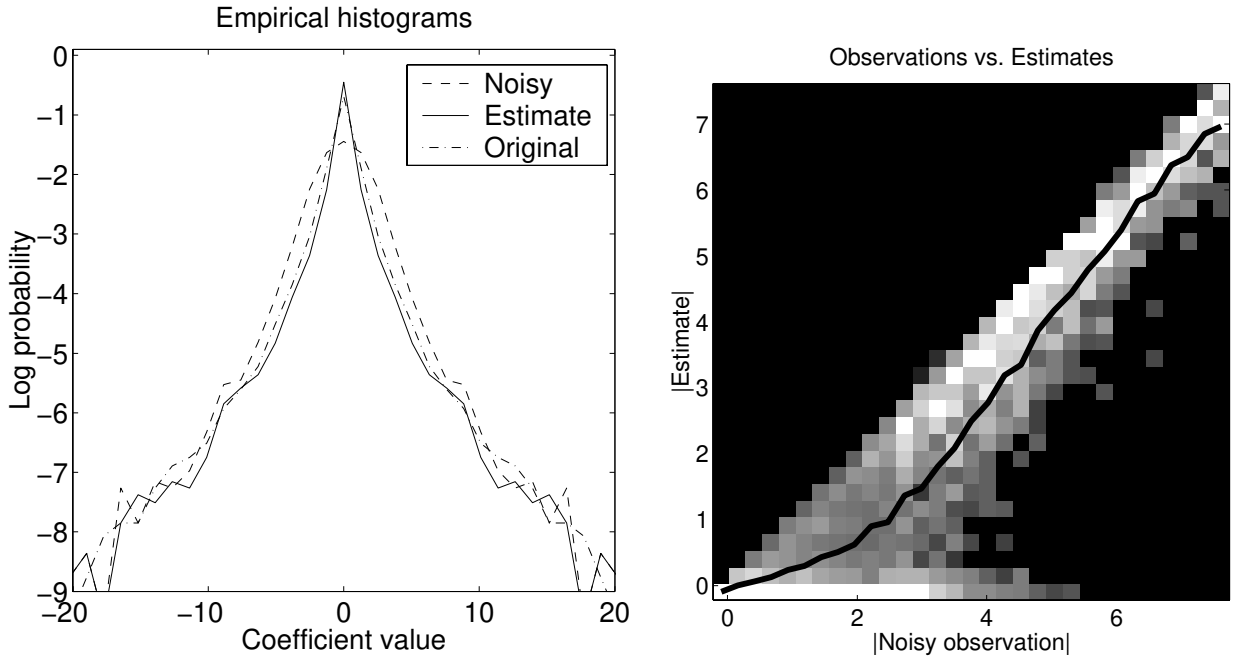


Figure 8. Estimation of a 1-D GSM processes based on observations contaminated by white Gaussian noise. (a) GSM process generated with $h(x) \triangleq \exp(1.5x)$. (b) GSM generated with $h(x) \triangleq [x^+]^3$.



(a) (b)

Figure 9. (a) Empirical histograms of original wavelet coefficients, estimated coefficients, and noisy observations, all plotted on a semilog scale. (b) Joint histogram of absolute value of noisy observations $y(s) = c(s) + v(s)$ versus absolute value of estimates $\widehat{c}(s)$. The overlaid solid line is the conditional mean $\mathbb{E}\big[|\widehat{c}(s)| \;\big|\; |y(s)|\big]$.

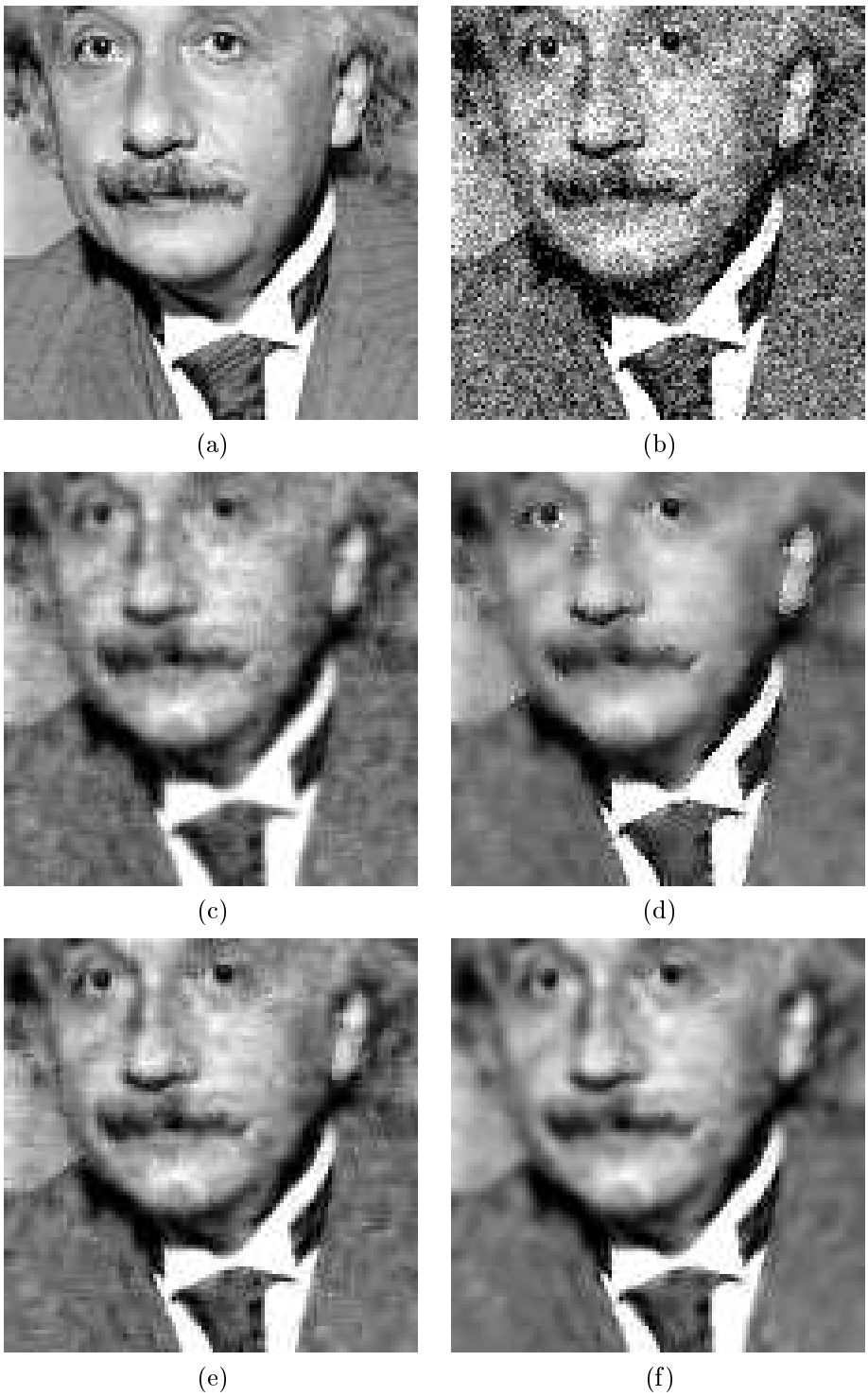


Figure 10. Cropped denoising results using a 4-orientation steerable pyramid. (a) Original image. (b) Noisy image. (c) Wiener subband denoising. (d) MATLAB adaptive. (e) Soft thresholding. (f) GSM-tree algorithm.

Mixing density	GSM density	GSM char. function
$\lambda Z(\alpha)$	symmetrized Gamma	$[1+\lambda^2 t^2]^{-\alpha}, \quad \alpha > 0$
$\lambda / Z(\alpha - \frac{1}{2})$	Student: $[1+t^2/\lambda^2]^{-\alpha} \quad \alpha > \frac{1}{2}$	No explicit form
Positive $\frac{\alpha}{2}$-stable	α-stable	$\exp(-\lvert\lambda t\rvert^{\alpha}), \quad \alpha \in (0,2]$
No explicit form	generalized Gaussian: $\exp(-\lvert\frac{c}{\lambda}\rvert^{\alpha}), \quad \alpha \in (0,2]$	No explicit form
$z \stackrel{d}{=} \lambda \exp(x/\alpha)$ $\alpha \geq 0$	Log multiplier No explicit form	No explicit form

Table 1. Example densities from the class of Gaussian scale mixtures. The notation $Z(\gamma)$ denotes a positive gamma variable z of index γ with density $p(z) = \frac{z^{\gamma-1}}{\Gamma(\gamma)} \exp(-z)$.

Noisy	Wiener subband	*wiener2.m*	Soft threshold	GSM Tree
1.59	9.28	10.19	10.11	10.54
4.80	10.61	11.86	11.47	12.31
9.02	12.58	13.37	13.24	14.68
13.06	14.96	14.23	15.41	16.83

Table 2. Denoising results (SNR in dB) for 256×256 Einstein image using a 4-orientation steerable pyramid. The original noisy SNR is given by $10\log_{10}[\mathrm{var}(\mathcal{I})/\sigma^2]$, and the cleaned SNR is given by $10\log_{10}[\mathrm{var}(\mathcal{I})/\mathrm{var}(\widehat{\mathcal{I}}-\mathcal{I})]$, where $\mathcal{I}$ and $\widehat{\mathcal{I}}$ denote the original and denoised images respectively.