Published in: Proceedings of the 7th International Conference on Image Processing, Vancouver, BC, Canada. 10-13 September 2000.

Random cascades of Gaussian scale mixtures and their use in modeling natural images with application to denoising

Martin J. Wainwright[1] Eero P. Simoncelli[2] Alan S. Willsky[1]

[1]Laboratory for Information & Decision Systems
Electrical Engineering & Computer Science
Massachusetts Institute of Technology
{*mjwain,willsky*}*@mit.edu*

[2]Center for Neural Science, and
Courant Institute of Mathematical Sciences
New York University
eero@cns.nyu.edu

Multiresolution representations play an important role in image processing and computer vision, as well as in modeling stochastic processes [e.g., 1, 2]. Our work lies in the intersection of these areas: we seek to develop and study a new class of multiscale stochastic processes that are capable of capturing the statistics of natural images. Such models play a central role in a variety of applications in image processing, such as compression, denoising, and enhancement.

Our work is motivated by empirical observations of natural image statistics when represented in multiscale bases. It is now well-known that wavelet marginal distributions are highly non-Gaussian, with heavy tails and high kurtosis [e.g., 3]. Moreover, in agreement with theoretical analysis of $1/f$ processes [e.g., 4], the detail coefficients of orthonormal wavelets applied to natural images tend to be approximately uncorrelated. Despite this approximate decorrelation, they are by no means independent. Indeed, they exhibit a strong 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 [5, 6]. Similar behaviors have been observed by authors modeling turbulent fluid flow [e.g., 7].

We have developed a class of non-Gaussian multiscale processes, defined by random coarse-to-fine cascades on trees of multiresolution coefficients, that exhibit precisely these types of behavior. Our cascade models represent a significant advance over linear models defined on multiscale trees [1]. Although linear models lead to exceptionally efficient algorithms for image processing, they cannot capture the significant types of non-Gaussianity and nonlinear cascade behavior present in wavelet coefficients of natural images. To capture such behavior, we define cascades that reproduce a rich semi-parametric class of random variables known as Gaussian scale mixtures (GSM). We demonstrate that this model class not only captures natural image statistics, but also facilitates efficient and optimal processing, which we illustrate by application to image denoising. More details of the work reported here can be found in [8, 9].

MW is supported by NSERC 1967 fellowship 16033; AW and MW by AFOSR grant F49620-98-1-0349 and ONR grant N00014-91-J-1004; ES by NSF CAREER grant MIP-9796040.

1. GAUSSIAN SCALE MIXTURES

A GSM vector $\mathbf{c}$ of length m has the representation as a product of two independent random variables $\mathbf{c} \stackrel{d}{=} \sqrt{z}\,\mathbf{u}$, where $\stackrel{d}{=}$ indicates equality in distribution. The positive scalar random variable z is the *mixing variable* with density p_z, whereas $\mathbf{u} \sim \mathcal{N}(0, Q)$ is a Gaussian random vector. Samples of $\mathbf{c}$ are obtained by sampling a Gaussian random vector $\mathbf{u}$ and multiplying each component by the mixing coefficient drawn independently from density p_z. As a consequence, any GSM vector has a density given by an integral:

$$p_{\mathbf{c}}(\mathbf{c}) = \int_0^\infty \frac{1}{(2\pi)^{m/2}\,|zQ|^{1/2}} \exp\left(-\frac{\mathbf{c}^T Q^{-1}\mathbf{c}}{2z}\right) p_z(z)dz.$$

Finite mixtures of Gaussians correspond to the special case where p_z is a discrete probability mass function. A number of well-known heavy-tailed distributions belong to the GSM class, including the generalized Gaussian (or stretched exponential) family, the α-stable family, and the Student t-variables. In previous work [8], we have shown that GSM vectors can account well for the non-Gaussian properties of natural images, including marginal (Fig. 1) and pairwise joint distributions (Fig. 2).

2. RANDOM CASCADES ON WAVELET TREES

In order to build global probability distribution on the space of images consistent with these local descriptions, it is natural to make use of a probabilistic graphical model. The simplest choice is the quad tree associated with the wavelet transform. Let $c(s)$ represent a d-vector of wavelet coefficients at node s, corresponding to the same spatial scale and position but different orientations. The multiplier variables for a large class of GSMs can be generated by passing a Gaussian random d-vector $x(s)$ through a nonlinearity $h : \mathbb{R}^d \to \mathbb{R}^d$. Thus, the GSM coefficient can be written as $c(s) = h\big(x(s)\big) \odot u(s)$, where $h\big(x(s)\big) \equiv \sqrt{z(s)}$ corresponds to the mixing variable; $u(s) \sim \mathcal{N}(0, Q(s))$ is a Gaussian random vector of length d; and $\odot$ denotes entry-wise multiplication.

We define coarse-to-fine stochastic dynamics on the state

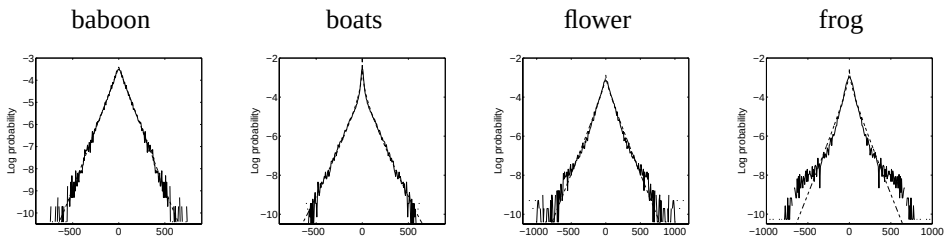

Fig. 1. Log histograms of GSM model fits (dotted line) to the log empirical histograms of steerable pyramid coefficients (a single subband) applied to natural images. Parameters are computed by numerical minimization of the Kullback-Leibler divergence.

variables $x(s)$ and $u(s)$:

$$\begin{aligned} x(s) &= A(s)x(s\bar{\gamma}) + B(s)w(s) \\ u(s) &= C(s)u(s\bar{\gamma}) + D(s)\zeta(s), \end{aligned}$$

where $s\bar{\gamma}$ denotes the parent node of s. Here $x(s)$ and $u(s)$ are d-vectors, and w and ζ are $\mathcal{N}(0, I)$ process noises of length d that are mutually uncorrelated and white. Observations of the vector $c(s)$ of wavelet coefficients are given by:

$$y(s) = h\big(x(s)\big) \odot u(s) + v(s),$$

where $v(s) \sim \mathcal{N}(0, R(s))$ is white observation noise. Note that while the state dynamics correspond to a multiscale autogressive (MAR) process [see 1], the nonlinear observation equation produces a GSM random vector at each node.

The tree structure imposes a powerful Markov property: Any two vectors of wavelet coefficients $c(s)$ and $c(t)$ are conditionally independent given their common state ancestors $x(s \wedge t)$ and $u(s \wedge t)$. Given this structure, it is straightforward to show that the joint distributions of any pair of wavelet vectors $c(s)$ and $c(t)$ are given by:

$$\begin{aligned} c(s) &\stackrel{d}{=} h\Big[\mathcal{A}(s,t)x(s \wedge t) + \nu_1(s)\Big] \odot u(s) \\ c(t) &\stackrel{d}{=} h\Big[\mathcal{A}(t,s)x(s \wedge t) + \nu_2(t)\Big] \odot u(t) \end{aligned}$$

where ν_1 and ν_2 are uncorrelated white noises, and

$$\mathcal{A}(s,t) \triangleq \prod_{r=s}^{s \wedge t} A(r). \tag{1}$$

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 left three panels of Fig. 2). Others proposed modeling these joint contours with a 2D generalized Gaussian [10]. 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 tree with $A(s) \equiv \gamma$ and $B(s) \equiv \sqrt{1-\gamma^2}$; with u white in scale (typical for natural images); and $h(x) \triangleq \|x\|$. For nodes s and t at the same scale and orientation but spatially separated by distance Δ, equation (1) yields $\mathcal{A}(s,t) \propto \gamma^{[\log 2(\Delta)+1]}$, which allows us to predict the form of any pairwise joint distributions. The top row of Fig. 2 shows the empirical behavior of steerable pyramid wavelet coefficients applied to natural images, as compared to coefficients of the simulated random cascade in the bottom row. The shapes of the joint contours of image data and simulated model (left three panels) are strikingly similar; the joint conditional histograms (right three panels) demonstrate that the relationship between wavelet coefficients ranges from strong dependence of quadrature phase pairs, to near-independence of distant pairs. Thus, a GSM cascade on a tree accounts well for pairwise joint dependencies of coefficients at a range of separations.

3. STATE ESTIMATION

An important problem is the estimation of the state process $x(s)$, which determines the mixing variables $h\big(x(s)\big)$. Here we outline an algorithm for maximum a posteriori (MAP) estimation of the state $x(s)$. Although the algorithm is generally applicable, it will be particularly efficient under the assumption that $u(s)$ is uncorrelated from node-to-node. We begin by forming a Gaussian vector $\mathbf{x}$ by stacking up the vectors $x(s)$ in a fixed order; define a vector of observations $\mathbf{y}$ in a similar fashion. The problem of MAP estimation corresponds to minimizing the negative log likelihood $f(\mathbf{x}) \triangleq -\log p(\mathbf{x}|\mathbf{y})$. We apply Newton's method to the problem. That is, we generate a sequence $\{\mathbf{x}^n\}$ according to the recursion:

$$\mathbf{x}^{n+1} = \mathbf{x}^n + \alpha^n \big[\nabla^2 f(\mathbf{x}^n)\big]^{-1} \nabla f(\mathbf{x}^n)$$

where α^n is a stepsize; ∇f (respectively $\nabla^2 f$) is the gradient, $\nabla^2 f$ is the Hessian of f. Associated with Newton's method is the hurdle that computing the descent direction $d^n = \big[\nabla^2 f(\mathbf{x}^n)\big]^{-1} \nabla f(\mathbf{x}^n)$ may be very costly, especially in application to images, where the dimension of the Hessian will be enormous ($\approx 10^5$).

Clearly, it is crucial to exploit structure inherent in the problem. We have observed that computing the descent di-

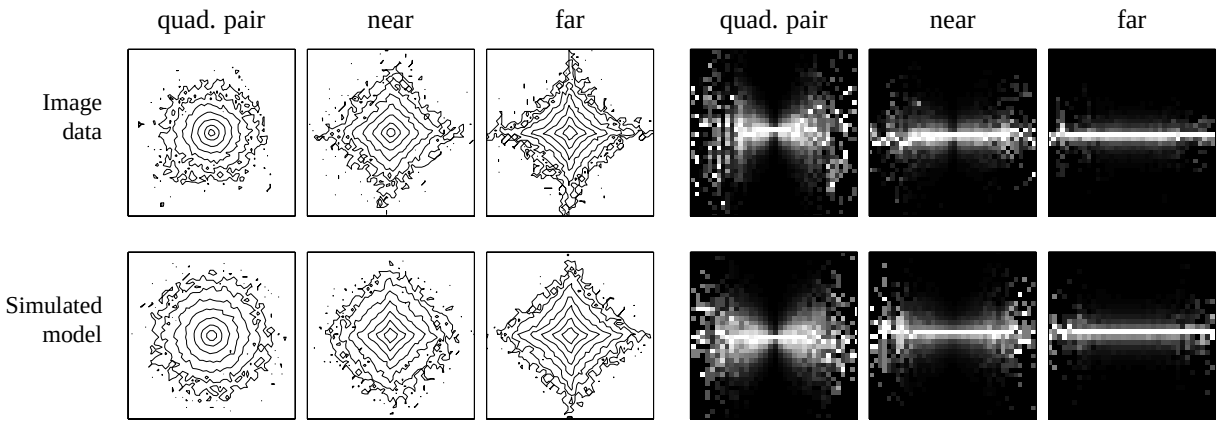

Fig. 2. Left three panels show log contours of joint probability for pairs of wavelet coefficients at same scale and orientation, but varying spatial separation. Right three panels show joint conditional histograms for the same pairs of coefficients. Here each column of the 2D histogram corresponds to a 1D conditional histogram of the wavelet coefficient conditioned on its neighbor. Intensity corresponds to frequency of occurrence, except that each column has been independently rescaled to fill the full range.

rection can be reformulated as an equivalent linear-Gaussian estimation problem. Such problems can be solved very quickly by $\mathcal{O}(d^3 N)$ algorithms [see 1], where d is the dimension of $x(s)$ and N is the total number of nodes in the tree. This leads to a hybrid algorithm, where the quadratic approximation inherent in Newton's method is performed globally on the entire graph, but the local tree structure is exploited in the computation of each descent direction. As a Newton-like method, the resulting algorithm has a number of desirable properties: namely, convergence to a stationary point is guaranteed, and under suitable regularity conditions, the rate of convergence is supergeometric [11]. This estimation algorithm is described in more detail in [9].

Lastly, when conditioned on $\widehat{x}(s)$, the Bayes least square estimate of wavelet coefficients $c(s) = h\big(x(s)\big) \odot u(s)$ from the noisy observations $y(s) = c(s) + v(s)$ is given by the standard formula:

$$\widehat{c}(s) = P_c \big[P_c + R(s)\big]^{-1} y(s)$$

where $R(s)$ is the covariance matrix of the observation noise, and $P_c \equiv P_c(s; \widehat{x}(s))$ is the covariance matrix of the vector $c(s)$ conditioned on the estimate $\widehat{x}(s)$.

A related problem is that of estimating the system matrices (e.g., A, B) involved in the state dynamics. For this purpose, we have developed an approximate EM technique that exploits the state estimation algorithm at each step. We refer the interested reader to [9].

4. RESULTS

Fig. 3 illustrates a typical sample path of a 1D GSM process on a chain (a special case of tree), as well as the behavior of the estimator based on noisy observations (SNR 2.76 dB). Observe that the sample path alternates 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. The SNR of the estimate $\widehat{c}(s)$ shown in Figure 3(a) is 9.71 dB, which is within 0.50 dB of the ideal performance assuming perfect knowledge of the premultiplier $x(s)$ at each node. Note that 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.

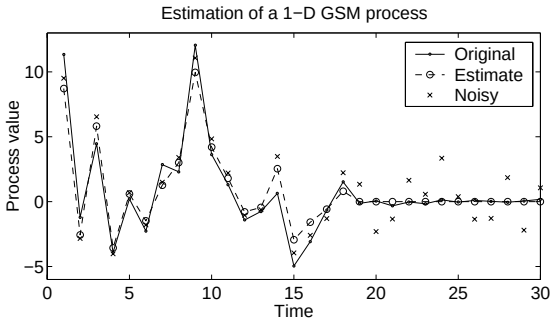

Fig. 3. Estimation of a 1-D GSM process generated with $h(x) \triangleq (x^3)^+$ from measurements contaminated by white Gaussian noise.

We have also applied the algorithm to denoise natural images, using a 4-orientation steerable pyramid [12], and various choices of the nonlinearity. Here we compare the results of our algorithm with $h(x) = (x^5)^+$ to Wiener filtering applied to each subband (linear), to MATLAB's adaptive filtering (*wiener2.m*), as well as to thresholding. Reported in Table 1 are SNR (dB) results for the 256×256 Einstein image for four levels of SNR. Our results are superior in both SNR and visual quality to the other estimators. However, denoising with local estimates of the multiplier [e.g. 13] yields higher SNR results, suggesting that the tree structure

Orig.	Linear	Adapt.	Thres.	GSM Tree
1.59	9.28	8.47	10.12	10.54
4.80	10.61	11.29	11.71	12.31
9.02	12.58	14.23	14.04	14.68
13.06	14.96	15.95	15.89	16.83

Table 1. Denoising results (SNR in dB) for 256×256 Einstein image using a 4-orientation steerable pyramid.

does not fully capture all dependencies.

5. DISCUSSION

In summary, we have developed a semi-parametric class of non-Gaussian multiscale statistical processes defined by random cascades on wavelet trees. This model class is rich enough to accurately capture the remarkably regular non-Gaussian features of natural images, but sufficiently structured to permit estimation of the underlying state variables. We showed that our models accurately fit both the marginal and joint histograms of wavelet coefficients from natural images. We developed a Newton-like method for exact MAP state estimation that exploits fast algorithms for tree estimation, and hence is very efficient. Applications of this algorithm to denoising of both 1D signals and natural images were presented.

The GSM-tree model class is related to a number of previous approaches to image coding and denoising. First, the coefficients generated by GSM cascades exhibit precisely the self-reinforcing property exploited by zerotree encoders [e.g., 14]. Second, GSM models induce a scission of a wavelet coefficient into an unknown multiplier or variance times another random component. Forms of this scission are seen in a number of approaches to image coding and denoising [e.g. 13, 15, 16, 6]. Most approaches have assumed the multiplier to be fixed but unknown, and estimated it in a local and suboptimal fashion. Our GSM framework allows a choice of prior on the multiplier, and the associated algorithm efficiently computes the exact MAP estimate. A close relative of the GSM-tree model class are discrete-state Gaussian mixtures on trees [17]. In contrast to these finite mixtures, we have advocated the use of infinite mixtures of Gaussians. Further discussion of links between GSM models and other work can be found in [9].

A number of extensions to the GSM-tree model presented here are possible. Although we have assumed a fixed parametric form of the nonlinearity, using a nonparametric form would give more flexibility with no loss of efficiency. Secondly, while this work has focused on trees, GSM processes can also be defined on non-tree graphs with additional connections betweeen spatial neighbors. Adding extra connections to the graph will increase modeling power, but also will necessitate different techniques for estimation. Lastly, although the current model assumes the same number of multipliers as coefficients, previous empirical work [8] shows that a smaller set of multipliers suffices to describe a number of wavelet coefficients. Estimating the order of the underlying multiplier process, though a challenging problem, could lead to more powerful models.

6. REFERENCES

[1] K. Chou, A. Willsky, and A. Benveniste, "Multiscale recursive estimation, data fusion, and regularization," *IEEE Trans. AC*, vol. 39, no. 3, pp. 464–478, Mar. 1994.

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

[3] D. J. Field, "Relations between the statistics of natural images and the response properties of cortical cells," *J. Opt. Soc. Am. A*, vol. 4, no. 12, pp. 2379–2394, 1987.

[4] A. H. Tewfik and M. Kim, "Correlation structure of the discrete wavelet coefficients of fractional Brownian motion," *IEEE Trans. IT*, vol. 38, pp. 904–909, Mar. 1992.

[5] E. P. Simoncelli, "Statistical models for images: Compression, restoration and synthesis," in *31st Asilomar Conf.* pp. 673–678, Nov. 1997.

[6] R. W. Buccigrossi and E. P. Simoncelli, "Image compression via joint statistical characterization in the wavelet domain," *IEEE Trans. IP*, vol. 8, no. 12, pp. 1688–1701, Dec. 1999.

[7] 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.*, vol. 80, pp. 1098–1101, 1998.

[8] M. J. Wainwright and E.P. Simoncelli, "Scale mixtures of Gaussians and the statistics of natural images," in *NIPS 12*, May 2000, vol. 12, pp. 855–861, Paper available at http://ssg.mit.edu/group/mjwain/mjwain.shtml.

[9] M. J. Wainwright, E. P. Simoncelli, and A. S. Willsky, "Random cascades on wavelet trees and their use in modeling and analyzing natural images," *Applied Computational and Harmonic Analysis*, 2000, Special issue on wavelets; to appear.

[10] J. Huang and D. Mumford, "Statistics of natural images and models," in *CVPR*, 1999, p. 216.

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

[12] E. P. Simoncelli and W. T. Freeman, "The steerable pyramid: A flexible architecture for multi-scale derivative computation," in *Proc. IEEE ICIP* , Oct. 1995.

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

[14] J.M. Shapiro, "Embedded image coding using zerotrees of wavelet coefficients," *IEEE Trans. on SP*, vol. 41, pp. 3445–3462, Dec. 1993.

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

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

[17] J.K. Romberg, H. Choi, and R.G. Baraniuk, "Bayesian wavelet domain image modeling using hidden Markov trees," in *Proc. IEEE ICIP*, Oct. 1999.