A&A 368, 730–746 (2001)
DOI: 10.1051/0004-6361:20000575

Astronomy & Astrophysics

Entropy and astronomical data analysis: Perspectives from multiresolution analysis

J.-L. Starck[1,2], F. Murtagh[3,4], P. Querre[1], and F. Bonnarel[4]

[1] DAPNIA/SEI-SAP, CEA-Saclay, 91191 Gif-sur-Yvette Cedex, France
[2] Statistics Department, Stanford University, Sequoia Hall, Stanford, CA 94305 USA
[3] School of Computer Science, The Queen's University of Belfast, Belfast BT7 1NN, Northern Ireland
[4] CDS, Observatoire Astronomique de Strasbourg, 11 rue de l'Université, 67000 Strasbourg, France

Received 30 October 2000 / Accepted 27 December 2000

Abstract. The Maximum Entropy Method is well-known and widely used in image analysis in astronomy. In its standard form it presents certain drawbacks, such an underestimation of the photometry. Various refinements of MEM have been proposed over the years. We review in this paper the main entropy functionals which have been proposed and discuss each of them. We define, from a conceptual point of view, what a good definition of entropy should be in the framework of astronomical data processing. We show how a definition of multiscale entropy fulfills these requirements. We show how multiscale entropy can be used for many applications, such as signal or image filtering, multi-channel data filtering, deconvolution, background fluctuation analysis, and astronomical image content analysis.

Key words. methods: data analysis – techniques: image processing

1. Introduction

The term "entropy" is due to Clausius (1865), and the concept of entropy was introduced by Boltzmann into statistical mechanics in order to measure the number of microscopic ways that a given macroscopic state can be realized. Shannon (1948) founded the mathematical theory of communication when he suggested that the information gained in a measurement depends on the number of possible outcomes out of which one is realized. Shannon also suggested that the entropy can be used for maximization of the bit transfer rate under a quality constraint. Jaynes (1957) proposed to use the entropy measure for radio interferometric image deconvolution, in order to select between a set of possible solutions that contains the minimum of information or, following his entropy definition, which has maximum entropy. In principle, the solution verifying such a condition should be the most reliable. A great deal of work has been carried out in the last 30 years on the use of entropy for the general problem of data filtering and deconvolution (Ables 1974; Bontekoe et al. 1994; Burg 1978; Frieden 1978a; Gull & Skilling 1991; Mohammad-Djafari 1994, 1998; Narayan & Nityananda 1986; Pantin & Starck 1996; Skilling 1989; Weir 1992). Traditionally information and entropy are determined from events and the probability of their occurrence. But signal and noise, rather than events and occurrences of events, are the basic building-blocks of signal and data analysis in the physical sciences. Instead of the probability of an event, we are led to consider the probabilities of our data being either signal or noise.

Observed data Y in the physical sciences are generally corrupted by noise, which is often additive and which follows in many cases a Gaussian distribution, a Poisson distribution, or a combination of both. Other noise models may also be considered. Using Bayes' theorem to evaluate the probability distribution of the realization of the original signal X, knowing the data Y, we have

$$p(X|Y) = \frac{p(Y|X).p(X)}{p(Y)} \tag{1}$$

$p(Y|X)$ is the conditional probability distribution of getting the data Y given an original signal X, i.e. it represents the distribution of the noise. It is given, in the case of uncorrelated Gaussian noise with variance σ^2, by:

$$p(Y|X) = \exp\left\{ -\sum_{\text{pixels}} \frac{(Y-X)^2}{2\sigma^2} \right\}. \tag{2}$$

The denominator in Eq. (1) is independent of X and is considered as a constant. This is the case of stationary noise. $p(X)$ is the a priori distribution of the solution X. In the absence of any information on the solution X except

Send offprint requests to: J.-L. Starck, e-mail: jstarck@cea.fr

its positivity, a possible course of action is to derive the probability of X from its entropy, which is defined from information theory.

The main idea of information theory (Shannon 1948) is to establish a relation between the received information and the probability of the observed event (Bijaoui 1984). If we denote $\mathcal{I}(E)$ the information related to the event E, and p the probability of this event happening, then we consider that

$$\mathcal{I}(E) = f(p). \tag{3}$$

Then we assume the two following principles:

- The information is a decreasing function of the probability. This implies that the more information we have, the less will be the probability associated with one event;
- Additivity of the information. If we have two independent events E_1 and E_2, the information $\mathcal{I}(E)$ associated with the occurrence of both is equal to the addition of the information of each of them:

$$\mathcal{I}(E) = \mathcal{I}(E_1) + \mathcal{I}(E_2). \tag{4}$$

Since E_1 (of probability p_1) and E_2 (of probability p_2) are independent, then the probability of both happening is equal to the product of p_1 and p_2. Hence

$$f(p_1 p_2) = f(p_1) + f(p_2). \tag{5}$$

Then we can say that the information measure is

$$\mathcal{I}(E) = k \ln(p) \tag{6}$$

where k is a constant. Information must be positive, and k is generally fixed at -1.

Another interesting measure is the mean information which is denoted

$$H = -\sum_i p_i \ln(p_i). \tag{7}$$

This quantity is called the entropy of the system and was established by Shannon (1948).

This measure has several properties:

- It is maximal when all events have the same probability $p_i = 1/N_\mathrm{e}$ (N_e being the number of events), and is equal to $\ln(N_\mathrm{e})$. It is in this configuration that the system is the most undefined;
- It is minimal when one event is sure. In this case, the system is perfectly known, and no information can be added;
- The entropy is a positive, continuous, and symmetric function.

If we know the entropy H of the solution (the next section describes different ways to calculate it), we derive its probability by

$$p(X) = \exp(-\alpha H(X)). \tag{8}$$

Given the data, the most probable image is obtained by maximizing $p(X|Y)$. Taking the logarithm of Eq. (1), we thus need to maximize

$$\ln(p(X|Y)) = -\alpha H(X) + \ln(p(Y|X)) - \ln(p(Y)). \tag{9}$$

The last term is a constant and can be omitted. Then, in the case of Gaussian noise, the solution is found by minimizing

$$J(X) = \sum_{\text{pixels}} \frac{(Y-X)^2}{2\sigma^2} + \alpha H(X) = \frac{\chi^2}{2} + \alpha H(X) \tag{10}$$

which is a linear combination of two terms: the entropy of the signal, and a quantity corresponding to χ^2 in statistics measuring the discrepancy between the data and the predictions of the model. α is a parameter that can be viewed alternatively as a Lagrangian parameter or a value fixing the relative weight between the goodness-of-fit and the entropy H.

For the deconvolution problem, the object-data relation is given by the convolution

$$Y = P * X \tag{11}$$

where P is the point spread function, and the solution is found (in the case of Gaussian noise) by minimizing

$$J(X) = \sum_{\text{pixels}} \frac{(Y - P*X)^2}{2\sigma^2} + \alpha H(X). \tag{12}$$

The way the entropy is defined is fundamental, because from its definition will depend the solution. The next section discusses the different approaches which have been proposed in the past. Multiscale Entropy, presented in Sect. 3, is based on the wavelet transform and noise modeling. It is a means of measuring information in a data set, which takes into account important properties of the data which are related to content. We describe how it can be used for signal and image filtering, and in Sect. 4 for image deconvolution. The case of multi-channel data is considered in Sect. 5. We then proceed to the use of multiscale entropy for description of image content. We pursue three directions of enquiry, respectively described in Sects. 6, 7 and 8. In Sects. 6 and 7, we determine whether signal is present in the image or not, possibly at or below the image's noise level; and how multiscale entropy is very well correlated with the image's content in the case of astronomical stellar fields. Knowing that multiscale entropy represents well the content of the image, we finally use it to define the optimal compression rate of the image. In all cases, a range of examples illustrate these new results.

2. The concept of entropy

We wish to estimate an unknown probability density $p(X)$ of the data. Shannon (1948), in the framework of the information theory, defined the entropy of an image X by

$$H_\mathrm{s}(X) = -\sum_{k=1}^{N_\mathrm{b}} p_k \log p_k \tag{13}$$

where $X = \{X_1, ..X_N\}$ is an image containing integer values, N_b is number of possible values which a given pixel X_k can take (256 for an 8-bit image), and the p_k values are derived from the histogram of X:

$$p_k = \frac{\#X_j = k}{N} \tag{14}$$

$\#X_j = k$ gives the number of pixels satisfying $X_j = k$.

If the image contains floating values, it is possible to to build up the histogram L of values L_i, using a suitable interval Δ, counting up how many times m_k each interval $(L_k, L_k + \Delta)$ occurs among the N occurrences. Then the probability that a data value belongs to an interval k is $p_k = \frac{m_k}{N}$, and each data value has a probability p_k.

The entropy is minimum and equal to zero when the signal is flat, and increases when we have some fluctuations. Using the entropy in Eq. (10) leads to the minimization of:

$$J(X) = \frac{\chi^2}{2} + \alpha H_\mathrm{s}(X). \tag{15}$$

This is a minimum entropy restoration method.

The trouble with this approach is that, because the number of occurrences is finite, the estimate p_k will be in error by an amount proportional to $m_k^{-\frac{1}{2}}$ (Frieden 1978b). The error becomes significant when m_k is small. Furthermore this kind of entropy definition is not easy to use for signal restoration, because the gradient of Eq. (15) is not easy to compute. For these reasons, other entropy functions are generally used. The main ones are as follows, where N is the number of pixels, and k represents a pixel index:

– Burg (1978):

$$H_\mathrm{b}(X) = -\sum_{k=1}^{N} \ln(X_k) \tag{16}$$

– Frieden (1978a):

$$H_\mathrm{f}(X) = -\sum_{k=1}^{N} X_k \ln(X_k) \tag{17}$$

– Gull & Skilling (1991):

$$H_\mathrm{g}(X) = \sum_{k=1}^{N} X_k - M_k - X_k \ln\left(\frac{X_k}{M_k}\right) \tag{18}$$

where M is a given model, usually taken as a flat image.

Each of these entropies can be used, and they correspond to different probability distributions that one can associate with an image (Narayan & Nityananda 1986). See (Frieden 1978a; Skilling 1989) for descriptions. The last of the above definitions of entropy has the advantage of having a zero maximum when X equals the model M. All of these entropy measures are negative (if $X_k > 1$), and maximum when the image is flat. They are decreasing when we introduce information, so minimizing the information is equivalent to maximizing the entropy for these definitions. They are negative because an offset term is omitted which has no importance for the minimization of the functional. The fact that we consider that a signal has maximum information value when it is flat is evidently a curious way to measure information. A consequence is that we must now maximize the entropy if we want a smooth solution, and the probability of X must be redefined by:

$$p(X) = \exp(\alpha H(X)). \tag{19}$$

The sign has been inverted (see Eq. (8)), which is natural if we want the best solution to be the smoothest. These three entropies, above, lead to the Maximum Entropy Method (MEM), for which the solution is found by minimizing (for Gaussian noise)

$$J(X) = \sum_{k=1}^{N} \frac{(Y_k - X_k)^2}{2\sigma^2} - \alpha H(X). \tag{20}$$

For the Shannon entropy (which is obtained from the histogram of the data), this is the opposite. The entropy is null for a flat image, and increases when the data contains some information. So, if the Shannon entropy were used for restoration, this would lead to a Minimum Entropy Method.

In 1986, Narayan & Nityananda (1986) compared several entropy functions, and finally concluded by saying that all were comparable if they have good properties, i.e. they enforce positivity, and they have a negative second derivative which discourages ripple. They showed also that results varied strongly with the background level, and that these entropy functions produced poor results for negative structures, i.e. structures under the background level (absorption area in an image, absorption band in a spectrum, etc.), and compact structures in the signal.

The Gull and Skilling entropy gives rise to the difficulty of estimating a model. Furthermore it has been shown (Bontekoe et al. 1994) that the solution is dependent on this choice. The α parameter determination is also not straightforward: see discussion in Pantin & Starck (1996).

In order to resolve these problems, Gull and Skilling proposed to limit the resolution of the solution, by introducing the concept of hidden image S and Intrinsic Correlation Function C (ICF, Gaussian- or cubic spline-like) in the Preblur MAXENT algorithm. The idea is to search for an image O which is the convolution product of a hidden solution S by C: $O = C * S$. Hence, the functional to minimize is:

$$H_\mathrm{g}(h) = \sum_{k=1}^{N} S_k - M_k - S_k \ln\left(\frac{S_k}{M_k}\right). \tag{21}$$

Since in astronomical images many scale lengths are present, the *Multi-channel Maximum Entropy Method*, developed by Weir (1991, 1992), uses a set of ICFs having different scale lengths, each defining a channel. Many new parameters appear in such an approch which lead to new

troubles in practice (Bontekoe et al. 1994). Most of them can be fixed by using the Pyramid Maximum Entropy (Bontekoe et al. 1994), or a wavelet approach (Pantin & Starck 1996).

As described above, many studies have been carried out in order to improve the functional to be minimized. But the question which should be raised is: what is a good entropy for signal restoration?

In Starck et al. (1998b), the benchmark properties for a good "physical" definition of entropy were discussed. Assuming that a signal X is the sum of several components:

$$X = S + B + N \tag{22}$$

where S is the signal of interest, B the background, and N the noise, we proposed that the following criteria should be verified:

1. The information in a flat signal is zero ($S = 0$, $N = 0$ and $B =$ Constant);
2. The amount of information in a signal is independent of the background (i.e., $H(X)$ is independent of B);
3. The amount of information is dependent on the noise (i.e., $H(X)$ is dependent on N); A given signal X does not furnish the same information in the different cases where the noise N is high or small;
4. The entropy must work in the same way for a pixel which has a value $B + \epsilon$, and for a pixel which has a value $B - \epsilon$. $H(X)$ must be a function of the absolute value of S instead of S;
5. The amount of information is dependent on the correlation in the signal. If the signal S presents large features above the noise, it contains a lot of information. By generating a new set of data from S, by randomly taking the pixel values in S, the large features will evidently disappear, and this new signal will contain less information. But the pixel values will be the same as in S.

Figure 1 illustrates the last point perfectly. The second image is obtained by distributing randomly the Saturn image pixel values, and the standard entropy definitions produce the same information measurement for both images. The concept of information becomes subjective, or at least it depends on the application domain. Indeed, for someone who is not involved in image processing, the second image contains *less* information than the first one. For someone working on image transmission, it is clear that the second image will require more bits for lossless transmission, and from this point of view, he/she will consider that the second image contains *more* information. Finally, for data restoration, all fluctuations due to noise are not of interest, and do not contain relevant information. From this physical point of view, the standard definitions of entropy seem badly adapted to information measurement in signal restoration.

These points are not axioms, but rather desirable properties that should be respected by the entropy functional in order to characterize well the data. We see that in these properties we are taking account of: (i) the background – very much a relative notion, associated with our understanding of the image or signal; and (ii) the noise – important when handling scientific images and signals. The background can also be termed continuum, or DC component, and is often very dependent on the semantics of the image. Our signal generation process could be conceived in terms of thermodynamics (Ferraro et al. 1999): the rate of variation of entropy is composed of internal heat changes, and heat transfers from external sources. The latter is our noise, N, and the former is signal including background.

Among all entropy functions proposed in the past, it is the Shannon one (Shannon 1948) which best respects the desiderata listed above. Indeed, if we assume that the histogram bin is defined as a function of the standard deviation of the noise, the first four points are verified, while none of these criteria is verified with other entropy functions (and only one of the desiderata is verified for the Gull and Skilling entropy by taking the model equal to the background).

Our critique of information measures is solely in view of our overriding goal, namely to define a demonstrably appropriate measure for image and signal processing in the physical sciences.

3. Multiscale entropy

3.1. Definition

Following on from the desirable criteria discussed in the previous section, a possibility is to consider that the entropy of a signal is the sum of the information at each scale of its wavelet transform (Starck et al. 1998b), and the information of a wavelet coefficient is related to the probability of it being due to noise. Let us look at how this definition holds up in practice. Denoting h the information relative to a single wavelet coefficient, we define

$$H(X) = \sum_{j=1}^{l} \sum_{k=1}^{N_j} h(w_{j,k}) \tag{23}$$

with $h(w_{j,k}) = -\ln p(w_{j,k})$. l is the number of scales, and N_j is the number of samples (pixels, time- or wavelength-interval values) in band (scale) j. For Gaussian noise, we get

$$h(w_{j,k}) = \frac{w_{j,k}^2}{2\sigma_j^2} + \text{Const.} \tag{24}$$

where σ_j is the noise at scale j. Below, when we use the information in a functional to be minimized, the constant term has no effect and we will omit it. We see that the information is proportional to the energy of the wavelet coefficients. The larger the value of a normalized wavelet coefficient, then the lower will be its probability of being noise, and the higher will be the information furnished by this wavelet coefficient. We can see easily that this entropy fulfills all the requirements listed in the previous section.

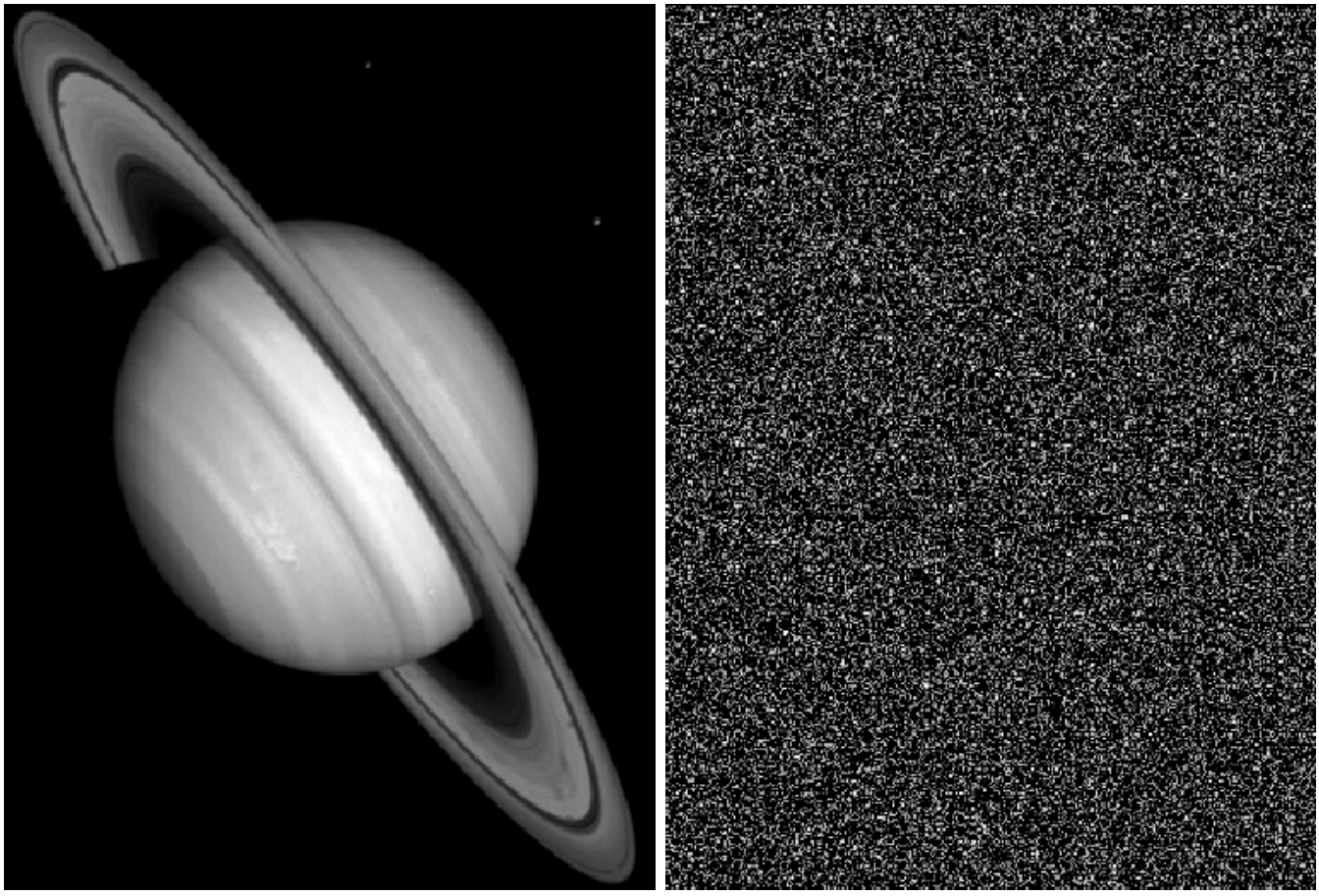

Fig. 1. Saturn image (left) and the same data distributed differently (right). These two images have the same entropy, using any of the standard entropy definitions

Just as for Shannon entropy, here information increases with entropy. Using such an entropy for optimization purposes will ordinarily lead to a minimum entropy method.

Equation (23) holds if the wavelet coefficients are statistically independent, which should imply that our approach is limited to an orthogonal or bi-orthogonal transform. This limitation may be addressed through the use of the so-called cycle-spinning algorithm (also named translation-invariant algorithm) (Donoho 1995), which consists of performing the process of "transform," "denoise," and "inverse transform" on every orthogonal basis corresponding to versions of the data obtainable by combinations of circular left-right and upwards-downwards translations. Donoho (Donoho 1995) has shown that using a non-decimating wavelet transform is equivalent to performing a set of decimated transforms with shift on the input signal. This means that Eq. (23) remains true for non-decimated wavelet transforms if it is normalized by the number of shifts. We will consider the orthogonal case in the following, knowing it can be generalized to non-decimated transforms.

3.2. Signal and noise information

Assuming that the signal X is still composed of the three components S, B, N ($X = S + B + N$), H is independent of B but not of N. Hence, our information measure is corrupted by noise, and we decompose our information measure into two components, one ($H_{\rm s}$) corresponding to the non-corrupted part, and the other ($H_{\rm n}$) to the corrupted part. We have (Starck et al. 1998b)

$$H(X) = H_{\rm s}(X) + H_{\rm n}(X). \tag{25}$$

We will define in the following $H_{\rm s}$ as the signal information, and $H_{\rm n}$ as the noise information. It is clear that noise does not contain any meaningful information, and so $H_{\rm n}$ describes a semantic component which is usually not informative to us. For each wavelet coefficient $w_{j,k}$, we have to estimate the proportions $h_{\rm n}$ and $h_{\rm s}$ of h (with $h(w_{j,k}) = h_{\rm n}(w_{j,k}) + h_{\rm s}(w_{j,k})$) which should be assigned to $H_{\rm n}$ and $H_{\rm s}$. Hence signal information and noise information are defined by

$$\begin{aligned} H_{\rm s}(X) &= \sum_{j=1}^{l}\sum_{k=1}^{N_j} h_{\rm s}(w_{j,k}) \\ H_{\rm n}(X) &= \sum_{j=1}^{l}\sum_{k=1}^{N_j} h_{\rm n}(w_{j,k}). \end{aligned} \tag{26}$$

If a wavelet coefficient is small, its value can be due to noise, and the information h relative to this single wavelet

coefficient should be assigned to $H_{\rm n}$. If the wavelet coefficient is high, compared to the noise standard deviation, its value cannot be due to the noise, and h should be assigned to $H_{\rm s}$. h can be distributed as $H_{\rm n}$ or $H_{\rm s}$ based on the probability $P_{\rm n}(w_{j,k})$ that the wavelet coefficient is due to noise, or the probability $P_{\rm s}(w_{j,k})$ that it is due to signal. We have $P_{\rm s}(w_{j,k}) = 1 - P_{\rm n}(w_{j,k})$. For the Gaussian noise case, we estimate $P_{\rm n}(w_{j,k})$ that a wavelet coefficient is due to the noise by

$$\begin{aligned} P_{\rm n}(w_{j,k}) &= \mathrm{Prob}(W > \mid w_{j,k} \mid) \\ &= \frac{2}{\sqrt{2\pi}\sigma_j} \int_{|w_{j,k}|}^{+\infty} \exp(-W^2/2\sigma_j^2) dW \\ &= \mathrm{erfc}\left(\frac{\mid w_{j,k} \mid}{\sqrt{2}\sigma_j}\right). \end{aligned}$$

The idea for deriving $h_{\rm s}$ and $h_{\rm n}$ is the following: we imagine that the information h relative to a wavelet coefficient is a sum of small information components $\mathrm{d}h$, each of them having a probability to be noise information. To understand this principle, consider two coefficients u and w ($w > u$) with Gaussian noise ($\sigma = 1$). The information relative to w is $h(w) = w^2$. When u varies from 0 to w with step $\mathrm{d}u$, the information $h(u)$ increases until it becomes equal to $h(w)$. When it becomes closer to w, the probability that the difference $w - u$ can be due to the noise increases, and the added information $\mathrm{d}h$ is more corrupted by the noise. By weighting the added information by the probability that the difference $w - u$ is due to the noise, we have:

$$h_{\rm n}(w_{j,k}) = \int_0^{|w_{j,k}|} P_{\rm n}(\mid w_{j,k} \mid -u) \left(\frac{\partial h(x)}{\partial x}\right)_{x=u} \mathrm{d}u \quad (27)$$

which is the noise information relative to a single wavelet coefficient, and

$$h_{\rm s}(w_{j,k}) = \int_0^{|w_{j,k}|} P_{\rm s}(\mid w_{j,k} \mid -u) \left(\frac{\partial h(x)}{\partial x}\right)_{x=u} \mathrm{d}u \quad (28)$$

which is the signal information relative to a single wavelet coefficient. For Gaussian noise, we have

$$\begin{aligned} h_{\rm n}(w_{j,k}) &= \frac{1}{\sigma_j^2} \int_0^{|w_{j,k}|} u\ \mathrm{erfc}\left(\frac{\mid w_{j,k} \mid -u}{\sqrt{2}\sigma_j}\right) \mathrm{d}u \\ h_{\rm s}(w_{j,k}) &= \frac{1}{\sigma_j^2} \int_0^{|w_{j,k}|} u\ \mathrm{erf}\left(\frac{\mid w_{j,k} \mid -u}{\sqrt{2}\sigma_j}\right). \end{aligned} \quad (29)$$

3.3. Filtering

The problem of filtering or restoring data D can be expressed by the following: we search for a solution $\tilde{D}$ such that the difference between D and $\tilde{D}$ minimizes the information due to the signal, and such that $\tilde{D}$ minimizes the information due to the noise.

$$J(\tilde{D}) = H_{\rm s}(D - \tilde{D}) + H_{\rm n}(\tilde{D}). \quad (30)$$

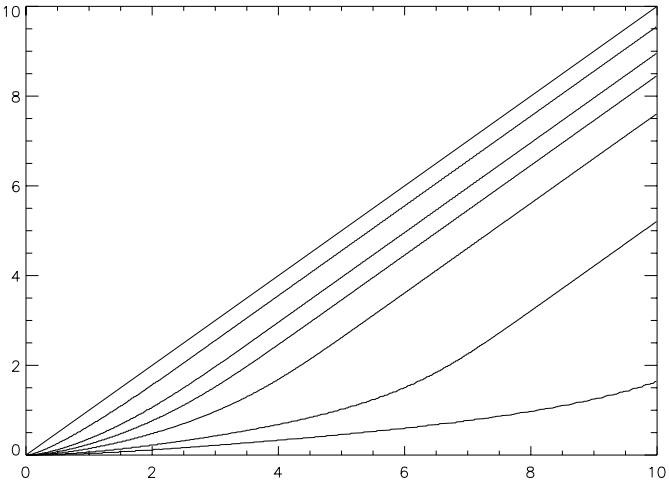

Fig. 2. Filtered wavelet coefficient versus the wavelet coefficient with different α values (from the top curve to the bottom one, α is equal, respectively, to 0, 0.1, 0.5, 1, 2, 5, 10)

Furthermore, the smoothness of the solution can be controlled by adding a parameter α:

$$J(\tilde{D}) = H_{\rm s}(D - \tilde{D}) + \alpha H_{\rm n}(\tilde{D}). \quad (31)$$

In practice (Chambolle et al. 1998), we minimize for each wavelet coefficient $w_{j,k}$:

$$j(\tilde{w}_{j,k}) = h_{\rm s}(w_{j,k} - \tilde{w}_{j,k}) + \alpha h_{\rm n}(\tilde{w}_{j,k}). \quad (32)$$

The solution is found by first computing the gradient $\nabla(J(\tilde{w}_{j,k}))$ (Starck 1999):

$$\begin{aligned} \nabla(j(\tilde{w}_{j,k})) = \int_0^{w_{j,k}-\tilde{w}_{j,k}} & P_{\rm s}(u)\mathrm{d}u \\ & + \alpha(\tilde{w}_{j,k} - \int_0^{\tilde{w}_{j,k}} P_{\rm s}(u)\mathrm{d}u) \end{aligned} \quad (33)$$

which gives for the Gaussian case:

$$\begin{aligned} \nabla(j(\tilde{w}_{j,k})) = & -\frac{w_{j,k} - \tilde{w}_{j,k}}{\sigma_j^2} \mathrm{erf}\left(\frac{w_{j,k} - \tilde{w}_{j,k}}{\sqrt{2}\sigma_j}\right) \\ & + \sqrt{\frac{2}{\pi}} \frac{1}{\sigma_j} \left[1 - \mathrm{e}^{-\frac{(w_{j,k} - \tilde{w}_{j,k})^2}{2\sigma_j^2}}\right] \\ & + \alpha \left(\frac{\tilde{w}_{j,k}}{\sigma_j^2} \mathrm{erfc}\left(\frac{\tilde{w}_{j,k}}{\sqrt{2}\sigma_j}\right) + \frac{1}{\sigma_j}\sqrt{\frac{2}{\pi}} \left[1 - \mathrm{e}^{-\frac{\tilde{w}_{j,k}^2}{2\sigma_j^2}}\right]\right). \end{aligned} \quad (34)$$

The solution of the equation $\nabla(j(\tilde{w}_{j,k})) = 0$ can be obtained by any minimization routine. In our examples below, we have used simple dichotomy.

Figure 2 shows the result when minimizing the functional j with different α values, and noise standard deviation equal to 1. The filtered wavelet coefficient is plotted versus the wavelet coefficient. From the top curve to the bottom, α is respectively equal to 0, 0.1, 0.5, 1, 2, 5, 10.

The higher the value of α, the more the filtered wavelet coefficient is reduced. When α is equal to 0, there is no regularization and the data are unchanged.

Simulations have shown (Starck & Murtagh 1999) that the MEF method produces a better result than the standard soft or hard thresholding, from both the visual aspect and PSNR (peak signal-to-noise ratio). Figures 3 and 4 show the filtering respectively on simulated noisy blocks and on a real spectrum.

A robust way to constrain α is to use the fact that we expect a residual with a given standard deviation at each scale j equal to the noise standard deviation σ_j at the same scale. Then rather than a single α we have an α_j per scale. A full description of the MEF algorithm can be found in Starck & Murtagh (1999).

4. Deconvolution

4.1. Introduction

Consider an image characterized by its intensity distribution (the "data") I, corresponding to the observation of a "real image" O through an optical system. If the imaging system is linear and shift-invariant, the relation between the data and the image in the same coordinate frame is a convolution:

$$I = O * P + N \tag{35}$$

P is the point spread function (PSF) of the imaging system, and N is additive noise. In practice $O*P$ is subject to non-stationary noise which one can tackle by simultaneous object estimation and restoration (Katsaggelos 1991). The issue of more extensive statistical modeling will not be further addressed here (see Llacer & Núñez 1990; Lorenz & Richter 1993; Molina 1994), beyond noting that multiresolution frequently represents a useful framework, allowing the user to introduce a priori knowledge of objects of interest.

We want to determine $O(x, y)$ knowing I and P. This inverse problem has led to a large amount of work, the main difficulties being the existence of: (i) a cut-off frequency of the point spread function, and (ii) the additive noise. See for example (Cornwell 1989).

Equation (35) is usually in practice an ill-posed problem. This means that there is no unique and stable solution.

4.2. The principle

The most realistic solution is that which minimizes the amount of information, but remains compatible with the data. By the MEM method, minimizing the information is equivalent to maximizing the entropy and the functional to minimize is

$$J(O) = \sum_{k=1}^{N} \frac{(I_k - (P * O)_k)^2}{2\sigma_I^2} - \alpha H(O). \tag{36}$$

where H is either the Frieden or the Gull and Skilling entropy.

Similarly, using the multiscale entropy, minimizing the information is equivalent to minimizing the entropy and the functional to minimize is

$$J(O) = \sum_{k=1}^{N} \frac{(I_k - (P * O)_k)^2}{2\sigma_I^2} + \alpha H(O). \tag{37}$$

We have seen that in the case of Gaussian noise, H is given by the energy of the wavelet coefficients. We have

$$J(O) = \sum_{k=1}^{N} \frac{(I_k - (P * O)_k)^2}{2\sigma_I^2} + \alpha \sum_{j=1}^{l} \sum_{k=1}^{N_j} \frac{w_{j,k}^2}{2\sigma_j^2} \tag{38}$$

where σ_j is the noise at scale j, N_j the number of pixels at the scale j, σ_I the noise standard deviation in the data, and l the number of scales.

Rather than minimizing the amount of information in the solution, we may prefer to minimize the amount of information which can be due to the noise. The function is now:

$$J(O) = \sum_{k=1}^{N} \frac{(I_k - (P * O)_k)^2}{2\sigma_I^2} + \alpha H_{\mathrm{n}}(O) \tag{39}$$

and for Gaussian noise, H_{n} has been defined by

$$H_{\mathrm{n}}(X) = \sum_{j=1}^{l} \sum_{k=1}^{N_j} \frac{1}{\sigma_j^2} \int_0^{|w_{j,k}|} u \ \mathrm{erf}\left(\frac{|\ w_{j,k}\ | - u}{\sqrt{2}\sigma_j} \right). \tag{40}$$

The solution is found by computing the gradient $\nabla(J(O))$ and performing the following iterative schema:

$$O^{n+1} = O^n - \gamma \nabla(J(O^n)). \tag{41}$$

We consider an α_j per scale, and introduce thereby an adaptive regularization which depends on the signal-to-noise ratio of the input data wavelet coefficients.

4.3. The parameters

In order to introduce flexibility in the way we restore the data, we introduce two parameters $\beta_{j,k}$ and $\alpha_{j,k}$ which allow us to weight, respectively, the two terms of the equation to be minimized:

$$J(O) = \frac{1}{2\sigma_I^2} \sum_{k=1}^{N} \left(\sum_j \sum_l \beta_{j,k} w_{j,l}(R) \psi_{j,l}(k) \right)^2 + \sum_{j=1}^{l} \sum_{k=1}^{N_j} \alpha_{j,k} h(w_{j,k}(O))$$

where $R = I - P * O$, and $R = \sum_j \sum_k w_{j,k}(R) \psi_{j,k}$ ($w_{j,k}(R)$ are the wavelet coefficients of R, and $w_{j,k}(O)$ are the wavelet coefficients of O).

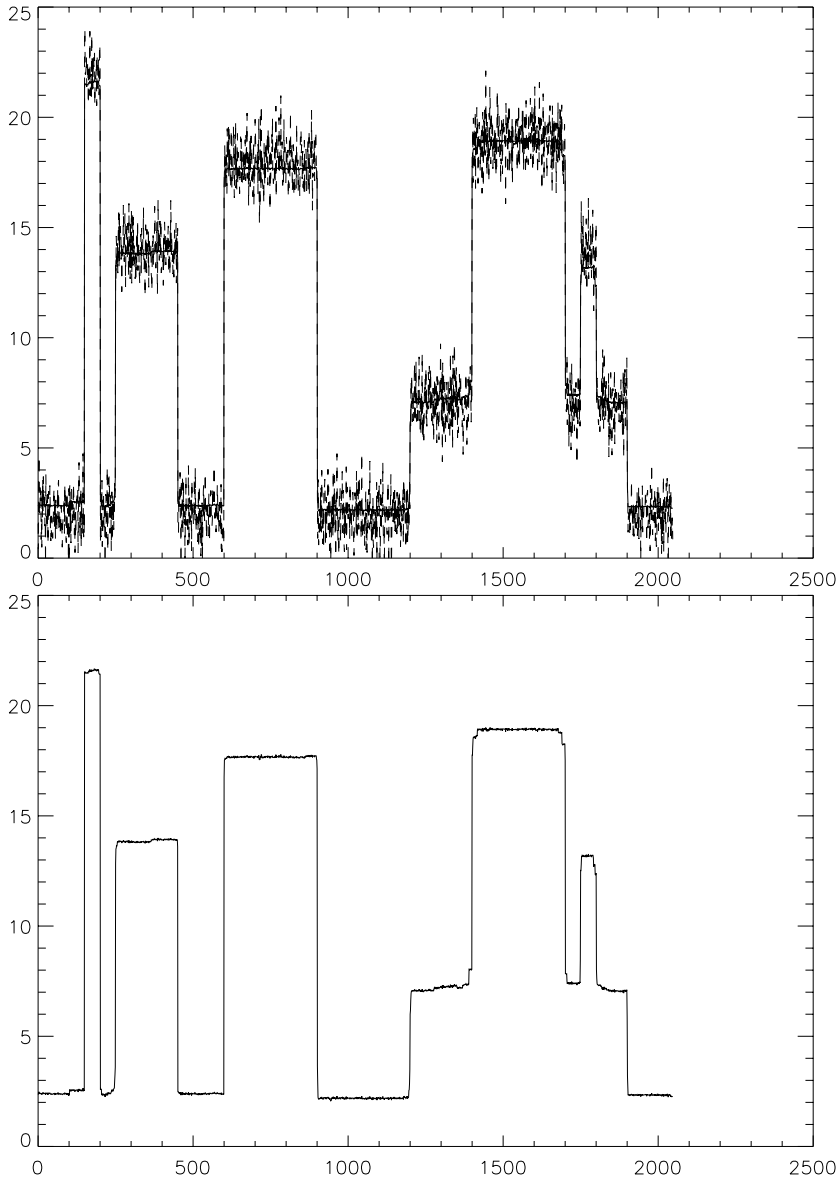

Fig. 3. Top, noisy blocks and filtered blocks overplotted. Bottom, filtered blocks

We consider three approaches for estimating $\beta_{j,k}$

1. No weighting: $\beta_{j,k} = 1$;
2. Soft weighting: $\beta_{j,k} = P_{\rm s}(w_{j,k}(I))$.
   In this case, $\beta_{j,k}$ is equal to the probability that the input data wavelet coefficient is due to signal (and not to noise);
3. Hard weighting: $\beta_{j,k} = 0$ or 1 depending on $P_{\rm n}(w_{j,k}(I))$ ($P_{\rm n}(w_{j,k}(I)) = 1 - P_{\rm s}(w_{j,k}(I))$). This corresponds to using only significant input data wavelet coefficients.

$\alpha_{j,k}$ is the product of two values: $\alpha_{j,k} = \alpha_{\rm u}\beta'_{j,k}$.

- $\alpha_{\rm u}$ is a user parameter (defaulted to 1) which allows us to control the smoothness of the solution. Increasing $\alpha_{\rm u}$ produces a smoother solution;
- $\beta'_{j,k}$ depends on the input data and can take the following value:

1. No regularization ($\beta'_{j,k} = 0$): only the first term of the functional is minimized;
2. No protection from regularization ($\beta'_{j,k} = 1$): the regularization is applied at all positions and at all the scales;
3. Soft protection ($\beta'_{j,k} = P_{\rm n}(w_{j,k}(I))$): the regularization becomes adaptive, depending on the probability that the input wavelet coefficient is due to noise;
4. Hard protection ($\beta'_{j,k} = 0$ or 1 depending on $P_{\rm n}(w_{j,k}(I))$);
5. Soft + hard protection: ($\beta'_{j,k} = 0$ or $P_{\rm n}(w_{j,k}(I))$ depending on $P_{\rm n}(w_{j,k}(I))$).

We see that choosing a hard weighting and no regularization leads to deconvolution from the multiresolution support (Starck et al. 1998a).

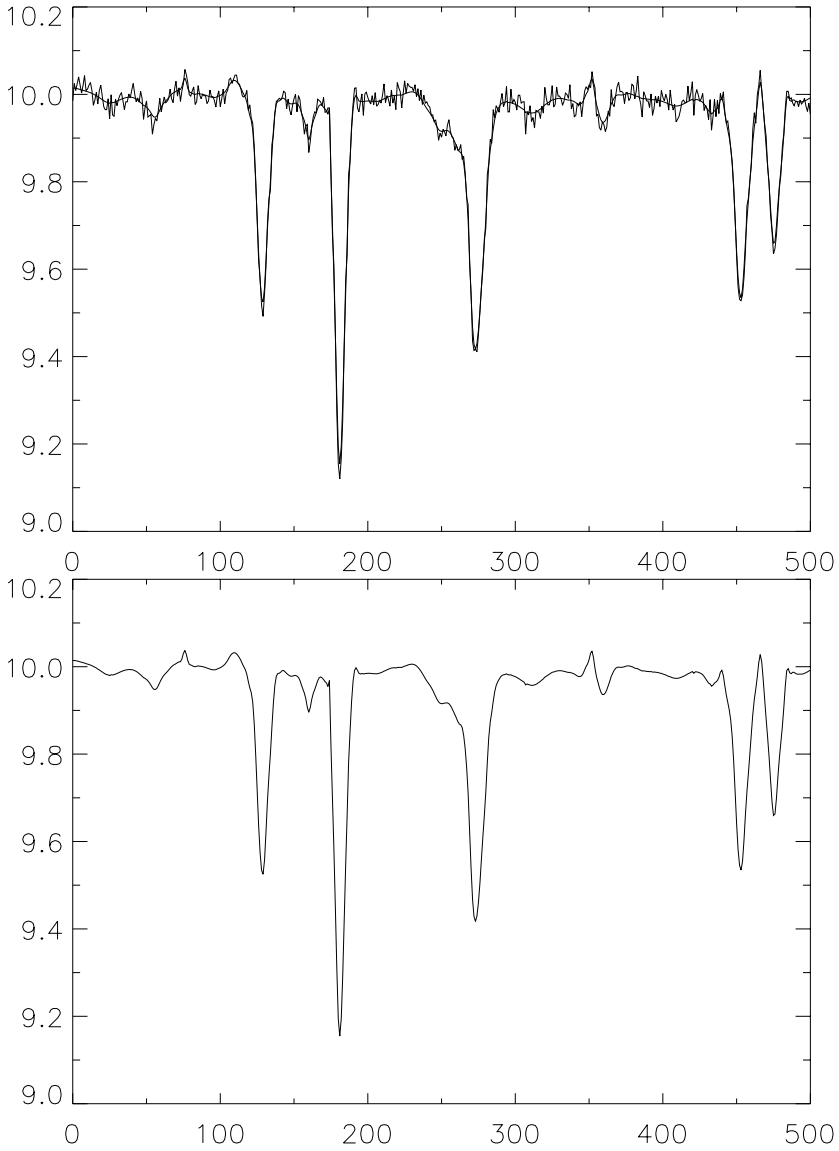

Fig. 4. Top, noisy blocks and filtered blocks overplotted. Bottom, filtered blocks

4.4. Examples

Figure 5 shows a simulation. The original image (panel (a)) contains stars and galaxies. Fig. 5b shows the data (blurred image + Gaussian noise), Fig. 5c shows the deconvolved image, and Fig. 5d the residual image (i.e. data minus solution reconvolved by the PSF). The blurred image SNR is 12dB, and the deconvolved image SNR is 23.11 dB.

5. Multichannel data

5.1. Introduction

The challenge for multichannel data restoration is to have a data representation which takes into account at the same time both the spatial and the spectral (or temporal) correlation. A three-dimensional transform-based coding technique has been proposed in Saghri et al. (1995), consisting of a one-dimensional spectral Karhunen-Loève transform (Karhunen 1947) (KLT) and a two-dimensional spatial discrete cosine transform (DCT). The KLT is used to decorrelate the spectral domain and the DCT is used to decorrelate the spatial domain. All images are first decomposed into blocks, and each block uses its own Karhunen-Loève transform instead of one single matrix for the whole image. Lee (Lee 1999) has improved this approach by introducing a varying block size. The block size is adapted using a quadtree and a bit allocation for each block. The DCT transform can also be replaced by a wavelet transform (WT) (Epstein et al. 1992; Tretter & Bouman 1995).

We introduce here the Wavelet-Karhunen-Loève transform (WT-KLT) and show how to use it for noise removal. Decorrelating first the data in the spatial domain using the WT and afterwards in the spectral domain using the KLT, allows us to derive robust noise modeling in the WT-KLT space, and hence to filter the transformed data in an

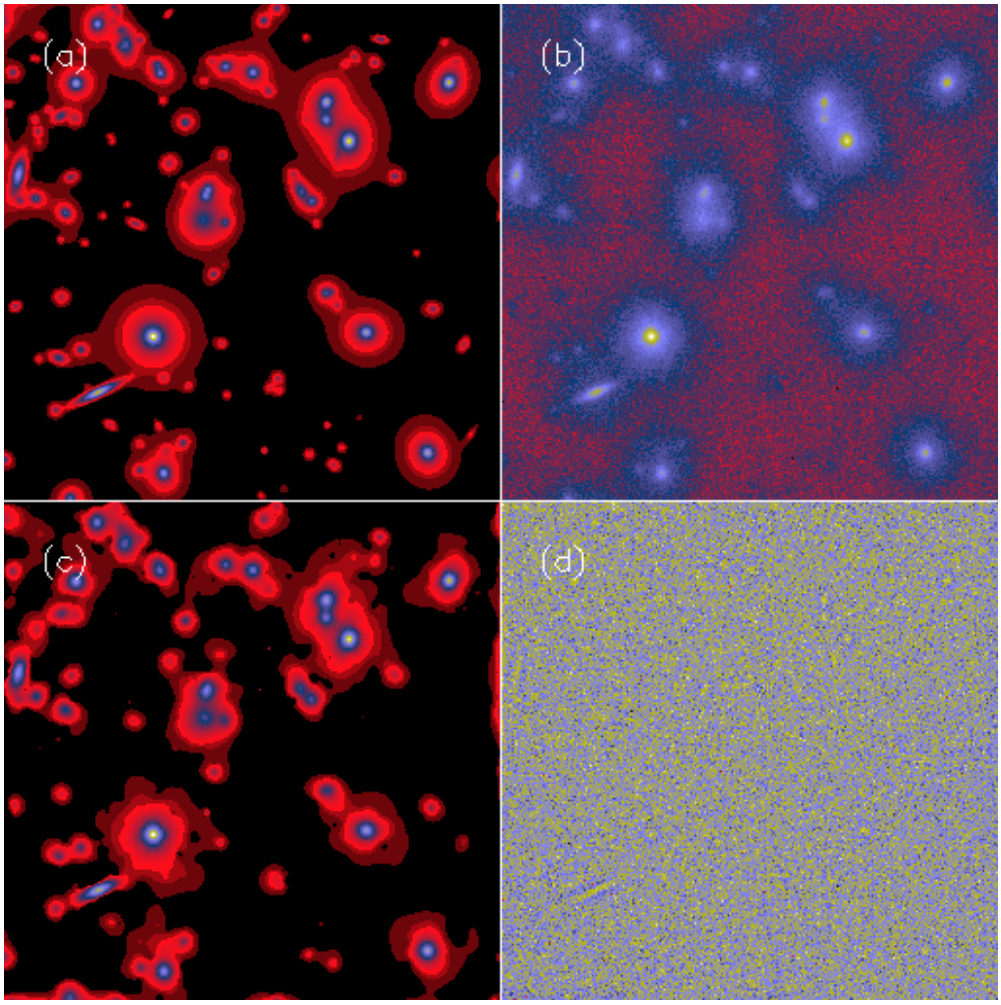

Fig. 5. a) Original image, **b)** blurred image + Gaussian noise, **c)** deconvolved image, and **d)** residual image

efficient way. We show also that the correlation matrix can be computed by different methods, taking into account the noise modeling.

5.2. The Wavelet-Karhunen-Loève transform

5.2.1. Definition

The Karhunen-Loève transform, also often referred to as eigenvector, Hotelling transform, or Principal Component Analysis (PCA) (Karhunen 1947; Loève 1948; Hotelling 1933) allows us to transform discrete signals into a sequence of uncorrelated coefficients. Considering a vector $D = d_1, ..., d_L$ of L signals or images of dimension N (i.e. N pixels per image), we denote $M = \{m_1, ..., m_L\}$ the mean vector of the population (m_i is the mean of the ith signal d_i). The covariance matrix C of D is defined by $C = (D - M)(D - M)^t$, and is of order $L \times L$. Each element $c_{i,i}$ of C is the variance of d_i, and each element $c_{i,j}$ is the covariance between d_i and d_j. The KLT method consists of applying the following transform to all vectors $x_i = \{d_1(i), ..., d_L(i)\}$ $(i = 1..N)$:

$$y_i = \Lambda^{-\frac{1}{2}} A(x_i - M) \tag{42}$$

where Λ is the diagonal matrix of eigenvalues of the covariance matrix C, and A is a matrix whose rows are formed from the eigenvectors of C (Gonzalez 1993), ordered following the monotonic decreasing order of eigenvalues.

Because the rows of A are orthonormal vectors, $A^{-1} = A^t$, and any vector x_i can be recovered from its corresponding y_i by:

$$x_i = \Lambda^{\frac{1}{2}} A^t y_i + M. \tag{43}$$

The Λ matrix multiplication can be seen as a normalization. Building A from the correlation matrix instead of the covariance matrix leads to another kind of normalization, and the Λ matrix can be suppressed ($y_i = A(x_i - M)$ and $x_i = A^t y_i + M$). Then the norm of y will be equal to the norm of x.

We suppose now that we have L observations of the same view, e.g. at different wavelengths (or at different epochs, etc.), and denote as d_l one observation, $W^{(l)}$ its wavelet transform, and $w_{l,j,k}$ one wavelet coefficient at scale j and at position k. The standard approach would be to use an orthogonal wavelet transform, and to calculate the correlation matrix C from the wavelet coefficients

instead of the pixel values:

$$C_{m,n} = \frac{\sum_{j=1}^{J}\sum_{k=1}^{N_j} w_{m,j,k} w_{n,j,k}}{\sqrt{\sum_{j=1}^{J}\sum_{k=1}^{N_j} w_{m,j,k}^2}\sqrt{\sum_{j=1}^{J}\sum_{k=1}^{N_j} w_{n,j,k}^2}} \quad (44)$$

where J is the number of bands, and N_j is the number of coefficients in band j (Murtagh 1998). In (Lee 1999), a more complex approach was proposed, which is to decompose the images into N_b blocks and apply a KLT for each block separately. We investigate here different approaches for data restoration.

5.2.2. Correlation matrix and noise modeling

We introduce a noise model into our calculation of the correlation matrix. Indeed, if the input sequence D contains noise, then the wavelet coefficient are noisy too. Eigenvalues at the high scales are computed with noisy WT coefficients and we may lose the true underlying correlation that may exist between the input images d_l. The expression of the correlation matrix has to be modified in order to allow us to take the noise into account. We add a weighting term to each wavelet coefficient which depends on the signal-to-noise ratio. The correlation matrix is calculated by

$$C_{m,n} = \frac{\sum_{j=1}^{J}\sum_{k=1}^{N_j} p_j(w_{m,j,k}) w_{m,j,k} p_j(w_{n,j,k}) w_{n,j,k}}{\sqrt{\sum_{j=1}^{J}\sum_{k=1}^{N_j} p_j^2(w_{m,j,k}) w_{m,j,k}^2}\sqrt{\sum_{j=1}^{J}\sum_{k=1}^{N_j} p_j^2(w_{n,j,k}) w_{n,j,k}^2}} \quad (45)$$

where p_j is a weighting function. The standard approach corresponds to the specific case where $p_j(w_m) = 1$ (no weighting). By considering that only wavelet coefficients with high signal-to-noise ratio should be used for the correlation matrix calculation, p_j can be defined by:

$$p_j(w) = \begin{cases} 1 & \text{if } w \text{ is significant} \\ 0 & \text{if } w \text{ is not significant} \end{cases} \quad (46)$$

and a wavelet coefficient w is said to be "significant" if its probability to be due to the noise is smaller than a given ϵ value. In the case of Gaussian noise, it suffices to compare the wavelet coefficients w to a threshold level t_j. t_j is generally taken as $\lambda\sigma_j$, where σ_j is the noise standard deviation at scale j, and λ is chosen between 3 and 5. The value of $\lambda = 3$ corresponds to a probability of false detection of 0.27%, for a Gaussian statistic.

Changes brought about in the first eigenvalue through such hard thresholding in wavelet space are studied in (Murtagh 1998). This hard weighting scheme may lead to problem if only a few coefficients are significant, and can be replaced by a soft weighting scheme, by defining $p_j(w)$ by:

$$p_j(w) = 1 - \mathrm{Prob}(W > \mid w \mid) \quad (47)$$

where $\mathrm{Prob}(W > \mid w \mid)$ is the probability that a wavelet coefficient is larger than w due to the noise. For Gaussian noise, we have:

$$\begin{aligned} p_j(w) &= 1 - \frac{2}{\sqrt{2\pi}\sigma_j}\int_{|w|}^{+\infty} \exp(-W^2/2\sigma_j^2)\mathrm{d}W \\ &= \mathrm{erf}\left(\frac{\mid w \mid}{\sqrt{2}\sigma_j}\right). \end{aligned} \quad (48)$$

5.2.3. Scale and Karhunen-Loève transform

We can also analyze separately each band of the wavelet transform, and then apply one KLT per resolution level. This implies calculating a correlation matrix $C^{(j)}$ for each band j.

$$C^{(j)}_{m,n} = \frac{\sum_{k=1}^{N_j} p_j(w_{m,j,k}) w_{m,j,k} p_j(w_{n,j,k}) w_{n,j,k}}{\sqrt{\sum_{k=1}^{N_j} p_j^2(w_{m,j,k}) w_{m,j,k}^2}\sqrt{\sum_{k=1}^{N_j} p_j^2(w_{n,j,k}) w_{n,j,k}^2}}. \quad (49)$$

This has the advantage of taking into account more complex behavior of the signal. Indeed, structures of different sizes may have a different spectral behavior (for example, stars and galaxies in astronomical images), and a band-by-band independent analysis allows us to represent better such data.

5.2.4. The WT-KLT transform

The final WT-KLT algorithm has the following steps:

1. Estimate the noise standard deviation $\sigma^{(l)}$ of each input data set d_l;
2. Calculate the wavelet transform $W^{(l)}$ of each input data set d_l;
3. For each band j of the wavelet transform, calculate the correlation matrix $C^{(j)}$ relative to the vector $x_j = \left\{W_j^{(1)}, W_j^{(2)}, ..., W_j^{(L)}\right\}$, where $W_j^{(l)}$ represents the band j of the wavelet transform $W^{(l)}$ of d_l;
4. For each band j, we diagonalize the matrix $C^{(j)}$ and build the transform matrix A_j from the eigenvectors of $C^{(j)}$;
5. For each band j and each position k, we apply the matrix A_j to the vector $x_{j,k} = \{w_{1,j,k}, w_{2,j,k}, ..., w_{L,j,k}\}$:
$$y_{j,k} = A_j x_{j,k}; \quad (50)$$
6. The WT-KLT coefficients $c_{l,j,k}$ are derived from $y_{j,k}$ by $c_{l,j,k} = y_{j,k}(l)$. The l index in the transformed coefficients no longer represents the observation number, but instead the eigenvector number. $l = 1$ indicates the main eigenvector while $l = L$ indicates the last one.

The mean vector M disappears in this algorithm because the wavelet coefficients are of zero mean.

Figure 6 shows the flowchart of the WT-KLT transform.

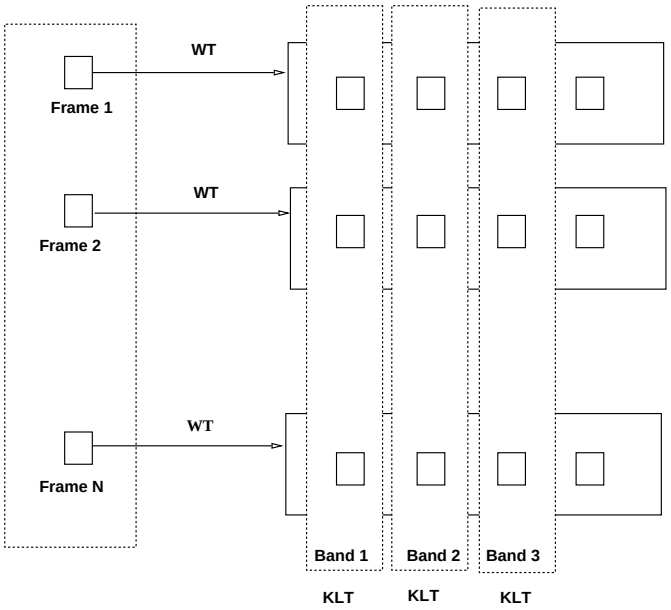

Fig. 6. WT-KLT transform flowchart. Each frame of the input dataset is first wavelet transformed, and a principal component analysis is applied at each resolution level

5.2.5. The WT-KLT reconstruction algorithm

The reconstruction algorithm has the following steps:

1. For each band j and each position k, we apply the matrix A_j^t to the vector $y_{j,k} = \{c_{1,j,k}, c_{2,j,k}, ..., c_{L,j,k}\}$

$$x_{j,k} = A_j^t y_{j,k}; \tag{51}$$

2. The wavelet coefficients $w_{l,j,k}$ are derived from $x_{j,k}$ by $w_{l,j,k} = x_{j,k}(l)$;
3. An inverse wavelet transform of $W^{(l)}$ furnishes d_l.

5.3. Noise modeling in the WT-KLT space

Since a WT-KLT coefficient c is obtained by two successive linear transforms, analytic noise modeling can be carried out in order to derive the noise standard deviation associated with the c value.

5.3.1. Non-Gaussian noise

If the noise in the data D is Poisson, the Anscombe transformation (Anscombe 1948)

$$t(D) = 2\sqrt{D + \frac{3}{8}} \tag{52}$$

acts as if the data arose from a Gaussian white noise model, with $\sigma = 1$, under the assumption that the mean value of I is sufficiently large. The arrival of photons, and their expression by electron counts, on CCD detectors may be modeled by a Poisson distribution. In addition, there is additive Gaussian read-out noise. The Anscombe transformation has been extended to take this combined noise into account. The generalization of the variance stabilizing Anscombe formula is derived as (Starck et al. 1998a):

$$t(D) = \frac{2}{g}\sqrt{gD + \frac{3}{8}g^2 + \sigma^2 - gm} \tag{53}$$

where g is the electronic gain of the detector, σ and m the standard deviation and the mean of the read-out noise.

This implies that for the filtering of an image with Poisson noise or a mixture of Poisson and Gaussian noise, we will first pre-transform the data D into another dataset $t(D)$ with Gaussian noise. Then $t(D)$ will be filtered, and the filtered data will be inverse-transformed.

For other kinds of noise, modeling must be performed in order to define the noise probability distribution of the wavelet coefficients (Starck et al. 1998a). In the following, we will consider only stationary Gaussian noise.

5.3.2. Noise level on WT-KLT coefficients

Assuming a Gaussian noise standard deviation σ_l for each signal or image d_l, the noise in the wavelet space follows a Gaussian distribution $\sigma_{l,j}$, j being the scale index. For a bi-orthogonal wavelet transform with an L^2 normalization, $\sigma_{l,j} = \sigma_l$ for all j. Since the WT-KLT coefficients are obtained from a linear transform, we can easily derive the noise standard deviation of a WT-KLT coefficient from the noise standard deviation of the wavelet coefficients, as follows. Considering the noise standard deviation vector $s = \{\sigma_1, ..., \sigma_L\}$, we apply the following transformation:

$$y_j = A_j^2 s^2 \tag{54}$$

and the noise standard deviation relative to a WT-KLT coefficient $c_l(j,k)$ is $\sqrt{y_j(l)}$.

5.4. Multiscale entropy and multichannel data

The multiscale entropy relative to a set of observations $D(1..M)$ can be written as:

$$H(D) = \sum_{l=1}^{L}\sum_{j=1}^{J}\sum_{k=1}^{N_j} h(c_{l,j,k}) \tag{55}$$

where J is the number of scales used in the wavelet transform decomposition, L the number of observations, k a pixel position, c a WT-PCA coefficient, and l denotes the eigenvector number.

The last scale of the wavelet transform is not used, in line with our development of this methodology, so this entropy measurement is background independent. This in turn is very important because the background can vary from one wavelength to another.

As for a wavelet coefficient in the case of single channel data, we know the noise standard deviation relative to a coefficient, and coefficients are of zero mean. Therefore, we can apply the same filtering method. The filtered WT-PCA coefficients are found by minimizing for each $c_{l,j,k}$:

$$j(\tilde{c}_{l,j,k}) = h_s(c_{l,j,k} - \tilde{c}_{l,j,k}) + \alpha h_n(\tilde{c}_{l,j,k}). \tag{56}$$

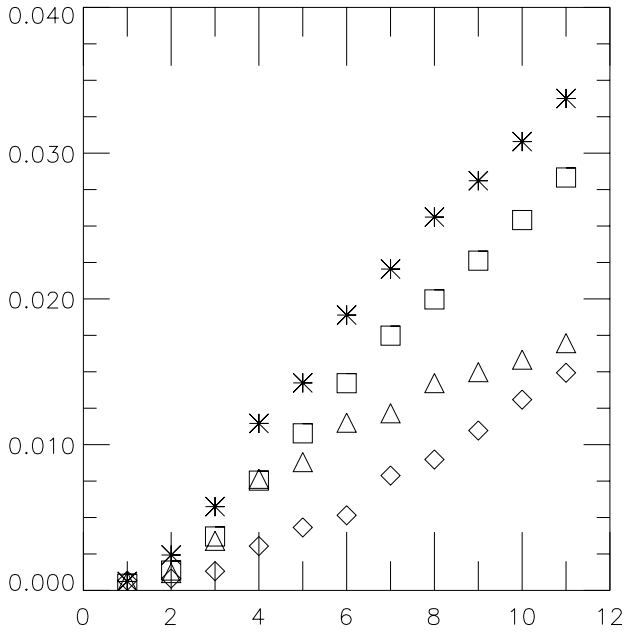

Fig. 7. Simulation: integrated Root Mean Square Error (vertical) versus the noise standard deviation (horizontal). See text

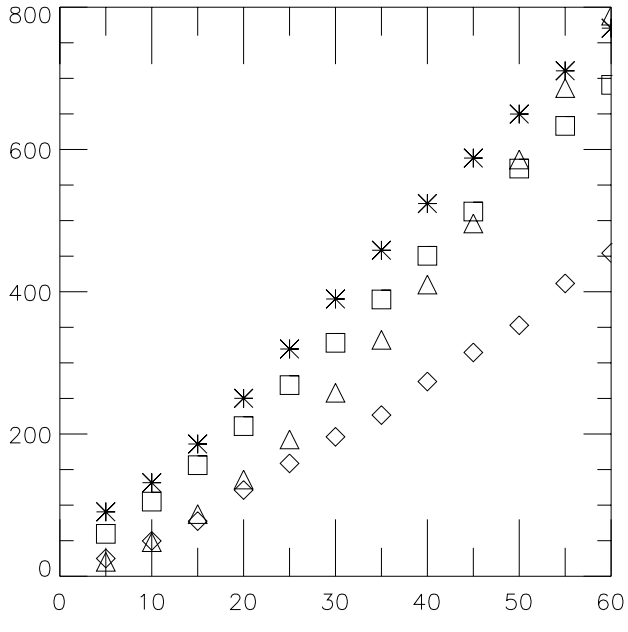

Fig. 8. Planck simulation: integrated Root Mean Square Error (vertical) versus the noise standard deviation (horizontal). See text

5.5. Example

Figure 7 shows a simulation. We created a dataset of 18 frames, each containing a source at the same position, but at different intensity levels. The source is a small Gaussian. The data cannot be coadded because the level of the source varies from one frame to another (variable source). Additive noise was used, and the data were filtered. We calculated the Root Mean Square Error (RMSE) on each individual frame using a 5×5 square centered on the source. Hence, the RMSE reflects well the photometric errors. The addition of the 18 RMSE values, which we call IRMSE (integrated RMSE), furnishes us with a reliable measurement of the filtering quality. The simulation was repeated with 12 noise levels, and four different filtering methods were compared. Figure 7 shows the IRMSE versus the noise standard deviation plot. The four methods are (i) multiscale entropy applied on the WT-KLT coefficients (diamond), (ii) reconstruction from a subset of eigenvectors of the KLT (triangle), (iii) multiscale entropy applied on each frame independently (square), and (iv) thresholding applied to the wavelet transform of each frame (star). This simulation shows clearly that the approach proposed here, multiscale entropy applied to the WT-KLT coefficients, outperforms all other methods.

The same experiments were performed using a simulated Planck dataset. The dataset contains ten images, each one being a linear combination of 6 sky components images (CMB, SZ, free-free, etc.). As in the previous simulation, noise was added, and the data were filtered by the four methods. The only difference is that the RMSE is calculated on the full frames. Figure 8 shows IRMSE versus the noise standard deviation plot. Diamonds, triangles, squares and stars represent the same methods as before. Again, the multiscale entropy applied on the WT-KLT coefficients outperforms the other methods.

6. Multiscale entropy applied to background fluctuation analysis

The mean entropy vector

The multiscale entropy has been defined by:

$$H(X) = \sum_{j=1}^{l} \sum_{k=1}^{N} h(w_j) \tag{57}$$

with $h(w_j = \ln(p(w_j(k))))$. In order to study the behavior of the information at a given scale, we prefer to calculate the mean entropy vector E defined by:

$$E(j) = \frac{1}{N} \sum_{k=1}^{N} h(w_j) \tag{58}$$

$E(j)$ gives the mean entropy at scale j. From the mean entropy vector, we have statistical information on each scale separately. Having a noise model, we are able to calculate (generally from simulations) the mean entropy vector $E^{(\mathrm{noise})}(j)$ resulting from pure noise. Then we define the normalized mean entropy vector by

$$E_n(j) = \frac{E(j)}{E^{(\mathrm{noise})}(j)}. \tag{59}$$

Figure 9 shows the result of a simulation. Five simulated images were created by adding n sources to a 1024×1024 image containing Gaussian noise of standard deviation equal to 1 (and arbitrary mean). The n sources are identical, with a maximum equal to 1, standard deviation equal

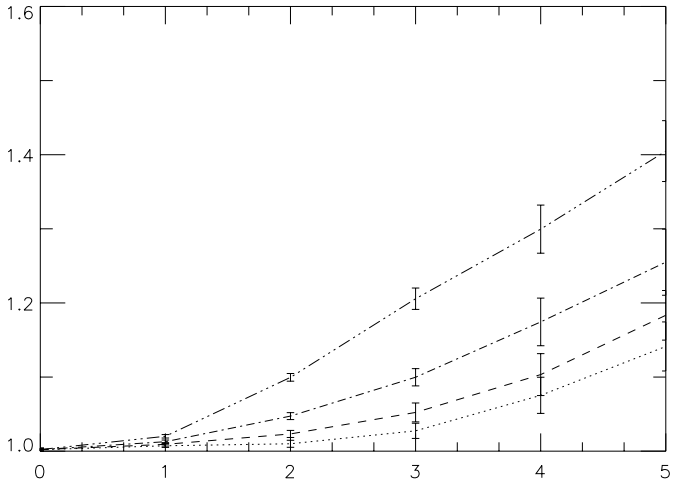

Fig. 9. Mean entropy versus the scale of 5 simulated images containing undetectable sources and noise. Each curve corresponds to the multiscale transform of one image. From top to bottom, the image contains respectively 400, 200, 100, 50 and 0 sources

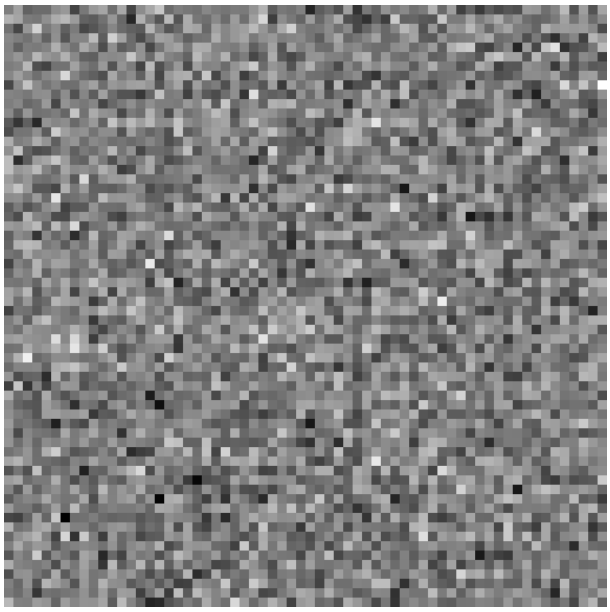

Fig. 10. Region of a simulated image containing an undetectable source at the center

to 2, and zero covariance terms. Defining the signal-to-noise ratio (SNR) as the ratio between the standard deviation in the smallest box which contains at least 90% of the flux of the source, and the noise standard deviation, we have a SNR equal to 0.25. The sources are not detectable in the simulated image, nor in its wavelet transform. Figure 10 shows a region which contains a source at the center. It is clear there is no way to find this kind of noisy signal. The five images were created using a number of sources respectively equal to 0, 50, 100, 200 and 400, and the simulation was repeated ten times with different noise maps in order to have an error bar on each entropy measurement. For the image which contains 400 sources, the number of pixels affected by a source is less than 2.5%.

When the number of sources increases, the difference between the multiscale entropy curves increases. Even if the sources are very faint, the presence of signal can be clearly detected using the mean entropy vector. But it is obvious that the positions of these sources remain unknown.

7. Multiscale entropy as a measure of relevant information in an image

Since the multiscale entropy extracts the information from the signal only, it was a challenge to see if the astronomical content of an image was related to its multiscale entropy.

For this purpose, we studied the astronomical content of 200 images of 1024×1024 pixels extracted from scans of 8 different photographic plates carried out by the MAMA digitization facility (Paris, France) (Guibert 1992) and stored at CDS (Strasbourg, France) in the Aladin archive (Bonnarel et al. 1999). We estimated the content of these images in three different ways:

1. By counting the number of objects in an astronomical catalog (USNO A2.0 catalog) within the image. The USNO (United States Naval Observatory) catalog was originally obtained by source extraction from the same survey plates as we used in our study. (A catalog is the term commonly applied in astronomy to a relational table, or a tabular array, of coordinate positions followed by object feature measurements;)
2. By counting the number of objects estimated in the image by the Sextractor object detection package (Bertin & Arnouts 1996). As in the case of the USNO catalog, these detections are mainly point sources (i.e. stars, as opposed to spatially extended objects like galaxies);
3. By counting the number of structures detected at several scales using the MR/1 multiresolution analysis package (MR/1 1999).

Figure 11 show the results of plotting these numbers for each image against the multiscale signal entropy of the image. The best results are obtained using the MR/1 package, followed by Sextractor and then by the number of sources extracted from USNO. The latter two basically miss the content at large scales, which is taken into account by MR/1. Unlike MR/1, Sextractor does not attempt to separate signal from noise.

Sextractor and multiresolution methods were also applied to a set of CCD (charge coupled detector, i.e. digital, as opposed to the digitized photographic plates used previously) images from CFH UH8K, 2MASS and DENIS near infrared surveys. Results obtained were very similar to what was obtained above. This lends support to (i) the quality of the results based on MR/1, which take noise and scale into account, and (ii) multiscale entropy being a good measure of content of such a class of images.

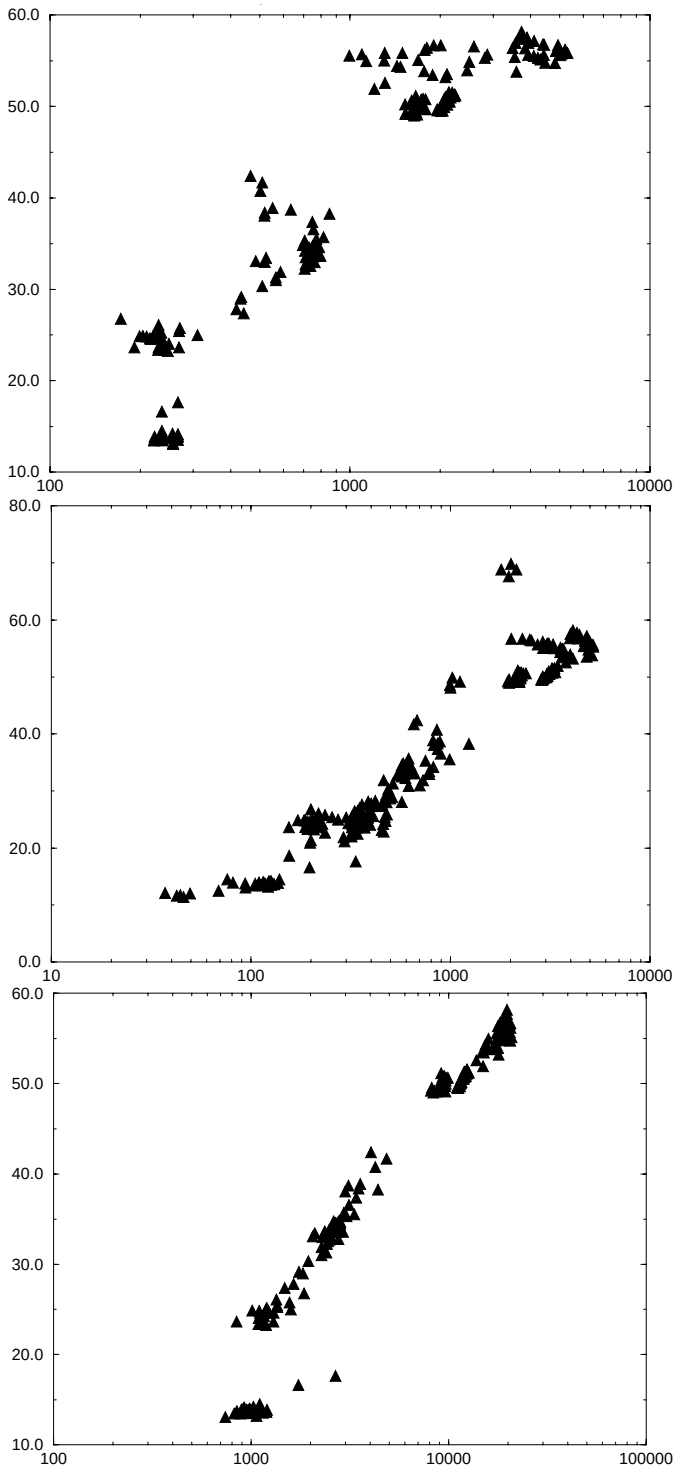

Fig. 11. Multiscale entropy versus the number of objects: the number of objects is, respectively, obtained from (top) the USNO catalog, (middle) the Sextractor package, and (bottom) the MR/1 package

8. Multiscale entropy and optimal compressibility

Subsequently we looked for the relation between the multiscale entropy and the optimal compression rate of an image which we can obtain by multiresolution techniques (Starck et al. 1998a). By optimal compression rate we mean a compression rate which allows all the sources to be preserved, and which does not degrade the astrometry (object positions) and photometry (object intensities). Louys et al. (Louys et al. 1999) and Couvidat (Couvidat 1999) have estimated this optimal compression rate using the compression program of the MR/1 package (MR/1 1999). We note that photometry and astrometry (and visual quality) were used in this work, given the crucial importance in astronomy of flux conservation and completeness. The cited references may be referred to for further details.

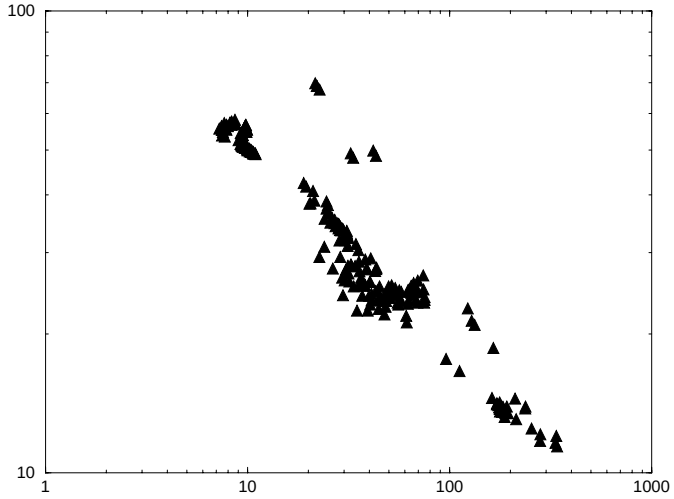

Fig. 12. Multiscale entropy of astronomical images versus the optimal compression ratio. Images which contain a high number of sources have a small ratio and a high multiscale entropy value. With logarithmic numbers of sources, the relation is almost linear

Figure 12 shows the relation obtained between the multiscale entropy and the optimal compression rate for all the images used in our previous tests, both digitized plate and CCD images. The power law relation is obvious thus allowing us to conclude that:

- The compression rate depends strongly on the astronomical content of the image. We can then say that compressibility is also an estimator of the content of the image;
- The multiscale entropy allows us to predict the optimal compression rate of the image.

9. Conclusion

We have seen that information must be measured from the transformed data, and not from the data itself. This is so that a priori knowledge of physical aspects of the data can be taken into account. We could have used the Shannon entropy, perhaps generalized (Sporring & Weickert 1999), to measure the information at a given scale, and derive the bins of the histogram from the standard deviation of the noise, but for several reasons we thought it better to directly introduce noise probability into our information measure. Firstly, we have seen that this leads, for Gaussian

noise, to a very physically meaningful relation between the information and the wavelet coefficients: information is proportional to the energy of the wavelet coefficients normalized by the standard deviation of the noise. Secondly, this can be generalized to many other kinds of noise, including such cases as multiplicative noise, non-stationary noise, or images with few photons/events. We have seen that the equations are easy to manipulate. Finally, experiments have confirmed that this approach gives good results.

For filtering, the multiscale entropy has the following advantages:

- It provides a good trade-off between hard and soft thresholding;
- No a priori model on the signal itself is needed as with other wavelet-based Bayesian methods (Chipman et al. 1997; Crouse et al. 1998; Vidakovic 1998; Timmermann & Nowak 1999);
- It can be generalized to many different noise distributions;
- The regularization parameter α can be easily fixed automatically. Cross-validation (Nason 1996) could be an alternative, but with the limitation to Gaussian noise.

Replacing the standard entropy measurements by the Multiscale Entropy avoids the main problems in the MEM deconvolution method.

We have seen also how our new information measure allows us to analyze image background fluctuation. In the example discussed, we showed how signal which was below the noise level could be demonstrated to be present. Our SNR was 0.25. This innovative analysis leads to our being able to affirm that signal is present, without being able to say where it is.

To study the semantics of a large number of digital and digitized photographic images, we took already prepared – external – results, and we also used two other processing pipelines for detecting astronomical objects within these images. Therefore we had three sets of interpretations of these images. We then used Multiscale Entropy to tell us something about these three sets of results. We found that Multiscale Entropy provided interesting insight into the performances of these different analysis procedures. Based on strength of correlation between Multiscale Entropy and analysis result, we argued that this provided circumstantial evidence of one analysis result being superior to the others.

We finally used Multiscale Entropy to provide a measure of optimal image compressibility. Using previous studies of ours, we had already available to us a set of images with the compression rates which were consistent with the best recoverability of astronomical properties. These astronomical properties were based on positional and intensity information, – astrometry and photometry. Papers cited contain details of these studies. Therefore we had optimal compression ratios, and for the corresponding images, we proceeded to measure the Multiscale Entropy. We found a very good correlation. We conclude that Multiscale Entropy provides a good measure of image or signal compressibility.

The breadth and depth of our applications lend credence to the claim that Multiscale Entropy is a good measure of image or signal content. Compared to previous work, we have built certain aspects of the semantics of such data into our analysis procedures. As we have shown, the outcome is a better ability to understand our data.

Acknowledgements. Work described in Sect. 7 was contributed to by S. Couvidat. We wish to thank F. Bouchet and R. Teyssier for providing us with the Planck simulated dataset. This work was partially supported by USA National Science Foundation grant DMS 98–72890 (KDI).

References

Ables, J. 1974, A&A, 15, 383
Anscombe, F. 1948, Biometrika, 15, 246
Bertin, E., & Arnouts, S. 1996, A&AS, 117, 393
Bijaoui, A. 1984, Introduction au Traitement Numérique des Images (Masson, Paris)
Bonnarel, F., Fernique, P., Genova, F., et al. 1999, Astronomical Data Analysis Software and Systems VIII, ASP Conf. Ser., 172, ed. D. M. Mehringer, R. L. Plante, & D. A. Roberts, ISBN: 1-886733-94-5, 229. 8, 229
Bontekoe, T., Koper, E., & Kester, D. 1994, A&A, 294, 1037
Burg, J. 1978, Annual Meeting International Society Exploratory Geophysics, Reprinted in Modern Spectral Analysis, ed. D. G. Childers (IEEE Press, New York), 34
Chambolle, A., DeVore, R., Lee, N., & Lucier, B. 1998, IEEE Trans. Sig. Proc., 7, 319
Chipman, H., Kolaczyk, E., & McCulloch, R. 1997, J. Am. Stat. Assoc., 92, 1413
Cornwell, T. 1989, in Diffraction-Limited Imaging with Very Large Telescopes, ed. D. Alloin, & J. Mariotti (Kluwer, Dordrecht)
Couvidat, S. 1999, DEA Dissertation, Strasbourg Observatory
Crouse, M., Nowak, R., & Baraniuk, R. 1998, IEEE Trans. Sig. Proc., 46, 886
Donoho, D. 1995, in Wavelets and Statistics, ed. A. Antoniadis, & G. Oppenheim (Springer-Verlag, Berlin)
Epstein, B., Hingorani, R., Shapiro, J., & Czigler, M. 1992, in Data Compression Conference, Snowbird, UT, 200
Ferraro, M., Boccignone, G., & Caelli, T. 1999, IEEE Trans. Patt. Anal. Mach. Intell., 21, 1199
Frieden, B. 1978a, Image Enhancement and Restoration (Springer-Verlag, Berlin)
Frieden, B. 1978b, Probability, Statistical Optics, and Data Testing (Springer-Verlag, Berlin)
Gonzalez, R. C. 1993, Digital Image Processing (Addison-Wesley)
Guibert, J. 1992, in Digitised Optical Sky Surveys, 103+
Gull, S., & Skilling, J. 1991, MEMSYS5 Quantified Maximum Entropy User's Manual
Hotelling, H. 1933, J. Educ. Psychol., 24, 417
Jaynes, E. 1957, Phys. Rev., 106, 171
Karhunen, H. 1947, English translation by I. Selin, On Linear Methods in Probability Theory, The Rand Corporation, Doc. T-131, August 11, 1960
Katsaggelos, A. K. 1991, Digital Image Restoration (Springer-Verlag, Berlin)

Lee, J. 1999, IEEE Trans. Im. Proc., 8, 453
Llacer, J., & Núñez, J. 1990, in The Restoration of HST Images and Spectra, ed. R. White, & R. Allen, Space Telescope Science Institute, Baltimore
Loève, M. 1948, Fonctions Aléatoires de Second Ordre, in P. Levy, Processus Stochastiques et Mouvement Bownien (Hermann, Paris)
Lorenz, H., & Richter, G. 1993, in Science with the Hubble Space Telescope, ed. P. Benvenuti, & E. Schreier, European Southern Observatory, Garching, 203
Louys, M., Starck, J.-L., Mei, S., Bonnarel, F., & Murtagh, F. 1999, A&AS, 136, 579
Mohammad-Djafari, A. 1994, Traitement du Signal, 11, 87
Mohammad-Djafari, A. 1998, Traitement du Signal, 15, 545
Molina, R. 1994, IEEE Trans. Patt. Anal. Mach. Intell., 16, 1122
MR/1 1999, Multiresolution Image and Data Analysis Software Package, Version 2.0, Multi Resolutions Ltd., `www.multiresolution.com`
Murtagh, F. 1998, J. Classif., 15, 161
Narayan, R., & Nityananda, R. 1986, ARA&A, 24, 127
Nason, G. 1996, J. Roy. Stat. Soc. B, 58, 463
Pantin, E., & Starck, J. 1996, A&AS, 315, 575
Saghri, J., Tescher, A., & Reagan, J. 1995, IEEE Sig. Proc. Mag., 12, 32
Shannon, C. 1948, Bell System Tech. J., 27, 379
Skilling, J. 1989, in Maximum Entropy and Bayesian Methods (Kluwer), 45
Sporring, J., & Weickert, J. 1999, IEEE Trans. Info. The., 45, 1051
Starck, J. 1999, Entropie Multiéchelle: Définition et Applications, Habilitation, Université de Paris XI, available at `http://jstarck.free.fr`
Starck, J. & Murtagh, F. 1999, Signal Proc., 76, 147
Starck, J., Murtagh, F., & Bijaoui, A. 1998a, Image Processing and Data Analysis: The Multiscale Approach (Cambridge University Press, Cambridge, UK)
Starck, J., Murtagh, F., & Gastaud, R. 1998b, IEEE Trans. Circ. Syst. II, 45, 1118
Timmermann, K. E., & Nowak, R. 1999, IEEE Trans. Sig. Proc., 46, 886
Tretter, D., & Bouman, C. 1995, IEEE Trans. Im. Proc., 4, 308
Vidakovic, B. 1998, J. Am. Stat. Assoc., 93, 173
Weir, N. 1991, in 3rd ESO/ST-ECF Data Analysis Workshop
Weir, N. 1992, in Astronomical Data Analysis Software and System 1, ed. D. Worral, C. Biemesderfer, & J. Barnes, Astronomical Society of the Pacific, 186