Mon. Not. R. Astron. Soc. **347,** 339–354 (2004)

Maximum-entropy image reconstruction using wavelets

Klaus Maisinger, M. P. Hobson[⋆] and A. N. Lasenby

Astrophysics Group, Cavendish Laboratory, Madingley Road, Cambridge CB3 0HE

Accepted 2003 September 15. Received 2003 August 27; in original form 2003 March 11

ABSTRACT

The maximum-entropy method (MEM) is often used for enhancing astronomical images and, in particular, has recently been applied to cosmic microwave background (CMB) observations. Wavelet functions are also now used widely in astronomy, since they allow the sparse and efficient representation of a signal at different scales, and the application of wavelets to the denoising of CMB maps has been investigated. In this paper, we give a systematic discussion of how to combine these two approaches by the use of the MEM in wavelet bases for the denoising and deconvolution of general images and, in particular, CMB maps. We find that the MEM in the à trous wavelet basis has lower reconstruction residuals than conventional pixel-basis MEM in the case when the signal-to-noise ratio is low and the point spread function is narrow. Furthermore, the Bayesian evidence for the wavelet MEM reconstructions is generally higher for a wide range of images. From a Bayesian point of view, the wavelet basis thus provides a better model of the image.

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

1 INTRODUCTION

Both the maximum-entropy method (MEM) and wavelet techniques are used for astronomical image enhancement. In particular, both methods have recently been applied to the analysis of cosmic microwave background (CMB) data (see, for instance, Hobson, Jones & Lasenby 1999; Sanz et al. 1999a,b; Tenorio et al. 1999). Maps of CMB anisotropies are a useful tool in the analysis of CMB data. Making maps is rarely straightforward, because a multitude of systematic instrumental effects, calibration uncertainties and other deficiencies in the modelling of the telescope come into play. For example, interferometric maps suffer from the incomplete sampling of the telescope in Fourier space and require the deconvolution of the synthesized beam (e.g. Thompson, Moran & Swenson 1994) and the suppression of receiver noise. CMB observations from single-dish telescopes use total power measurements and scan across the observed fields to assemble a map. Here, it is the effect of the finite primary beam that needs to be deconvolved in order that a high-resolution map may be recovered. Beyond the area of the CMB, the task of image reconstruction is generic and occurs in virtually any type of astronomical map-making.

In a general imaging problem, we assume that the data d observed by an experiment are given by a convolution of the true sky signal, or image h, with the point spread function P of the instrument, plus some Gaussian random noise n:

$$d = P * h + n.$$

In the discretized version, the data vector $\boldsymbol{d}$ is given by a multiplication of the vector $\boldsymbol{h}$ of the image pixels with the instrumental response matrix $\mathbf{R}$ that describes the convolution with the point spread function, and the additive noise vector $\boldsymbol{n}$:

$$\boldsymbol{d} = \mathbf{R}\boldsymbol{h} + \boldsymbol{n}.$$

To solve the inverse problem of estimating the image $\boldsymbol{h}$ from the data, some type of regularization is usually required. A common technique is to use an entropic function S for the regularization. The best reconstruction is then found by minimizing the function $F(\boldsymbol{h}) = \frac{1}{2}\chi^2(\boldsymbol{h}) - \alpha S(\boldsymbol{h})$ that determines a suitable trade-off between a good fit to the data enforced by the χ^2-statistic and a strong regularization given by the entropy $S(\boldsymbol{h})$ of the reconstruction. The maximum-entropy method has proven to be very successful for the deconvolution of a wide range of noisy images.

Despite its capabilities, the MEM suffers from several shortcomings. For example, the appropriate entropy functional depends on the properties of the distribution of image pixels, but it is not always evident what the theoretical distribution should be. For positive additive distributions, one uses the entropy

$$S(\boldsymbol{h}) = \sum_{i=1}^{N_h} h_i - m_i - h_i \log \frac{h_i}{m_i}, \qquad (1)$$

where the sum is over all image pixels and m_i is a measure assigned to pixel i. Even in this case, problems can arise when there is no appropriate background level, or if the image brightness falls below the background level in some areas.

Another defect of the MEM is that, in its simplest forms, correlations between image pixels are not taken into account properly. This problem manifests itself in several guises. Because of correlations

[⋆] E-mail: mph@mrao.cam.ac.uk

between image pixels, the effective number of 'degrees of freedom' in the data is often much smaller than the number of parameters in the minimization problem, making effective regularization more difficult. In fact, the MEM is inherently based on the assumption that image pixels are independent. Furthermore, ignoring correlations leads to the introduction of spurious features in the map, such as the characteristic ringing artifact on uniform backgrounds. There is no provision in the MEM algorithm to reward local smoothness of the image. It appears to be quite difficult to regularize in such a way as to reconstruct faithfully sharp features and uniform areas at the same time.

Several solutions have been proposed to remedy the problem of image correlations. In particular, Gull & Skilling (1999) have introduced the concept of an intrinsic correlation function (ICF) that is used to decorrelate the reconstructed image. The ICF framework has been extended to allow reconstructions of objects on different scales. Weir (1992) proposes a multichannel approach, which allows for multiple scales of pixel-to-pixel correlations. In pyramidal maximum entropy (Bontekoe, Koper & Kester 1994), the number of pixels retained in the low-resolution channels is decimated. Despite these improvements, choosing an ICF is not straightforward. It is clear that there is no single set of ICFs that is universally optimal for all possible types of data. Choosing suitable scalelengths and weights is of great importance.

A slightly different approach to tackling the correlation problem is to use a representation of the image that is more efficient in identifying its information content. In other words, the task is to find an optimal *basis set* for the representation of the image. Furthermore, it is desirable to have a representation that can efficiently capture information present on different length-scales in the image. For instance, for CMB observations, several foreground components – such as radio point sources or SZ-clusters – are very localized on the sky. Some theories for structure formation also predict localized non-Gaussian imprints at arcmin scales on the CMB itself – for example, temperature fluctuations produced in the wake of cosmic strings. On the other hand, the primordial CMB itself shows more diffuse structure that peaks on angular scales close to a degree. Representing these signals in real space (i.e. the image plane) requires large numbers of basis functions (i.e. the pixels) for a given image. Similarly, a reconstruction in Fourier space requires the determination of a large number of modes, which are often very poorly constrained by the data, because each localized feature on a map is expanded into an infinite number of basis functions in Fourier space. The MEM has been applied to reconstructions in both real and Fourier space. A Fourier space approach has been developed by Hobson et al. (1998) and has been applied to simulated *Planck* data (see also Stolyarov et al. 2002), while Jones, Hobson & Lasenby (1999) simulate its use for the *Wilkinson Microwave Anisotropy Probe* (*WMAP*) satellite mission.

The application of *wavelets* to CMB data analysis has recently been investigated (see, for instance, Hobson et al. 1999; Sanz et al. 1999a,b; Tenorio et al. 1999; Cayón et al. 2000; Vielva et al. 2001). Wavelets are special sets of functions that allow the efficient representation of signals both in real and in Fourier space. Furthermore, they can represent different objects of greatly varying sizes simultaneously. The term 'wavelet' does not refer to a single unique function. Instead, it comprises a whole class of functions with similar properties. In the context of CMB analysis, wavelets have predominantly been used for noise filtering or the separation of localized foreground sources. A combination of MEM and certain types of wavelets has been discussed by Pantin & Starck (1996) and Starck et al. (2001).

In this paper, we investigate further the transformation of the image reconstruction problem into wavelet bases, and discuss two related methods for obtaining a solution. We also discuss the possible forms for the entropic priors in wavelet bases and methods for determining the appropriate level of regularization. Consideration is also given to how the different approaches can be viewed in the ICF framework. In Section 2, the maximum-entropy method is introduced. This followed in Section 3 by a general discussion of how an inverse problem can be transformed into a new basis to obtain an improved reconstruction. In Section 4, we give an introduction to wavelet transforms, and, in particular, to the à trous algorithm. The use of such a wavelet basis in the transformation of an inverse problem is discussed in Section 5, where we combine wavelets with the MEM. In Section 6, we test the techniques presented by applying them to simulated image reconstruction problems. Our conclusions are presented in Section 7.

2 THE MAXIMUM-ENTROPY METHOD

The task of recovering the original image from blurred and noisy data is a typical inverse problem. In this section, we focus on the maximum-entropy method (MEM) of image reconstruction and give a brief discussion of the background to the standard MEM technique. In Section 3, we highlight the enhancements to the standard method provided by performing reconstructions in alternative bases.

2.1 Inverse problems and regularization

In image reconstruction, the inverse problem consists of estimating the N_h image pixels h_i $(i = 1, \ldots, N_h)$ from the N_d data samples d_i. One may find a solution by optimizing some measure of the goodness of fit to the data (see, for example, Titterington 1985). For Gaussian noise, the χ^2-statistic is the preferred measure:

$$\chi^2(\boldsymbol{h}) = (\mathsf{R}\boldsymbol{h} - \boldsymbol{d})^{\mathrm{t}}\mathsf{N}^{-1}(\mathsf{R}\boldsymbol{h} - \boldsymbol{d}). \tag{2}$$

By minimizing χ^2, the vector $\hat{\boldsymbol{h}}$ that best fits the data can be found.

In image reconstructions, however, the number N_h of parameters is the number of image pixels, which can be very large. In this case, the parameter estimates are usually poorly constrained by the data, and overfitting leads to a bad reconstruction of the original image. In order to avoid overfitting and wildly oscillating solutions, some kind of additional information or regularization is required. The regularization is achieved by an additional function $S(\boldsymbol{h})$ which penalizes 'roughness' in the image. The choice of $S(\boldsymbol{h})$ is determined by what exactly one considers as roughness. A compromise between the goodness of fit $\chi^2(\boldsymbol{h})$ and the regularization $S(\boldsymbol{h})$ can then be found by minimizing the function

$$F(\boldsymbol{h}) = \frac{1}{2}\chi^2(\boldsymbol{h}) - \alpha S(\boldsymbol{h}), \tag{3}$$

where α is a Lagrange multiplier which determines the degree of smoothing. As α is varied, solutions lie on a trade-off curve between optimal fit to the data and maximal smoothness (see, for example, Titterington 1985).

2.2 Bayes' theorem and the entropic prior

The discussion above can be more coherently expressed in a Bayesian framework. Bayes' theorem states that the conditional probability $\Pr(\boldsymbol{h}\,|\,\boldsymbol{d})$ for a hypothesis $\boldsymbol{h}$ to be true given some data $\boldsymbol{d}$, the so-called *posterior* probability, is given by

$$\Pr(\boldsymbol{h}\,|\,\boldsymbol{d}) = \frac{\Pr(\boldsymbol{d}\,|\,\boldsymbol{h})\Pr(\boldsymbol{h})}{\Pr(\boldsymbol{d})}. \tag{4}$$

The probability $\Pr(\boldsymbol{d}\,|\,\boldsymbol{h})$ is called the *likelihood* of the data, and $\Pr(\boldsymbol{h})$ is the *prior* probability of the hypothesis. It can be used to incorporate our prior beliefs or expectations on possible solutions. The probability $\Pr(\boldsymbol{d})$, the *evidence*, only depends on the data and can, for the time being, be viewed as a normalization constant.

For Gaussian-distributed errors on the data points, the likelihood is then given by

$$\Pr(\boldsymbol{d}|\boldsymbol{h}) = \frac{1}{(2\pi)^{N_d/2}\sqrt{|\mathsf{N}|}} \exp\left[-\frac{1}{2}(\mathsf{R}\boldsymbol{h}-\boldsymbol{d})^t\mathsf{N}^{-1}(\mathsf{R}\boldsymbol{h}-\boldsymbol{d})\right], \quad (5)$$

where $\chi^2(\boldsymbol{h}) = (\mathsf{R}\boldsymbol{h}-\boldsymbol{d})^{\mathrm{t}}\mathsf{N}^{-1}(\mathsf{R}\boldsymbol{h}-\boldsymbol{d})$ is the standard misfit statistic introduced in (2). If the parameters are well constrained by the data, i.e. the likelihood function is narrow, the posterior probability will be sufficiently peaked and the errors on the parameters $\boldsymbol{h}$ will be small. Unfortunately, that is usually not the case in image reconstruction, where the number of parameters or image pixels is large. One then has to make use of the prior $\Pr(\boldsymbol{h})$ to obtain a solution.

There are many possible choices for the prior, but it is usually of the exponential form (e.g. Skilling 1989)

$$\Pr(\boldsymbol{h}) \propto \exp\left[\alpha S(\boldsymbol{h})\right], \quad (6)$$

where $S(\boldsymbol{h})$ is a regularization function and α is a constant. Assuming a likelihood function given by (5), the posterior probability from (4) then becomes

$$\Pr(\boldsymbol{h}|\boldsymbol{d}) \propto \exp\left[-\frac{1}{2}\chi^2(\boldsymbol{h}) + \alpha S(\boldsymbol{h})\right]. \quad (7)$$

Clearly, the optimal choice of the regularization has to reflect our knowledge of the expected solution (see Frieden 1983). A list of commonly used regularization functions can be found in Titterington (1985). These are either functions that are quadratic in the hypothesis $\boldsymbol{h}$ or that use some kind of logarithmic entropy.

An efficient prior is provided by the (Shannon) information entropy of the image. If one restricts oneself to images whose pixel values are strictly positive, the image can be considered as a positive additive distribution (PAD). In this case, the appropriate entropy function is (1), which is a generalization of the Shannon entropy; the entropy functions for more general images are discussed in Section 3.3. In (1), the measure $\boldsymbol{m}$ is often called the *model*, because the entropy is maximized by the default solution $h_i = m_i$ $(i = 1, \ldots, N_h)$. Given an entropy function $S(\boldsymbol{h})$, the posterior probability can be maximized by minimizing its negative logarithm

$$-\ln[\Pr(\boldsymbol{h}|\boldsymbol{d})] = F(\boldsymbol{h}) = \frac{1}{2}\chi^2(\boldsymbol{h}) - \alpha S(\boldsymbol{h}). \quad (8)$$

This is the maximum entropy method (MEM; see, for example, Gull 1989; Skilling 1989; Gull & Skilling 1999). The minimization is usually performed using some form of local downhill search algorithm for which one must calculate the gradient $\boldsymbol{g} \equiv \nabla_{\boldsymbol{h}} F$; some minimization algorithms also require one to evaluate the curvature (or Hessian) matrix of the functional given by $\mathsf{H} \equiv \nabla_{\boldsymbol{h}}\nabla_{\boldsymbol{h}} F$. A discussion of the calculation of derivatives is given in Appendix A.

2.3 Nuisance parameters and the evidence

In (8), we have ignored that fact that probabilities are implicitly conditional on the values of a set of nuisance (hyper)parameters, such as the constant α and the model $\boldsymbol{m}$ in the prior (and, formally, the assumed response matrix R and noise covariance matrix N in the likelihood). Thus, the posterior in (7) should more properly be written as $\Pr(\boldsymbol{h}\,|\,\boldsymbol{d}, \alpha, \boldsymbol{m}, \ldots)$, and similarly for the likelihood, prior and evidence in Bayes' theorem.

The formal method for removing such nuisance parameters from one's analysis is to marginalize over them. Thus, assuming for convenience that α and $\boldsymbol{m}$ are the only nuisance parameters, one should properly maximize the posterior

$$\Pr(\boldsymbol{h}|\boldsymbol{d}) = \int \Pr(\boldsymbol{h}|\boldsymbol{d}, \alpha, \boldsymbol{m})\Pr(\alpha, \boldsymbol{m}|\boldsymbol{d})\mathrm{d}\alpha\mathrm{d}\boldsymbol{m}. \quad (9)$$

It is clear, however, that this approach is prohibitively expensive in terms of computation for image reconstruction problems. Thus, it is usual instead to assign to the nuisance parameters 'optimal' values, which maximize their posterior distribution $\Pr(\alpha, \boldsymbol{m}\,|\,\boldsymbol{d})$. Using Bayes' theorem again, we may write this posterior as

$$\Pr(\alpha, \boldsymbol{m}|\boldsymbol{d}) = \frac{\Pr(\boldsymbol{d}|\alpha, \boldsymbol{m})\Pr(\alpha, \boldsymbol{m})}{\Pr(\boldsymbol{d})}. \quad (10)$$

However, if the evidence $\Pr(\boldsymbol{d}\,|\,\alpha, \boldsymbol{m})$ is sufficiently peaked, it will overwhelm any priors on α and $\boldsymbol{m}$, in which case one may simply maximize the value of the evidence to determine the optimal 'Bayesian' values of the nuisance parameters.

In the MEM, it is unusual to determine both the regularization constant α and the model $\boldsymbol{m}$ by maximizing the evidence $\Pr(\boldsymbol{d}\,|\,\alpha, \boldsymbol{m})$ simultaneously with respect to both. Most commonly, in image reconstruction, the model $\boldsymbol{m}$ is simply set uniformly across the image to the value of the expected image background (although, in a more refined analysis, the evidence can be used to discriminate between models). With a fixed model, it is then reasonably straightforward to determine a Bayesian value of the regularization constant α by maximizing the evidence with respect to it, as we discuss below.

2.4 The regularization parameter α

The parameter α introduced in (6) determines the amount of regularization on the image. It is clear that minimizing χ^2 only (by setting $\alpha = 0$) would lead to a closer agreement with the data, and thus to noise fitting. On the other hand, maximizing the entropy alone by setting $\alpha = \infty$ would lead to an image which equals the default model $\boldsymbol{m}$ everywhere. Indeed, for every choice of α, there is an image $\hat{\boldsymbol{h}}(\alpha)$ corresponding to the minimum of $F(\boldsymbol{h})$ for that particular choice. The images $\hat{\boldsymbol{h}}(\alpha)$ vary along a trade-off curve as α is varied.

There are several methods for assigning an optimal value to α. In early MEM applications, α was chosen such that for the final reconstruction $\hat{\boldsymbol{h}}$, the misfit statistic χ^2 equalled its expectation value $\chi^2(\hat{\boldsymbol{h}}) = N_d$, i.e. the number N_d of data values. This choice is often referred to as *historic MEM*. It can be shown that it leads to systematic underfitting of the data (Titterington 1985). However, an optimum value of α can be assigned within the Bayesian framework itself, as discussed above, by maximizing the evidence (Gull 1989; Hobson et al. 1998; Gull & Skilling 1999). In particular, by approximating the posterior $\Pr(\boldsymbol{h}\,|\,\boldsymbol{d}, \alpha, \boldsymbol{m})$ by a multivariate Gaussian in $\boldsymbol{h}$-space about its peak, one may derive a simple implicit relation for the Bayesian estimate $\hat{\alpha}$ of the regularization parameter. One finds that α must satisfy

$$-2\alpha S[\hat{\boldsymbol{h}}(\alpha)] = N_h - \alpha\mathrm{Tr}(\mathsf{M}^{-1}), \quad (11)$$

where N_h is the number of pixels in the reconstruction. The $N_h \times N_h$ matrix M is given by

$$\mathsf{M} = \mathsf{G}^{-1/2}\mathsf{H}\mathsf{G}^{-1/2}, \quad (12)$$

where $\mathsf{H} = \nabla_{\boldsymbol{h}}\nabla_{\boldsymbol{h}} F$ is the Hessian matrix of the functional (8) and $\mathsf{G} = -\nabla_{\boldsymbol{h}}\nabla_{\boldsymbol{h}} S$ is the negative Hessian matrix of the entropy functional, both evaluated at the peak $\hat{\boldsymbol{h}}(\alpha)$.

This choice of α is called *classic maximum entropy* (Gull 1989). It may also be shown that this Bayesian value of α may be reasonably approximated by choosing its value such that the value of $F(\boldsymbol{h})$ at its minimum is equal to half the number of data points, i.e. $F \approx N_d/2$ (MacKay 1992).

2.5 Errors on maximum-entropy estimates

There are two principal methods of quantifying errors on maximum-entropy reconstructions. Firstly, one may evaluate the Hessian matrix $\mathsf{H} = \nabla_{\boldsymbol{h}}\nabla_{\boldsymbol{h}} F$ at the optimal solution $\hat{\boldsymbol{h}}$, from which the covariance matrix of the parameters is given by $\mathsf{C} = \mathsf{H}^{-1}$. However, this typically requires the inversion of a large matrix, which is unfortunately non-diagonal and so requires a large computational effort. Secondly, once one has estimated the Hessian matrix H, one may approximate the posterior as the corresponding multivariate Gaussian centred on the optimal solution. Then, by taking 'samples' from this approximate posterior probability distribution, one can quantify the error on any particular parameter h_i (or combinations of parameters) by examining how h_i varies between samples from the distribution (Gull & Skilling 1999).

3 TRANSFORMING AN INVERSE PROBLEM

3.1 Hidden-space MEM

In applying the MEM to an image reconstruction problem, it is usually the case that the vector of parameters $\boldsymbol{h}$ simply comprises the pixel values in the reconstructed image. There is, however, no requirement in the method itself that this be so. Indeed, in many cases, a much improved reconstruction can be obtained by choosing the parameters $\boldsymbol{h}$ to be the coefficients of some other (pixelized) basis set that is more efficient at describing the image under consideration.

In this approach, one takes the reconstructed image (or 'visible space'), which we now denote by $\boldsymbol{v}$, to be related to the parameters $\boldsymbol{h}$ (or 'hidden space') by some linear operator K_f, traditionally termed the ICF:

$$\boldsymbol{v} = \mathsf{K}_\mathrm{f}\boldsymbol{h}. \tag{13}$$

The predicted noiseless data can then be calculated using the response matrix R and is given by $\boldsymbol{d} = \mathsf{R}\boldsymbol{v} = \mathsf{R}\mathsf{K}_\mathrm{f}\boldsymbol{h}$. Therefore, instead of reconstructing the image vector directly, we may continue to work entirely in terms of the hidden image $\boldsymbol{h}$.

Let us consider more closely the relationship between the data and the visible and hidden image vectors. The $N_\mathcal{D}$-dimensional data vector $\boldsymbol{d}$ is an element of the data space $\mathcal{D}$. Similarly, the visible vector $\boldsymbol{v} \in \mathcal{V}$, $\dim \mathcal{V} = N_\mathcal{V}$ and the hidden vector $\boldsymbol{h} \in \mathcal{H}$, $\dim \mathcal{H} = N_\mathcal{H}$. The data vector is then given by the transform

$$\begin{tikzcd}
\boldsymbol{h} \arrow[r, mapsto] \arrow[d, phantom, "\in"] & \boldsymbol{v} \arrow[r, mapsto] \arrow[d, phantom, "\in"] & \boldsymbol{d} \arrow[d, phantom, "\in"] \\
\mathcal{H} \arrow[r, "\mathsf{K}_\mathrm{f}"] & \mathcal{V} \arrow[r, "\mathsf{R}"] & \mathcal{D},
\end{tikzcd} \tag{14}$$

where the K_f is the ICF (the subscript 'f' simply denotes that the transformation is in the 'forward' direction – we shall consider transformations in the opposite, or 'backward', direction in Section 3.2). It is important to note that $N_\mathcal{H}$ and $N_\mathcal{V}$ need *not* be equal, in which case K_f is rectangular and hence is not invertible. We also note here that there is no reason, in general, for the data space $\mathcal{D}$ and visible space $\mathcal{V}$ to be identical. For example, Bridle et al. (1998) and Marshall et al. (2002) have applied maximum entropy to the reconstruction of lensing mass profiles of galaxy clusters from shear or magnification data. In this case, $\mathcal{D}$ and $\mathcal{V}$ do not even share the same physical dimensions. However, in image reconstruction, the data image and the reconstruction are often not only given in the same physical units but on the same discretized grid, and we have $\mathcal{D} = \mathcal{V}$. In this case, it is possible to incorporate information gleaned directly from the data into the choice of the regularization.

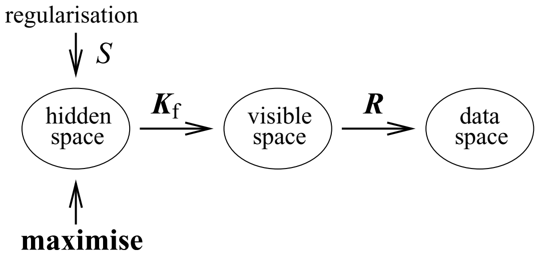


Figure 1. Schematic depiction of hidden-space MEM. The data are predicted by a composition of the ICF K_f and the instrumental response R. The entropic regularization is applied to the hidden space, and the posterior functional is maximized with respect to the hidden coefficients.

The theoretical framework of the MEM itself suggests that, in most cases, some form of ICF *should* be used. One of the fundamental axioms on which the MEM is based is that it should not by itself introduce correlations in the reconstructions (Gull & Skilling 1999). In other words, the parameters $\boldsymbol{h}$ should be a priori uncorrelated. However, it is often the case that there is reason a priori to believe that, for a specific application, the reconstructed quantities are correlated. For example, in the reconstruction of CMB fluctuations, we often have specific prior knowledge of the correlations between pixels.

If such a priori statistical knowledge of the correlation structure of the image exists, then this may be straightforwardly exploited within the MEM to improve the quality of the reconstruction. If the covariance matrix $\mathsf{C} = \langle v_i v_j^\dagger \rangle$ is known a priori, the ICF can be constructed straightforwardly by the Cholesky decomposition $\mathsf{C} = \mathsf{L}\mathsf{L}^\mathrm{t}$ and setting $\mathsf{K}_\mathrm{f} = \mathsf{L}$ (note that, in this case, K_f is square and so $N_\mathcal{H} = N_\mathcal{V}$). Unfortunately, the full correlation structure of an image is usually not known in advance, and a suitable ICF has to be chosen on empirical or heuristic grounds.

The structure of a hidden-space MEM algorithm is illustrated in Fig. 1. The χ^2 $(\boldsymbol{v})$-function is naturally defined on the space $\mathcal{V}$ of visible vectors $\boldsymbol{v}$, whereas the regularizing entropy function $S(\boldsymbol{h})$ is defined on $\mathcal{H}$. Thus, to obtain the optimal hidden-space vector, one maximizes with respect to $\boldsymbol{h}$ the functional

$$F(\boldsymbol{h}) = \frac{1}{2}\chi^2(\mathsf{K}_\mathrm{f}\boldsymbol{h}) - \alpha S(\boldsymbol{h}).$$

The resulting hidden image $\hat{\boldsymbol{h}}$ is then transformed into the optimal visible-space vector $\hat{\boldsymbol{v}}$ (or reconstructed image) using (13).

A hidden-space algorithm of this sort is, in fact, the most straightforward way to implement a transformed MEM. Such an implementation is given, for example, in the widely used software package MEMSYS5 (Gull & Skilling 1999). The main reason for the relative simplicity of this approach is that one need only replace the response matrix R with the combined operation $\mathsf{R}\mathsf{K}_\mathrm{f}$ throughout. Thus, a pre-existing MEM implementation is easily modified. Another desirable feature of a hidden-space approach is that both the minimization and regularization are performed in the same space. In particular, this means that both the gradient vector and the curvature matrix of the regularizing term $\alpha S(\boldsymbol{h})$ in (8) take simple forms. Most notably, for any of the entropic regularization functions discussed in Section 3.3, the curvature matrix of the regularizing term is diagonal. Because the curvature matrix of the misfit term $\chi^2(\mathsf{K}_\mathrm{f}\boldsymbol{h})$ is independent of the parameters $\boldsymbol{h}$, and so need only be calculate once, one may therefore calculate the curvature matrix of the full functional F straightforwardly. This has significant advantages in determining the errors on the reconstruction and enables the straightforward evaluation of the Bayesian evidence, and hence the Bayesian value for the regularization parameter α.

Depending on the relative values of $N_\mathcal{H}$ and $N_\mathcal{V}$, a hidden-space algorithm may require minimization in a space of dimensionality that is either smaller or larger than the visible space of the image itself. For the applications discussed in Section 6, $N_\mathcal{H} > N_\mathcal{V}$, but we find that the larger dimensionality of the minimization does not present a problem for the cases considered, and can be performed in a similar amount of CPU time as the corresponding visible-space minimization.

3.2 Hidden-space-regularized MEM

An alternative approach to transforming an inverse problem is instead to retain the definition of F in terms of the visible space quantities $\boldsymbol{v}$, and to perform only the regularization in some hidden space $\boldsymbol{h}$. In this method, one defines the hidden space in terms of the visible space by some linear operator, which we denote by K_b, such that

$$\boldsymbol{h} = \mathsf{K}_\mathrm{b}\boldsymbol{v}. \tag{15}$$

Comparing this expression with (13), we note that, for the case in which $N_\mathcal{H} = N_\mathcal{V}$ and K_f is invertible, then one may choose $\mathsf{K}_\mathrm{b} = \mathsf{K}_\mathrm{f}^{-1}$. In general, however, this is not possible; indeed, the two linear operators need not be related at all. In particular, the transform K_b is not an ICF in the same sense as discussed in Section 3.1, since the parameters that are reconstructed by the MEM, the visible image pixels, are still correlated.

We call this alternative approach *hidden-space-regularized* MEM, and an illustration of the method is shown in Fig. 2. In this case, one minimizes the functional

$$F(\boldsymbol{v}) = \frac{1}{2}\chi^2(\boldsymbol{v}) - \alpha S(\mathsf{K}_\mathrm{b}\boldsymbol{v})$$

from which one sees that the entropy $S(\mathsf{K}_\mathrm{b}\boldsymbol{v})$ can be viewed as a function on $\mathcal{V}$ and defines a new effective 'hidden entropy'. We note that, even in the case in which $\mathsf{K}_\mathrm{b} = \mathsf{K}_\mathrm{f}^{-1}$, hidden-space-regularized MEM is *not* equivalent to hidden-space MEM because of the non-linearity of the entropy functional. In Appendix B, we show in this case that both methods only become equivalent if a quadratic regularization function is used.

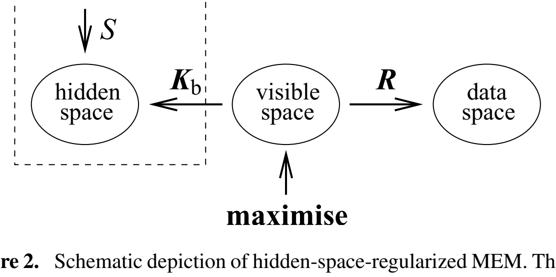


Figure 2. Schematic depiction of hidden-space-regularized MEM. The entropic regularization is applied to the hidden coefficients obtained from the visible (image) space, but the posterior functional is maximized with respect to the visible (image) pixels. The transform and entropy functional act as a new effective 'hidden entropy'.

One advantage of the hidden-space-regularized MEM is that the dimensionality of the minimization is always $N_\mathcal{V}$, irrespective of the dimensionality of the hidden space. Nevertheless, we have found in practical applications that this does not have significant effect on the CPU time required to locate the optimum. A significant disadvantage of this approach is that the regularization and optimization are performed in different spaces. This leads to more complicated forms for the gradient vector and curvature matrix of the regularization term $\alpha S(\mathsf{K}_\mathrm{b}\boldsymbol{v})$. Most importantly, the curvature matrix is no longer diagonal, which makes it significantly more computationally demanding to calculate evidences and errors on the reconstruction.

3.3 The entropy function

In both of the methods outlined above, the regularization is performed on the hidden-space vector $\boldsymbol{h}$ rather than the visible-space image $\boldsymbol{v}$. Thus, even if the image has only positive pixel values, the hidden-space parameters $\boldsymbol{h}$ may take positive, negative or even complex values depending on the nature of K_f or K_b.

From (1), for a positive additive distribution $\boldsymbol{h}$, the cross-entropy $S_+(\boldsymbol{h}, \boldsymbol{m})$ of the image $\boldsymbol{h}$ with some model $\boldsymbol{m}$ of the image is given by the sum

$$S_+(\boldsymbol{h}, \boldsymbol{m}) = \sum_i s_+(h_i, m_i),$$

where

$$s_+(h, m) = h - m - h \log \frac{h}{m}. \tag{16}$$

The function $s_+(h, m)$ reaches a global maximum of zero at $h = m$. Thus, in the absence of data, the reconstruction takes on the default value m for all pixels; in practice, the value of m is often set to the level of the expected background in hidden space. If one is dealing with image reconstructions $\boldsymbol{v}$ that are strictly positive, then it can be advantageous to choose K_f or K_b, so that positivity is maintained in the hidden space, and use (16) as the regularizing function.

In general, however, either the image itself or the corresponding hidden space may take positive consist of both positive and negative pixels. The entropy (16) is then inapplicable to such images. A simple generalization of (16) to negative values is

$$s_{|\cdot|}(h, m) = |h| - m - |h| \log \frac{|h|}{m}, \tag{17}$$

which has been used by Pantin & Starck (1996). This function does not have a unique maximum, since both $h = \pm\,|\,m\,|$ maximize the entropy. For $h \to 0$, one finds $s_{\,|\,\cdot\,|}(h) \to -m$. However, the entropy is not defined at zero, which is generally the expected default state of the image if the mean has been subtracted. In practice, the model values m have to be close to zero and small compared to any realistic data in order to avoid the introduction of a spurious background signal. Pantin & Starck (1996) use the value $m = k\sigma$, where σ is the rms of the data and $k = 1/100$ is an arbitrarily chosen, small constant.

The proper way to extend the entropy (16) to images that take both positive and negative values is the entropy definition (Hobson & Lasenby 1998; Gull & Skilling 1999)

$$s_\pm(h, m) = \Psi - 2m - h \log \frac{\Psi + h}{2m}, \tag{18}$$

where $\Psi = \sqrt{h^2 + 4m^2}$. The entropy $s_\pm$ has a maximum at $h = 0$. In the positive/negative entropy (18), the role of the model m is different from that in (16). The value of m determines the width of the entropy function and thus controls the magnitude of the allowed deviations

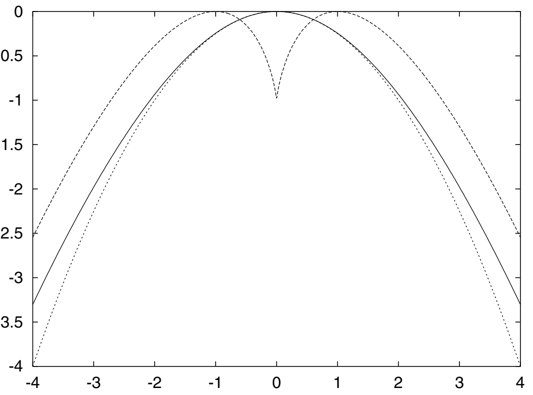


Figure 3. The entropy function $s_{\pm}(h) = \Psi - 2m - h\log[(\Psi + h)/(2m)]$ (solid line), its quadratic approximation $s(h) = -h^2/(4m)$ (short dashed line) and $s_{|\cdot|}(h) = |h| - m - |h|\log(|h|/m)$ (long dashed line). In all cases, the model was chosen to be $m = 1$. The function $s_{|\cdot|}$ is not defined at $h = 0$.

from the default value. Hence the model can be considered as a level of 'damping' imposed on the image. From a dimensional analysis, the obvious choice for m is to set it to the expected signal rms. If data and visible space are identical and the signal-to-noise ratio is sufficiently high, the rms of the observed data can be a good approximation to that of the signal.

Fig. 3 shows the entropy functions $s_{|\cdot|}$ and $s_{\pm}$. From the definitions (17) and (18), we see that, for a given model, both functions differ only by a constant offset in the limit of large h: $s_{|\cdot|}(h) \to s_{\pm}(h) + m\ (h \to \infty)$. However, in order to minimize the cusp of $s_{|\cdot|}$ around zero, the models will generally be chosen differently, and the maximum of $s_{|\cdot|}$ will be significantly narrower than that of $s_{\pm}$.

Expanding $s_{\pm}$ around zero,

$$s_{\pm}(h, m) = \sum_{n=1}^{\infty} \frac{\left[\prod_{i=1}^{n}(2i-3)\right]^2 (-1)^n}{(2m)^{2n-1}} \frac{h^{2n}}{(2n)!}$$
$$= -\frac{h^2}{2(2m)} + \frac{h^4}{4!(2m)^3} - \frac{3^2 h^6}{6!(2m)^5} + \frac{(3\times 5)^2 h^8}{8!(2m)^7} - \cdots$$
$$= -\frac{h^2}{4m} + \mathcal{O}(h^4), \tag{19}$$

one obtains the quadratic approximation

$$s_2(h) = -\frac{h^2}{4m}, \tag{20}$$

which is plotted in Fig. 3. With this entropy, MEM reduces to a scaled least-squares. The quadratic entropy can also be obtained from the large-α limit $h \simeq m$ of the standard entropy $s_+(h)$, but with a different scale factor $1/2m$. In a MEM reconstruction, one evaluates the product αS of the entropy S and a regularization constant α. The constant α is not dimensionless; its dimension $[\alpha] = 1/[h]$ is given by the dimension $[h]$ of h. If the model is chosen proportional to the signal rms σ_S, then, from (19), the product αs becomes invariant under a rescaling of h if α is also rescaled by $1/\sigma_S$. In fact, to first order, any change in the model m can be absorbed by a reciprocal change in the regularization constant α:

$$\alpha s \propto -\frac{\alpha}{m} h^2. \tag{21}$$

This explains why $s_{|\cdot|}$ produces reconstructions similar to those obtained for $s_{\pm}$ despite its narrow shape for small models. It should be remembered, however, that the quadratic approximation is only formally applicable when the value of the parameter h is small. For large values of h, the quadratic approximation provides a much stronger regularization than the proper expression (18). We note that in a more recent work, Starck et al. (2001) use a quadratic approximation of the same form as s_2 but with the model set to the variance of the noise on the corresponding hidden-space variable. Starck et al. (2001) also propose a different alternative form for the regularizing function (their equation 40), which is claimed to possess better properties.

4 THE WAVELET TRANSFORM

In this paper, we explore the use of the wavelet transform in converting between the visible and hidden spaces in a transformed MEM algorithm. To simplify our discussion, we begin by considering the wavelet transform in one dimension. Rather than consider some function $f(x)$ defined at all points x, it is more relevant to our discussion to restrict our attention to some digitized form, which we represent by the column vector $\boldsymbol{f}$ (whose length N must an integer power of 2). The discrete wavelet transform (DWT), like the discrete Fourier transform (DFT), is a linear operation that transforms $\boldsymbol{f}$ into another vector $\widetilde{\boldsymbol{f}}$ that contains the wavelet coefficients of the (digitized) function. The action of the DWT can therefore be described as a multiplication of the original vector by the $N \times N$ wavelet matrix W:

$$\widetilde{\boldsymbol{f}} = \mathsf{W}\boldsymbol{f}. \tag{22}$$

The original basis vectors $\boldsymbol{e}_i$ have unity as the ith element and the remaining elements equal to zero, and hence correspond to the pixels in the original image, the coefficients of which are contained in the vector $\boldsymbol{f}$. Therefore the original basis is the most localized basis possible in real space. For the DFT, the new basis vectors $\widetilde{\boldsymbol{e}}_i$ are (digitized) complex exponentials and represent the opposite extreme, since they are completely non-local in real space but localized in frequency space. For the DWT, the new basis vectors are the wavelets, which enjoy the characteristic property of being fairly localized both in real space and in frequency space, thus occupying an intermediate position between the original pixel basis and the Fourier basis of complex exponentials. Indeed, it is the simultaneous localization of wavelets in both spaces that makes the DWT such a useful tool for analysing data in wide range of applications.

In many applications, such as data compression or spectral decomposition, it is common to use wavelet bases that are orthogonal and non-redundant (i.e. the number of wavelet coefficients equals the number of points at which the original function is sampled). In this case, the wavelet transform matrix has useful property of being orthogonal, so that $\mathsf{W}^{\mathrm{t}} = \mathsf{W}^{-1}$. In image reconstruction, however, it is well-known that redundant transforms produce much better results (Lang et al. 1996). Moreover, Langer, Wilson & Anderson (1993) have suggested that non-orthogonal, translationally and rotationally invariant (*isotropic*) wavelet transforms are better suited to the task of image reconstruction than orthogonal ones. Bearing these observations in mind, in this paper we concentrate on the use of the à trous algorithm, which we now discuss.

4.1 The à trous wavelet transform

The à trous ('with holes') algorithm (Holschneider et al. 1989; Bijaoui, Starck & Murtagh 1994), is a non-orthogonal, redundant discrete wavelet transform that is widely used in image analysis. Starting from a data vector $c_{J,l}$ ($l = 1, \ldots, N$, $2^J \leqslant N$), iteratively smoothed vectors are obtained by

$$c_{j-1,k} = \sum_l H_l c_{j,k+2^{(J-j)}l}, \tag{23}$$

where each step is effectively a convolution of the image with the filter mask H_l using varying step sizes 2^{J-j}. At each scale, the detail wavelet coefficients contain the difference between the smoothed image $c_{j-1,k}$ and the image $c_{j,k}$ at the previous scale:

$$w_{j-1,k} = c_{j,k} - c_{j-1,k}.$$

Since no decimation is carried out between consecutive filter steps, the à trous transform is redundant. Thus the final wavelet transformed vector has length $J \times N$. The inverse à trous transform is simply the sum over the coefficients at all scales:

$$c_{J,l} = c_{0,l} + \sum_{j=0}^{J-1} w_{j,l}.$$

4.1.1 *Properties of the à trous transform*

As discussed above, the à trous transform constructs the wavelet coefficients by successive applications of the same filter mask with different spacings between pixels, followed by a subtraction of the smooth component in each step. This procedure is computationally efficient because of the compactness of the mask, as the wavelet coefficients in each domain can be constructed by a sum over only a small fraction of the coefficients on the previous scale. Perhaps not entirely obvious from this construction is that this algorithm is equivalent to a convolution of the original image with a series of point spread functions, which not only have different widths, but can also assume slightly different shapes on different scales. Wavelet coefficients are then given by the convolution

$$w_{j-1,k} = \sum_l \psi_{j-1,l-k} c_{J,l} \tag{24}$$

of the input vector and the wavelet function $\psi_{j-1}(-x)$ at the $(j-1)$th scale. For a symmetric wavelet $\psi_{j-1}(x) = \psi_{j-1}(-x)$, this is identical to a convolution with the wavelet itself. Popular choices for the corresponding scaling function $\phi(x)$ are the triangle function

$$\phi(x) = \begin{cases} 1 - |x| & \text{if} \quad x \in [-1, 1] \\ 0 & \text{otherwise} \end{cases} \tag{25}$$

or a B_3-spline (Starck, Murtagh & Bijaoui 1998).

The effective point spread functions for the triangular scaling function (25) are shown in Fig. 4. The horizontal axis denotes pixel offsets. The first wavelet scale is given by a convolution with the narrow wavelet function $\psi(x)$ (dotted line). The next scale is given by the same function, but scaled by a factor of 2 in width and 1/2 in height (dash–dotted line). The higher scales are produced from increasingly wider convolution masks. Finally, the last scale is the smooth component obtained from a convolution with the broad scaling function $\phi(x)$ (solid line).

Fig. 5 shows the effective point spread functions for the B_3-spline given by the convolution mask

$$\left(\frac{1}{16}, \frac{1}{4}, \frac{3}{8}, \frac{1}{4}, \frac{1}{16}\right).$$

Note how the wavelet functions become smoother as their width increases relative to the pixel resolution.

We note that, for the à trous transform, the normalization condition $\int |\psi(x)|^2\,\mathrm{d}x = 1$ does not hold. This means, for instance, that the transform of Gaussian white noise does not have a constant dispersion in all wavelet domains.

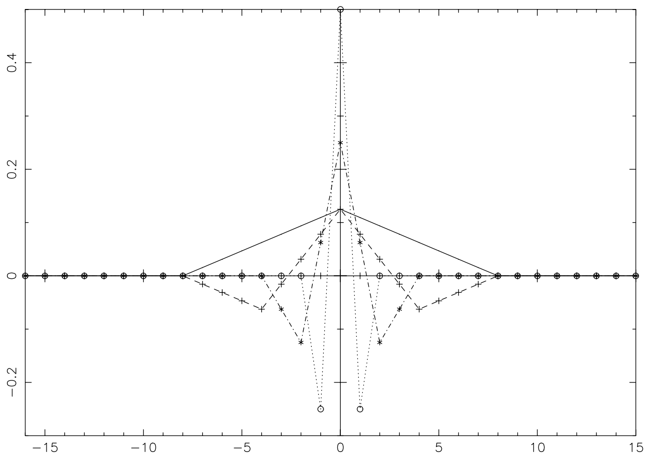


Figure 4. The effective point spread functions of the à trous transform for the triangular scaling function (25). The wavelet coefficients at each scale are given by convolutions with the increasingly wider wavelet functions (dotted line, dash–dotted line, dashed line). The smooth component is given by a convolution with the scaling function (solid line).

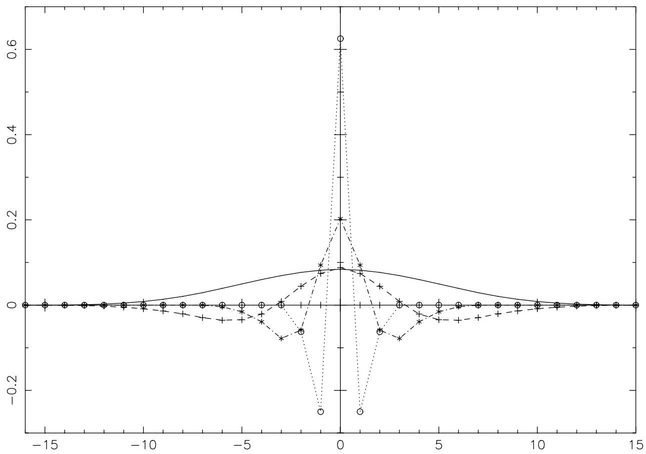


Figure 5. The effective point spread functions of the à trous transform for a B_3-spline. The wavelet coefficients at each scale are given by convolutions with the increasingly wider wavelet functions (dotted line, dash–dotted line, dashed line). The smooth component is given by a convolution with the scaling function (solid line). Compare the triangular mask in Fig. 4.

4.1.2 *The à trous transform as a Fourier filter*

The convolution functions shown in Figs 4 and 5 act as filters corresponding to different spatial frequency bands. The corresponding window functions can be obtained from a Fourier transform of the convolution masks. The window functions for the triangular transform from Fig. 4 are plotted in Fig. 6. The detail coefficients are given by the high-pass filter extending to the large Fourier modes (dotted line with circles). The coarser scales are given by filter functions moving to smaller and smaller Fourier modes. The smooth component can be obtained from the low-pass filter centred around the smallest frequencies (solid line), or longest wavelengths. The fact that the inverse transform is simply given by a sum of the detail scales and the smooth components ensures that all filters add up to

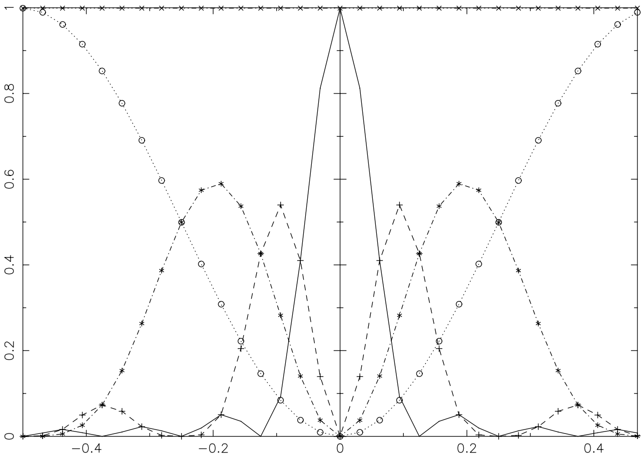


Figure 6. The window functions for the triangular transform (25) at different scales in Fourier space. These functions have been obtained by a Fourier transform of the functions in Fig. 4. The horizontal scale denotes spatial frequencies.

a constant rectangular window of unit height (crosses on top of the diagram). The corresponding window functions for a B_3-spline look very similar, but their sidelobes are much smaller.

4.2 Two-dimensional wavelet transforms

The non-orthogonal, redundant à trous transform lends itself particularly well to generalizations to higher dimensions. The convolution mask for the two-dimensional à trous transform can be obtained from a two-dimensional rotation of the one-dimensional wavelet function. One of the most appealing features of the à trous algorithm is that it does not single out any special direction in the image plane. In this sense, the à trous transform is isotropic. Furthermore, the à trous transform is also invariant under translations (the transform of the dilated function is simply the dilation of the transform). In practice, the convolution mask for the two-dimensional à trous transform can even be obtained from a simple product of two one-dimensional masks, while retaining the above properties to a sufficient approximation.

5 MEM IN A WAVELET BASIS

We are now in a position to combine the MEM with the wavelet transform. In particular, we will consider the two alternative transformed MEM algorithms discussed in Section 3.

5.1 Wavelet MEM

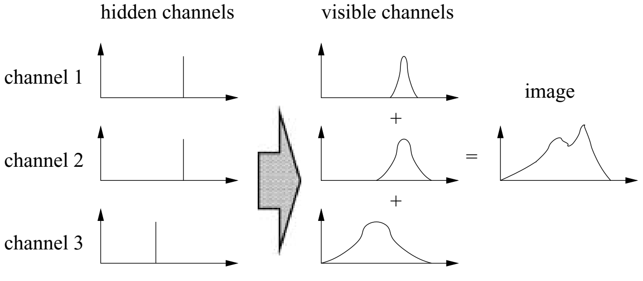


Figure 7. The multichannel ICF. The hidden image consists of several images or channels (left). The visible channels are obtained by convolutions of the hidden parameters with point spread functions of different widths (middle). The visible image is a weighted sum of the image channels (right).

Let us begin by considering the use of the wavelet transform in the hidden-space approach of Section 3.1. In this case, it is most convenient to take the intrinsic correlation function to be the transpose of the wavelet transform, i.e. $\mathsf{K}_\mathrm{f} = \mathsf{W}^\mathrm{t}$. For orthogonal wavelet transforms, $\mathsf{W}^\mathrm{t} = \mathsf{W}^{-1}$, and this choice of K_f ensures that the transformation $\boldsymbol{v} \mapsto \boldsymbol{h}$ is given by the wavelet transform. Thus, the hidden space simply consists of the wavelet coefficients of the reconstruction. For non-orthogonal transforms such as the à trous transform, the relation $\mathsf{W}^\mathrm{t} = \mathsf{W}^{-1}$ does not hold. Nevertheless, choosing $\mathsf{K}_\mathrm{f} = \mathsf{W}^\mathrm{t}$ enables the straightforward evaluation of the derivatives of F, since the transpose of K_f is required for the evaluation of the gradient. This is why our convention is a better choice than $\mathsf{K}_\mathrm{f} = \mathsf{W}^{-1}$ for non-orthogonal transforms. Furthermore, as we show below, the use of the transpose has a very simple interpretation for the non-orthogonal à trous transform. We call the resulting approach the *wavelet MEM* algorithm.

5.1.1 *Multichannel ICF*

The advantage of transforming the inverse problem to wavelet space in this way is that it allows for a natural multiresolution description of the image. Indeed, our approach may be be considered in the context of the multichannel ICF proposed by Weir (1992), which we now discuss.

Traditionally the ICF K_f is taken to be a convolution with some point spread function $P(x)$. Thus the visible image is a blurred version of the hidden variables. Popular point spread functions comprise B-splines of various orders (Gull & Skilling 1999) or Gaussians (e.g. Weir 1992). One deficiency of the 'blurring' ICF is that the width of the point spread function introduces a characteristic scale for the pixel correlations. In most applications, however, objects and correlations of varying sizes and scales are present in the same image. A straightforward generalization of the blurring ICF is the multichannel ICF (Weir 1992). This method is illustrated in Fig. 7. The hidden variables consist of a number of different images or channels $\boldsymbol{h}_i$. Each hidden channel is blurred with a different point spread function $\mathsf{K}_\mathrm{f}^{(i)}$ such that $\boldsymbol{v}_i = \mathsf{K}_\mathrm{f}^{(i)}\boldsymbol{h}_i$. The visible image $\boldsymbol{v}$ is obtained by a weighted sum $\boldsymbol{v} = \sum_i w_i \boldsymbol{v}_i$ of all blurred image channels $\boldsymbol{v}_i$, where the weight on the kth channel is denoted by w_i. The entropy function is applied to the hidden variables. If the weights on different scales are chosen appropriately, it is entropically favourable to represent extended structure in the visible image by a single coefficient (or very few) in the hidden domain.

One weakness of this approach is that it introduces a large number of new free parameters, like the widths of the point spread functions and the weighting factors w_k of the different channels (and a complete new hidden image for each scale). If there is a priori knowledge on the expected correlation scales of objects, the width of the point spread function can be chosen accordingly. However, this will rarely be the case. The reconstructions can become strongly sensitive to the chosen set of ICF widths, the number of channels, the weights and the models in different channels (Bontekoe et al. 1994). In the 'pyramidal MEM' (Bontekoe et al. 1994), the different channels contain different numbers of hidden parameters. With decreasing resolution, the number of parameters in each consecutive channel is reduced by a factor of a half in each image dimension. Bontekoe et al. (1994) find empirically that uniform models and weights (i.e. $w_k = 1$ for all channels k) and the same point spread function (scaled by factors of 2) can be used for all channels. They also note that the pyramid images can be interpreted as spatial bandpass filters.

5.1.2 The à trous transform as an ICF

If one writes the non-orthogonal à trous transform as a series of convolutions with scaling or wavelet functions, as discussed in Section 4.1, one can show that, for the à trous transform, the ICF $\mathbf{K}_{\mathrm{f}} = \mathbf{W}^{\mathrm{t}}$ used in wavelet MEM is just a special case of a multichannel ICF (see, for example, Rue & Bijaoui 1997). From (24), the à trous wavelet coefficients w_i^j at the *j*th scale are given by the (discretized) convolution

$$w_i^j = \sum_k W_{ik}^j f_k = \sum_k \psi_{k-i}^j f_k \tag{26}$$

of the original function or image $\boldsymbol{f}$ with a version of the wavelet ψ^j at the *j*th scale that was mirrored at the origin. For a symmetric wavelet function, this corresponds to a convolution with the wavelet itself. Examples of wavelet functions are shown in Figs 4 and 5. [For an orthogonal wavelet, the equivalent to (26) would be $w_i = \sum_k \psi_{k-2i} h_k$, where the factor of 2 in the index accounts for the decimation carried out between consecutive filter steps.] The transpose of the à trous transform operates on the wavelet coefficients and produces a new image $\boldsymbol{g}$, which is given by

$$g_k = \sum_{j,i} W_{ki}^j w_i^j = \sum_{j,i} \psi_{k-i}^j w_i^j . \tag{27}$$

Thus the transpose consists of a convolution of the *j*th wavelet domain with the corresponding wavelet ψ^j and a subsequent summation over all scales. The operation of the à trous transform and of its transpose is illustrated in Fig. 8.

The effect of the transpose of the à trous transform is just the same as that of a multichannel ICF. Its set of ICFs consists of the scaling function and the rescaled wavelet functions. Unlike in the standard multichannel approach, the blurring functions take negative values in some areas. As in Section 5, one can introduce different weights for all channels in the à trous transform, or one can simply adapt the default model on different scales to enhance the reconstruction quality. However, if one introduces scale-dependent weights w_k, the inverse transform also needs to be rescaled. The nice property that the Fourier transforms of all wavelet functions add up to 1 is lost.

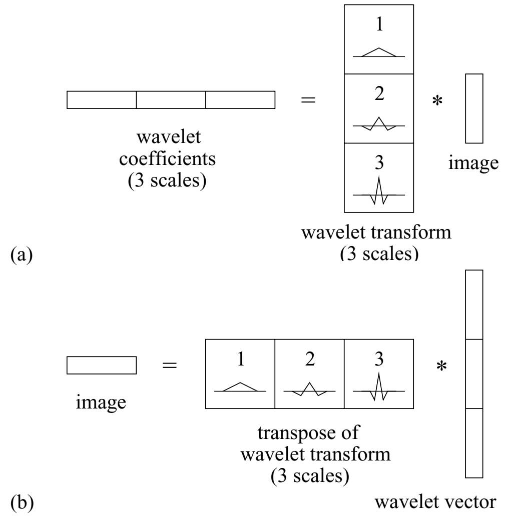


Figure 8. (a) The operation of a three-level à trous transform on an image. (b) The transpose of the à trous transform operates on a wavelet vector with three scales. Each scale is convolved with the corresponding wavelet, and the resulting vector is obtained by a sum over all scales.

5.2 Wavelet-regularized MEM

The wavelet transform may also be used in the hidden-space-regularized MEM algorithm discussed in Section 3.2. In this case, it is most convenient to choose $\mathbf{K}_{\mathrm{b}} = \mathbf{W}$, so that the hidden-space vector $\boldsymbol{h}$ simply contains the wavelet coefficents of the visible-space image $\boldsymbol{v}$. This method has been pursued by, for instance, Pantin & Starck (1996) and Starck et al. (2001). In fact, the entropy $S(\mathbf{W}\boldsymbol{v})$ can be viewed as a function on $\mathcal{V}$, and its proponents call it the 'multiresolution entropy'. We will call this approach *wavelet-regularized MEM*.

5.3 Choosing the regularization parameter and the model

In both wavelet and wavelet-regularized MEM, the entropic regularization is performed on the wavelet coefficients. The choice of the regularization parameter α has been discussed in general in Section 2.4. For wavelet MEM, we use the 'classic' Bayesian choice of α, as implemented in the software package MEMSYS5. For wavelet-regularized MEM, however, the maximization takes place in visible space and the Hessian of the effective entropy is not diagonal, which means it can not be implemented in MEMSYS5. In this case, we will simply employ the 'historic' maximum entropy criterion $\chi^2 = N_{\mathcal{D}}$ to fix α.

We note that, instead of setting a constant α globally, Pantin & Starck (1996) suggest choosing a regularization parameter α that is different for each scale *j* in the wavelet domain, or even one that varies from one wavelet coefficient to another. This can be achieved by introducing, for each pixel in the hidden space, an additional weighting factor α_i that is given by

$$\alpha_i \propto \sigma_{\mathrm{N}}^j (1 - M_i). \tag{28}$$

In this expression, the factor $\sigma^j{}_{\mathrm{N}}$ is the noise dispersion in the *j*th wavelet domain (i.e. the domain to which the *i*th wavelet coefficient h_i belongs) and M_i is the multiresolution support defined by

$$M_i = \begin{cases} 1 & \text{if} \quad |h_i| \geqslant k\sigma_{\mathrm{N}}^j \\ 0 & \text{if} \quad |h_i| < k\sigma_{\mathrm{N}}^j \end{cases} . \tag{29}$$

A common choice for the threshold factor in this expression is $k = 3$. The factor $(1 - M_i)$ in (28) reduces the regularization of coefficients with a high signal-to-noise ratio, and the factor $\sigma^j{}_{\mathrm{N}}$ introduces an additional pixel-dependent regularization. The pixel-dependent factors α_i are scaled by the global parameter α. From (21), however, we see that this procedure is, to first order, equivalent to a making corresponding change in the model value m_i, and so can be achieved by adopting a non-constant model, which we now discuss.

In real-space MEM, there is often no a priori reason to assign a varying model to particular image regions, because in the absence of any additional information one would expect a constant signal dispersion across the image. The wavelet transform, however, is designed to enhance the signal-to-noise ratio on some wavelet coefficients to allow a sparse representation of the signal. Consequently, one would expect different signal dispersions in each wavelet domain, and a uniform choice of the model seems appropriate only within domains. If data and image space are identical, the signal dispersion in the *j*th wavelet domain σ_{S}^j can be estimated from the dispersion of the data σ_{D}^j and of the noise σ_{N}^j:

$$\sigma_{\mathrm{S}}^j = \sqrt{\sigma_{\mathrm{D}}^{j2} - \sigma_{\mathrm{N}}^{j2}}.$$

Now the model can be set to $m_i = \sigma_{\mathrm{S}}^j$, where the ith pixel lies in the jth wavelet domain. If no analytic noise model is available, the noise dispersion can be obtained from Monte Carlo simulations.

6 APPLICATION TO SIMULATED DATA

In this section, we compare the two wavelet basis maximum-entropy methods discussed above by applying them to some simulated two-dimensional images. One of the problems of assessing the capabilities of different reconstruction methods is that there is no single unique criterion for the quality of a reconstruction. Furthermore, the outcome of any quality measurement depends on a large number of variables, such as the type and content of the image, properties of the instrument with which the data were observed (e.g. the point spread function and the noise) and the model and regularization constant used in the maximum-entropy algorithms, etc. In this section, we use the rms differences between the original image and the reconstruction as a measure of the reconstruction quality. Of course, in real applications, one does not have the original image available for comparison, but nevertheless one would usually hope to minimize the expected difference between true and reconstructed image. In Section 6.1, we use an observation of Saturn made by the *Voyager 2* spacecraft as our test image, which contains features typical of many astronomical images. In Section 6.2, we then turn to the investigation of the reconstruction of CMB maps that are realizations obtained from an inflationary CDM model, and provide a useful astronomical test image that is complementary in its properties to the Saturn image.

6.1 Saturn image

As a first test image, we use the high-resolution *Voyager 2* image of Saturn shown in Fig. 9(a). The image contains 256 × 256 pixels; it has had its mean subtracted and has been rescaled to an rms of 1. It is highly non-Gaussian, and its sharp lines and contrasting continuous extended regions provide useful characteristics to assess the visual impression of different reconstruction methods. We simulate observations of this image by convolving it with different Gaussian point spread functions with FWHMs of 3, 5 and 10 pixels and adding Gaussian white noise with an rms of 0.1, 0.5 or 2. Bearing in mind that the original image had unity rms, this corresponds to signal-to-noise ratios of 10, 2 and 0.5. For each FWHM and noise level, we use Monte Carlo simulations of 15 different noise realizations. One of the realizations obtained for a FWHM of 5 pixels and a noise level of 0.5 is shown in Fig. 9(b).

To each simulation, we apply several different reconstruction algorithms. First, we use a real-space MEM algorithm whose image model is set to be constant across the image plane to the value of the estimated signal rms. The regularization parameter α discussed in Section 2.4 is chosen from the historic MEM criterion which demands that the χ^2-statistic for the final reconstruction equal the number $N_{\mathcal{D}}$ of data points – in this case, 256 × 256. We also apply the wavelet MEM and wavelet-regularized MEM implementations using à trous wavelets derived from the triangle function (25) with four levels. In all cases, the model is chosen to be constant within a given wavelet domain. The model values for each domain or scale are found by setting them to the expected signal dispersions, as discussed in Section 6.3. Fig. 10 plots data, signal and noise

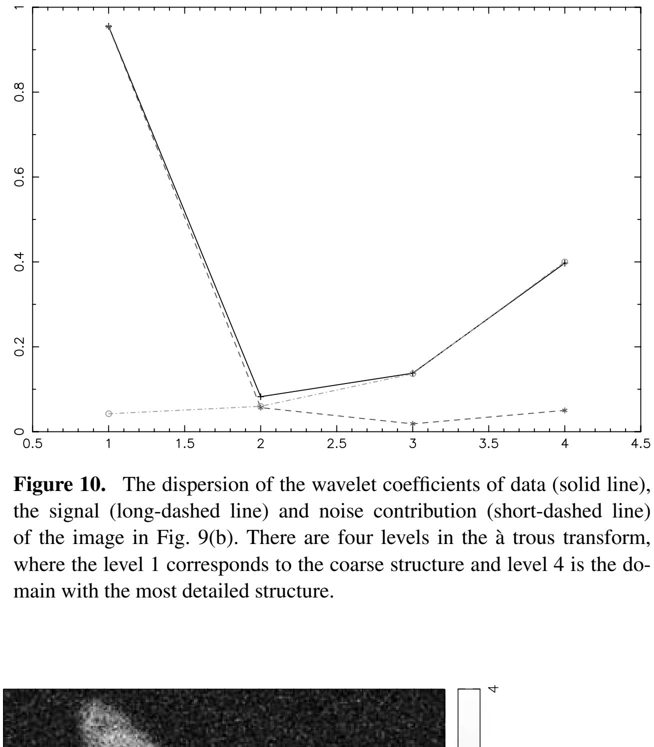


Figure 10. The dispersion of the wavelet coefficients of data (solid line), the signal (long-dashed line) and noise contribution (short-dashed line) of the image in Fig. 9(b). There are four levels in the à trous transform, where the level 1 corresponds to the coarse structure and level 4 is the domain with the most detailed structure.

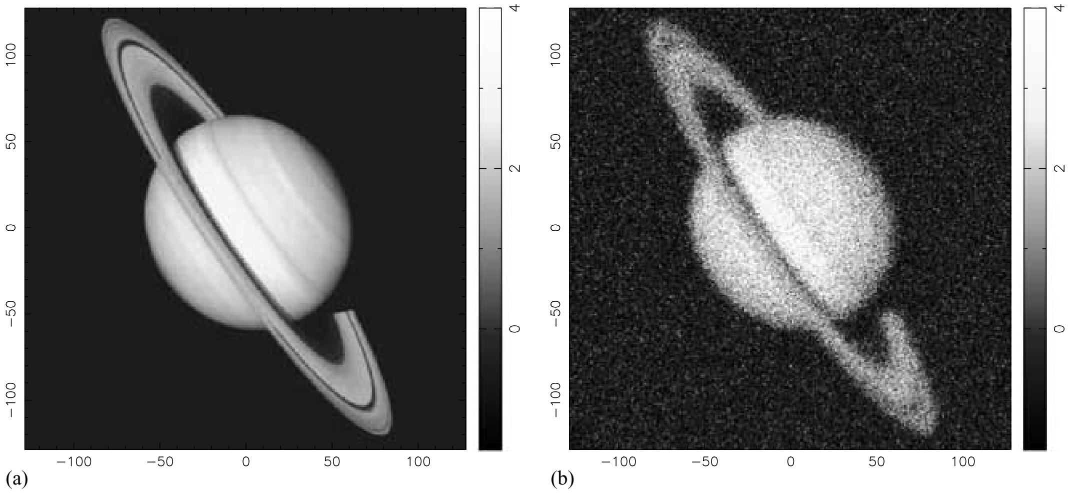


Figure 9. The original image (a) used in the simulations described in the text, and the 'data image' (b) obtained after a convolution with a Gaussian point spread function with a FWHM of 5 pixels and the addition of Gaussian random noise whose rms is half the rms of the original image.

dispersions for the four-level à trous transform. One can clearly see that on coarse scales (level 1), the data are dominated by the signal, and on the finest scales (level 4), they are dominated by the noise.

The reconstructions produced from the 'data' in Fig. 9(b) are presented in Fig. 11. The first image (a) shows the results from real-space MEM. The image is dominated by the 'ringing' or little speckles that are characteristic for the real-space method. Visually, this image is clearly the poorest of the three MEM reconstructions. The second image (b) was produced using the wavelet MEM algorithm with a four-level à trous transform, and is visually more appealing than the real-space reconstruction. The image is very smooth, but owing to the high noise levels on the data, there is little deconvolution of the point spread function. The third image (c) was produced using the wavelet-regularized MEM algorithm, again with a four-level à trous transform. This image is marginally less visually appealing than image (b); again, no real attempt at a deconvolution is apparent.

For a more quantitative comparison of the different methods, Table 1 lists the reconstruction errors for different FWHMs of the point spread function (3, 5 and 10 pixels) and different noise levels (0.1, 0.5 and 2). The quoted errors are the rms differences between the original image and the reconstruction averaged over 15 different noise realizations. The standard deviation of the error values is usually much less than 0.01 for low noise values and not more than 0.03–0.04 in the high-noise case, so the error on the mean (of the errors) is expected to be not more than 0.01 in most cases. There are several trends apparent from the numbers, as follows.

(i) For large point spread functions, all methods yield similar results (except perhaps for the real-space MEM at low signal-to-noise).

(ii) For high signal-to-noise, the methods also have a similar performance.

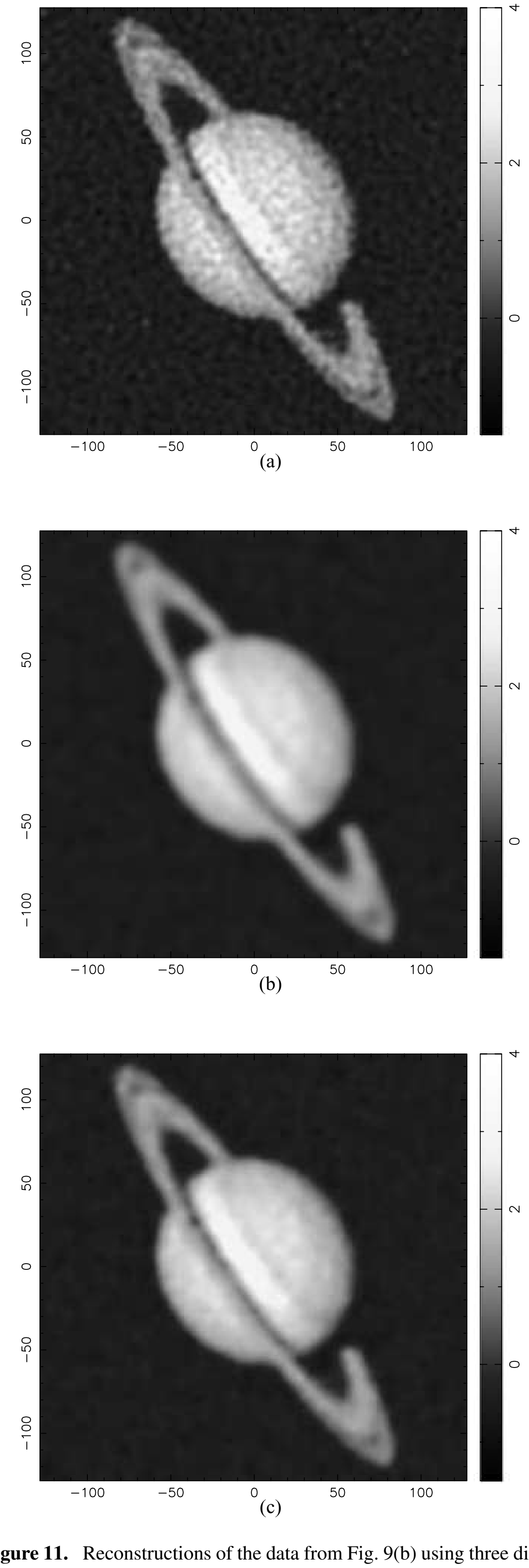
(iii) For sufficiently narrow point spread functions and poor signal-to-noise ratios, real-space MEM performs clearly worse than wavelet MEM and wavelet-regularized MEM.

(iv) In all cases, wavelet MEM performs as well as or better than wavelet-regularized MEM.

(v) In the case when the point spread function is narrow and the noise is dominant, it is most difficult for the MEM algorithms to distinguish between signal and noise, and the performance differences become most prominent.

Within the wavelet algorithms, there are of course many free parameters that may influence the reconstruction errors. We find, however, that for the à trous algorithm, the triangle function is equally efficient as the B_3-spline, and that the number of levels in the transform has only marginal effects on the reconstruction quality, provided there are more than three or four. We have also tested different models for the wavelet coefficients. Generally, there are many models suppressing power on small wavelet scales that provide a similar performance. A stronger relative penalty on small-scale structure, for example, by using the variance instead of the dispersion of the signal, can be advantageous for poor signal-to-noise ratios, because most of the fine structure will be noise. We find no benefits from methods that attempt a more data-dependent regularization – for instance, using the regularization constant (28).

Tracing back the reconstruction errors to the wavelet domain, it appears that, not surprisingly, the dominant contribution comes from those domains where the signal-to-noise is close to unity. The other domains are either perfectly reconstructed (the coarse structure) or contribute little to the image (the noise-dominated small-scale structure).



(a)

(b)

(c)

Figure 11. Reconstructions of the data from Fig. 9(b) using three different MEM algorithms: (a) real-space MEM; (b) wavelet MEM using a four-level à trous transform; (c) wavelet-regularized MEM using a four-level à trous transform. In each case, the value of the regularization constant α is fixed using the 'historic' criterion $\chi^2(\hat{\boldsymbol{h}}) = N_{\mathcal{D}}$.

Table 1. Reconstruction errors for different historic MEM algorithms as a function of the FWHM (in image pixels) of the convolution mask and of the noise level on the data. The numbers quote the rms differences between the reconstruction and the original image averaged over a number of noise realizations as described in the text.

PSF FWHM		3			5			10	
noise rms	0.1	0.5	2.0	0.1	0.5	2.0	0.1	0.5	2.0
real-space MEM	0.11	0.29	0.62	0.11	0.23	0.49	0.16	0.21	0.37
wavelet MEM	0.09	0.16	0.34	0.11	0.17	0.32	0.15	0.20	0.31
wavelet-regularized MEM	0.09	0.18	0.42	0.11	0.17	0.35	0.15	0.20	0.32

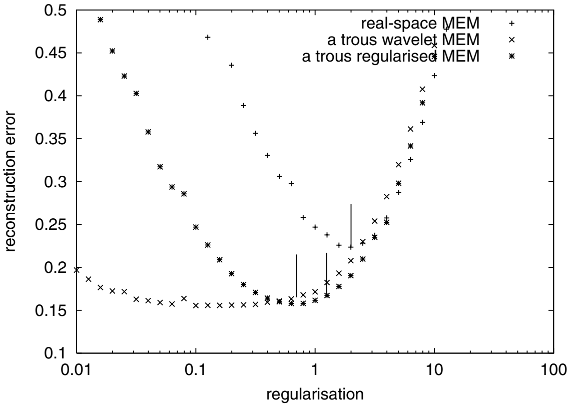


Figure 12. The rms reconstruction errors as a function of the regularization constant α, for a simulation with a five-pixel FWHM blurring function and a signal-to-noise of 2. For real-space MEM, wavelet MEM and wavelet-regulared MEM using à trous wavelets, the plotted points show the final errors for a reconstruction with a fixed value of α. For each method, the small lines indicate the value of α that would have been obtained from the historic MEM criterion $\chi^2 = N_\mathcal{D}$.

Table 2. Reconstruction errors as in Table 1, but for a Bayesian choice of the regularization constant α (classic MEM). It is not possible to implement such a choice of α straightforwardly for wavelet-regularized MEM, and so the corresponding reconstruction errors are not given.

PSF FWHM		3			5			10	
noise rms	0.1	0.5	2.0	0.1	0.5	2.0	0.1	0.5	2.0
real-space MEM	0.48	0.64	0.89	0.43	0.56	0.77	0.37	0.49	0.65
wavelet MEM	0.13	0.23	0.37	0.12	0.18	0.32	0.16	0.23	0.38

6.1.1 The influence of the regularization constant α

The selection of the regularization constant α is one of the major problems in the maximum entropy method, as already discussed in Sections 2.4 and 5.3. By setting the χ^2-statistic to the number of data points, $\chi^2 = N_\mathcal{D}$, as we did for the above reconstructions, one essentially fixes the 'goodness of fit', which may result in very similar error levels on the reconstructions independent of the basis functions or regularization. Fig. 12 illustrates the dependence of the reconstruction error on the regularization constant α for real-space MEM, wavelet MEM and wavelet-regularized MEM using the à trous wavelet. These results are obtained from a simulation with a FWHM of 5 pixels and a noise level of 0.5. Each point corresponds to the rms difference between the final reconstructions and the original image when the regularization parameter α has been fixed to the given value. One can clearly see that there is a trade-off curve between too close a fit to the data (and thus to spurious noise) and too strong a regularization (and thus to suppression of real structure in the image). Both extremes result in high reconstruction errors. For each reconstruction method, a short horizontal line marks the point along the curve that is preferred by the $\chi^2 = N_\mathcal{D}$ criterion. The following two features of these curves are worth pointing out.

(i) The global minimum of the curves for the wavelet algorithms is indeed lower than that for real-space MEM. This means that the lower errors of the wavelet method quoted in Table 1 are not merely artefacts of a poorly chosen regularization constant. On the contrary, for the real-space algorithm, the historic MEM criterion actually picks out points that are very close to the global minimum of the curve, despite the visual ringing evident from the image.

(ii) The curves for the wavelet algorithms are less narrowly peaked than that for real-space MEM. This means that the quality of the reconstruction is less sensitive to α. For the wavelet-regularized method, one can vary α by about an order of magnitude around the minimum without significantly affecting the reconstruction errors. For wavelet MEM, α can vary by over 2 orders of magnitude.

6.1.2 Bayesian α and relative evidences

The proper Bayesian way to determine the regularization constant α is to treat it as an additional model parameter and marginalize the posterior probability over α, i.e. integrate over all possible values of α (see Section 2.4). Because the evidence is often strongly peaked, it is usually sufficient to maximize the evidence to find the optimal value of α. The reconstruction errors for MEM with a Bayesian α (using MEMSYS5) are shown in Table 2. It is evident that for the real-space MEM, the reconstruction errors are very poor compared with those obtained with the historic MEM criterion in Table 1. This is expected from the curve in Fig. 12: the historic MEM criterion picks a value of α that is close to the minimum of the trade-off curve. The reconstruction errors can only get worse for a different choice of α. For comparison with Fig. 12, the Bayesian value for the real-space MEM is $\alpha = 0.13$. In fact, it is this poor performance of the real-space method that historically leads to the introduction of ICFs and the search for a better image model (e.g. MacKay 1992). The wavelet MEM ($\alpha = 0.04$ in Fig.12) improves the reconstruction errors dramatically. Nevertheless, for narrow point spread functions, the à trous transform tends to allow too much image structure on small scales. However, by reweighting the individual scales of the transform suitably, the reconstructions can be much improved. The optimal weighting factors will be investigated in future work.

The evidence for classic real-space and wavelet MEM are presented in Table 3. The values are the logarithms ln Pr($\boldsymbol{d}$) of the

Table 3. The logarithm ln Pr(***d***) of the evidence Pr(***d***) for a classic MEM reconstruction with a Bayesian choice of the regularization constant α (classic MEM). It is not possible to implement a Bayesian choice for α straightforwardly for wavelet-regularized MEM, and so the corresponding evidences are not given. The logarithms are normalized such that they equal zero for the real-space MEM for each dataset.

PSF FWHM		3			5			10	
noise rms	0.1	0.5	2.0	0.1	0.5	2.0	0.1	0.5	2.0
real-space MEM	0	0	0	0	0	0	0	0	0
wavelet MEM	17345	6468	1675	6994	2910	872	1467	626	203

Table 5. Reconstruction errors between reconstructions and original CMB maps. The errors are averaged over a set of 25 simulations with five different noise and image realizations. In each case, the regularization parameter α is fixed using the 'Bayesian' criterion. It is not possible to implement such a choice of α straightforwardly for wavelet-regularized MEM and so the corresponding reconstruction errors are not given.

PSF FWHM		3			5			10	
noise rms	0.1	0.5	2.0	0.1	0.5	2.0	0.1	0.5	2.0
real-space MEM	0.47	0.64	0.87	0.45	0.58	0.76	0.50	0.60	0.74
wavelet MEM	0.28	0.43	0.64	0.28	0.39	0.55	0.42	0.50	0.62

evidence Pr(***d***), as introduced in Section 2. The values have been normalized such that for a given FWHM and noise rms, the evidence for the real-space MEM equals 1. It is apparent that the evidence for the wavelet MEM is significantly higher than for the real-space MEM, as would be expected from the improved reconstruction errors. From a Bayesian point of view, the wavelet basis is therefore the favoured model for the reconstruction of this image. Unfortunately, a similar calculation of the evidence for the wavelet-regularized MEM is not possible, because the software package MEMSYS5 that is used for the reconstruction is limited to regularization functionals whose curvature matrix is diagonal.

6.2 CMB maps

Having applied the wavelet MEM techniques to a general image, we now turn to the reconstruction of CMB maps. We apply the different methods to five different CMB maps that are realizations obtained from the inflationary cold dark matter (CDM) model. The map size is 16 × 16 deg^2 with 256 × 256 pixels, corresponding to a pixel size of 3.75 arcmin. The maps are convolved with Gaussian beams of different beam sizes with FWHMs of 11.25 arcmin (3 pixels), 18.75 arcmin (5 pixels) and 37.5 arcmin (10 pixels), respectively. Gaussian white noise is added with different signal-to-noise ratios of $\sigma_{\rm S}/\sigma_{\rm N} = 10$, 2 and 0.5. This range of simulated observations includes the typical observational parameters characteristic of both the *WMAP* and *Planck* satellites for 12 months of observation. In particular, for *WMAP*, $\sigma_{\rm S}/\sigma_{\rm N} \approx 2$ and the FWHM of the beams at 40, 60 and 90 GHz are 32, 21 and 14 arcmin, respectively. For *Planck*, $\sigma_{\rm S}/\sigma_{\rm N} \approx 15$ and the FWHM of the beams at 33, 44, 70 and 100 GHz are 33, 23, 14 and 10 arcmin, respectively.

For each signal-to-noise ratio, we create five different noise realizations. With the five different input CMB maps, there are thus 25 different combinations of sky and noise realizations available for a given FWHM and noise level. The maps are reconstructed on the same grid as the simulated data. Again, we use the historic MEM criterion to determine the regularization constant instead of classic MEM, because it produces sufficiently good reconstruction errors and is also easily applicable to wavelet-regularized MEM.

The reconstruction errors for different methods are quoted in Table 4. Because of the wealth of information present on all scales in the CMB maps, the errors are generally larger than for the photographic image (compare Table 1). Although the differences in reconstruction errors are less conspicuous, the results confirm the conclusions drawn from the Saturn reconstructions. For low signal-to-noise ratios and narrow point spread functions, the wavelet-based methods are superior to real-space MEM. However, there is no significant difference between wavelet-regularized and wavelet methods.

For the same CMB simulation, we also performed real-space and wavelet MEM reconstructions using the 'Bayesian' value for the regularization parameter α. The corresponding reconstruction errors are quoted in Table 5. We see that, for the real-space MEM, the reconstruction errors are somewhat worse than those for the reconstructions produced using the historic criterion for α. For the wavelet MEM, however, the reconstruction errors are similar to the historic case, illustrating again that wavelet MEM is more robust to the choice of the regularization parameter.

For illustration, one of the realizations of the CMB maps used in the simulations is shown in Fig. 13(a). A simulated observation of this map using a point spread function with a FWHM of 5 pixels and Gaussian random noise of 0.5 is shown in Fig. 13(b). Reconstructions using real-space MEM and wavelet MEM with the Bayesian α criterion are shown in (c) and (d), respectively. The reconstruction quality of the real-space MEM is clearly visibly inferior compared with the wavelet MEM.

The averaged logarithms ln Pr(***d***) of the Bayesian evidence Pr(***d***) obtained from reconstructions of the CMB maps with the classic MEM criterion are shown in Table 6. They confirm the results from Table 3 and the visual impression from Fig. 13. The evidences for the wavelet method are again significantly higher than for the real-space MEM, even though the relative evidence ratios are not quite as pronounced as for the reconstructions of the photographic image.

Table 4. Reconstruction errors between reconstructions and original CMB maps. The errors are averaged over a set of 25 simulations with five different noise and image realizations. In each case, the regularization parameter α is fixed using the 'historic' criterion.

PSF FWHM		3			5			10	
noise rms	0.1	0.5	2.0	0.1	0.5	2.0	0.1	0.5	2.0
real-space MEM	0.14	0.34	0.66	0.26	0.39	0.62	0.44	0.51	0.66
wavelet MEM	0.15	0.32	0.55	0.27	0.40	0.57	0.45	0.52	0.65
wavelet-regularized MEM	0.15	0.32	0.59	0.27	0.39	0.59	0.45	0.52	0.65

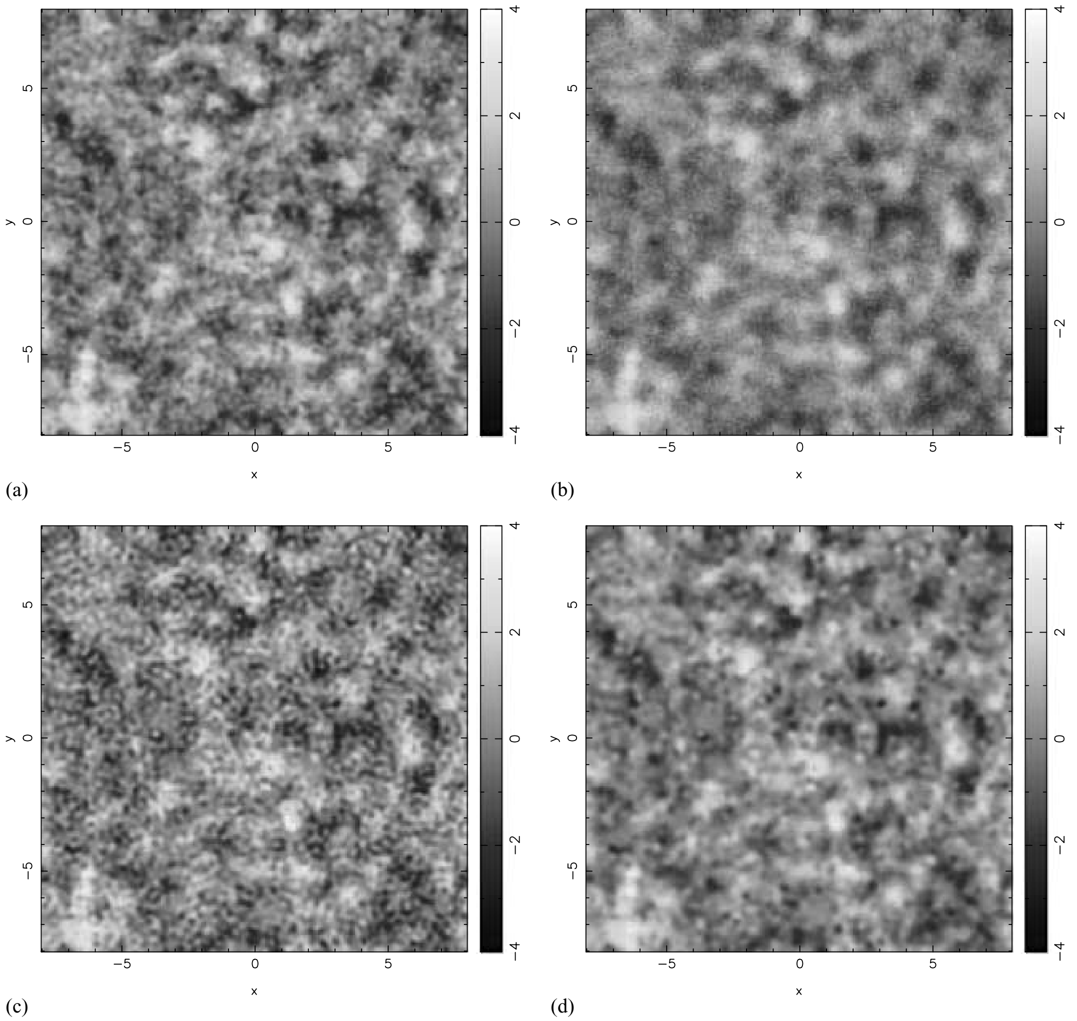


Figure 13. (a) One realization of the CMB maps used in the simulations described in the text. The map is normalized to unity rms, and the axes are labelled in degrees. (b) The 'data image' obtained after a convolution with a Gaussian point spread function with a FWHM of 5 pixels and the addition of Gaussian random noise whose rms is half the rms of the original image. (c) Reconstruction using real-space MEM. (d) Reconstruction using wavelet MEM with a four-level à trous transform. In both cases, the regularization parameter α is fixed using the 'Bayesian' criterion.

Table 6. The logarithm ln Pr($\boldsymbol{d}$) of the Bayesian evidence Pr($\boldsymbol{d}$) for differnt reconstructions of the CMB maps. The values are averaged over a set of 25 simulations with five different noise and image realizations and are normalized such that they equal zero for the real-space MEM for each dataset.

PSF FWHM	3			5			10		
noise rms	0.1	0.5	2.0	0.1	0.5	2.0	0.1	0.5	2.0
real-space MEM	0	0	0	0	0	0	0	0	0
wavelet MEM	6197	2526	631	2305	1262	423	542	296	75

7 CONCLUSIONS

Wavelets are functions that enable an efficient representation of signals or images. They help to identify and compress the information content of the signal into a small number of parameters. Because wavelet functions span a whole range of spatial scales, they can be used to describe signal correlations of different characteristic lengths. In this paper, we have investigated how wavelets can be combined with the maximum entropy method to improve the reconstruction of images from blurred and noisy data.

There are two principal ways to incorporate wavelets into the maximum entropy method. First, the wavelet transform can be treated as an intrinsic correlation function that is used to decorrelate the data (wavelet MEM). Secondly, the wavelet transform can be combined with the entropy functional into a new effective wavelet entropy (wavelet-regularized entropy). We have implemented both approaches for the à trous wavelet transform, which is non-orthogonal and has the benefit of being invariant under translations and rotations of the image. We show that the à trous transform can be considered as a special case of a multichannel ICF, in which the image is produced from a linear combination of images convolved with point spread functions of different widths. The quality of the reconstruction depends on the relative weights assigned to each channel or scale.

We have applied MEM implementations using à trous wavelets to simulated observations of CMB temperature anisotropies. We find that while the relative weighting of scales or channels is important, there is a range of different weightings that can yield roughly similar results as long as they suppress small-scale structure in the image. It does not matter much whether the weighting is introduced by setting different channel weights or by rescaling the entropy expressions or default models. Weightings that suppress small-scale structure more efficiently can perform better for low signal-to-noise, whereas they are worse for high signal-to-noise. Furthermore, we also find that for images containing structure on different scales, like CMB maps, methods that try to improve the reconstruction by ad hoc assignments of pixel- and data-dependent weights usually do not enhance the reconstruction quality. The more complicated reconstruction prescriptions given by Pantin & Starck (1996) have no benefits for CMB map-making.

As far as reconstruction errors are concerned, wavelet-based maximum entropy algorithms seem to match the standard MEM in pixel space (real-space MEM) for large point spread functions or low noise levels. There exist sufficient well-determined degrees of freedom in the data to make the image basis irrelevant. On the other hand, for poor signal-to-noise and narrow convolution masks, wavelet methods outperform real-space MEM. Thus the use of wavelet techniques can improve the reconstruction of images in many cases, while there is no disadvantage of using these methods in other situations. The improvement seems to be genuinely related to the basis set or ICF and not just an artefact of an improper choice of the regularization. A Bayesian treatment of the regularization constant and a comparison of the different reconstruction methods shows a much higher evidence for ICF methods than for the simple real-space MEM. In a Bayesian context, the wavelet basis can thus be interpreted as a better 'model' for the image. In this regard, it fulfills the promise of an improvement over the real-space method that the ICF was designed to address (see e.g. MacKay 1992).

The isotropic à trous transform or the multichannel ICF can be implemented within the MEMSYS5 maximum entropy kernel and thus be applied in a proper Bayesian maximization scheme. To summarize, we conclude that the use of wavelets in MEM image reconstructions is a successful technique that can improve the quality of image reconstructions.

ACKNOWLEDGMENTS

We thank Belen Barreiro, Steve Gull, Phil Marshall and Charles McLachlan for helpful discussions. Klaus Maisinger acknowledges support from an EU Marie Curie Fellowship.

REFERENCES

Bijaoui A., Starck J.-L., Murtagh F., 1994, Traitement du Signal, 11, 229
Bontekoe T. R., Koper E., Kester D. J. M., 1994, A&A, 284, 1037
Bridle S. L., Hobson M. P., Lasenby A. N., Saunders R., 1998, MNRAS, 299, 895
Cayón L. et al., 2000, MNRAS, 315, 757
Frieden B. R., 1983, J. Opt. Soc. Am., 73, 927
Gull S. F., 1989, in Skilling J., ed., Maximum Entropy and Bayesian Methods Developments in Maximum Entropy data analysis. Kluwer, Dordrecht, p. 53
Gull S. F., Skilling J., 1999, Quantified Maximum Entropy, MEMSYS5 Users' Manual. Maximum Entropy Data Consultants Ltd., Bury St Edmunds
Hobson M. P., Lasenby A. N., 1998, MNRAS, 298, 905
Hobson M. P., Jones A. W., Lasenby A. N., Bouchet F. R., 1998, MNRAS, 300, 1
Hobson M. P., Jones A. W., Lasenby A. N., 1999, MNRAS, 309, 125
Holschneider M., Kronland-Martinet R., Morlet J., Tchamitchian P., 1989, in Combes J. M., Grossman A., Tchamitchian P., eds, Wavelets: Time-Frequency Methods and Phase Space. Springer-Verlag, Berlin, p. 286
Jones A. W., Hobson M. P., Lasenby A. N., 1999, MNRAS, 305, 898
Lang M., Guo H., Odegard J. E, Burrus C. S., Wells R. O., 1996, IEEE Sig. Proc. Lett., 3, 10
Langer W. D., Wilson R. W., Anderson C. H., 1993, ApJ, 408, L45
MacKay D. J. C., 1992, Neural Comput., 4, 415
Marshall P. J., Hobson M. P., Gull S. F., Bridle S. L., 2002, MNRAS, 335, 1037
Pantin E., Starck J.-L., 1996, A&AS, 118, 575
Rue F., Bijaoui A., 1997, Exp. Astron., 7, 129
Sanz J. L., Argüeso F., Cayón L., Martínez-González E., Barreiro R. B., Toffolatti L., 1999a, MNRAS, 309, 672
Sanz J. L. et al., 1999b, A&AS, 140, 99
Skilling J., 1989, in Skilling J., ed., Maximum entropy and Bayesian methods. Kluwer, Dordrecht,, p. 45
Starck J.-L., Murtagh F., Bijaoui A., 1998, Image processing and data analysis: the multiscale approach. Cambridge University Press, Cambridge
Starck J.-L., Murtagh F., Querre P., Bonnarel F., 2001, A&A, 368, 730
Stolyarov V., Hobson M. P., Ashdown M. A. J., Lasenby A. N., 2002, MNRAS, 336, 97
Tenorio L., Jaffe A. H., Hanany S., Lineweaver C. H., 1999, MNRAS, 310, 823
Thompson A. R., Moran J. M., Swenson G. W., 1994, Interferometry and Synthesis in Radio Astronomy. Krieger Publishing Company, New York
Titterington D. M., 1985, A&A, 144, 381
Vielva P., Barreiro R. B., Hobson M. P., Martínez-González E., Lasenby A. N., Sanz J. L., Toffolatti L., 2001, MNRAS, 328, 1
Weir N., 1992, in ASP Conf. Ser. 25, Astronomical Data Analysis Software and Systems I, Vol. 1. Astron. Soc. Pac., San Francisco, p. 186
Zhuang Y., Baras J. S., 1994, report CSHCN TR 94-7/ISR-TR-94-3. Center for Satellite and Hybrid Communication Networks, MD

APPENDIX A: CALCULATION OF DERIVATIVES

In an implementation of a transformed maximum-entropy algorithm, the posterior functional has to be minimized numerically. Some minimization routines, like the MEMSYS5 package (Gull & Skilling 1999), require first derivatives of F with respect to the (visible or hidden) image pixels, whereas others, like the simple Newton–Raphson method, additionally require second derivatives. The derivatives of the χ^2- and entropy terms can be calculated separately.

For hidden-space MEM, one must minimize

$$F(\boldsymbol{h}) = \frac{1}{2}\chi^2(\mathsf{K}_\mathrm{f}\boldsymbol{h}) - \alpha S(\boldsymbol{h}) \tag{A1}$$

with respect to $\boldsymbol{h}$. The resulting optimal hidden image $\hat{\boldsymbol{h}}$ is then transformed into the optimal visible-space vector $\hat{\boldsymbol{v}} = \mathsf{K}_\mathrm{f}\hat{\boldsymbol{h}}$. For hidden-space-regularized MEM, one minimizes

$$F(\boldsymbol{v}) = \frac{1}{2}\chi^2(\boldsymbol{v}) - \alpha S(\mathsf{K}_\mathrm{b}\boldsymbol{v}) \tag{A2}$$

with respect to $\boldsymbol{v}$ directly. In these expressions, for a linear instrumental response matrix R, the misfit term has the functional form

$$\chi^2(\boldsymbol{y}) = (\mathsf{R}\boldsymbol{y} - \boldsymbol{d})^\mathrm{t}\mathsf{N}^{-1}(\mathsf{R}\boldsymbol{y} - \boldsymbol{d}). \tag{A3}$$

The regularization function has the form

$$S(\boldsymbol{y}) = \sum_{i=1} s(y_i, m_i), \tag{A4}$$

where the sum extends over all the pixels in the relevant space, and where $s(y, m)$ may take either of the forms (16) or (18), depending on nature of the space in which the regularization is performed (see Section 3.3).

From (A3), we find that the gradient vector and curvature matrix of the misfit function with respect to its argument $\boldsymbol{y}$ are given by

$$\boldsymbol{g}_{\chi^2}(\boldsymbol{y}) = 2\mathsf{R}^\mathrm{t}\mathsf{N}^{-1}(\mathsf{R}\boldsymbol{y} - \boldsymbol{d}), \tag{A5}$$

$$\mathsf{H}_{\chi^2} = 2\mathsf{R}^\mathrm{t}\mathsf{N}^{-1}\mathsf{R}. \tag{A6}$$

Similarly, we find that the elements of the gradient vector and curvature matrix of the regularizing function (A4) with respect to its argument $\boldsymbol{y}$ are given by

$$[\boldsymbol{g}_S(\boldsymbol{y})]_i = \frac{\partial s(y_i, m_i)}{\partial y_i}, \tag{A7}$$

$$[\mathbf{H}_S(\boldsymbol{y})]_{ij} = \begin{cases} \dfrac{\partial^2 s(y_i, m_i)}{\partial y_i^2} & \text{if} \quad i = j \\ 0 & \text{otherwise.} \end{cases} \tag{A8}$$

From (16), the derivatives of the entropy functional for positive-only distributions are

$$\frac{\partial s_+(y, m)}{\partial y} = -\ln \frac{y}{m}, \qquad \frac{\partial^2 s_+(y, m)}{\partial y^2} = -\frac{1}{y},$$

whereas for the positive-negative entropy functional (18), we obtain

$$\frac{\partial s_\pm(y, m)}{\partial y} = -\ln \frac{\psi + y}{2m}, \qquad \frac{\partial^2 s_\pm(y, m)}{\partial y^2} = -\frac{1}{\psi},$$

where $\psi = \sqrt{y^2 + 4m^2}$.

A1 Hidden-space MEM

For hidden-space MEM, we see from (A1) that the gradient vector and curvature matrix of F with respect to $\boldsymbol{h}$ are given by

$$\nabla_{\boldsymbol{h}} F = \frac{1}{2}\mathbf{K}_\mathrm{f}^\mathrm{t} \boldsymbol{g}_{\chi^2}(\mathbf{K}_\mathrm{f}\boldsymbol{h}) - \alpha \boldsymbol{g}_S(\boldsymbol{h}), \tag{A9}$$

$$\nabla_{\boldsymbol{h}} \nabla_{\boldsymbol{h}} F = \frac{1}{2}\mathbf{K}_\mathrm{f}^\mathrm{t} \mathbf{H}_{\chi^2} \mathbf{K}_\mathrm{f} - \alpha \mathbf{H}_S(\boldsymbol{h}). \tag{A10}$$

We note that we may return to a standard maximum-entropy algorithm operating in real space by simply setting K_f equal to the identity.

A2 Hidden-space-regularized MEM

For hidden-space-regularized MEM, we see from (A2) that the gradient vector and curvature matrix of F with respect to $\boldsymbol{v}$ are given by

$$\nabla_{\boldsymbol{v}} F = \frac{1}{2}\boldsymbol{g}_{\chi^2}(\boldsymbol{v}) - \alpha \mathbf{K}_\mathrm{b}^\mathrm{t} \boldsymbol{g}_S(\mathbf{K}_\mathrm{b}\boldsymbol{v}), \tag{A11}$$

$$\nabla_{\boldsymbol{v}} \nabla_{\boldsymbol{v}} F = \frac{1}{2}\mathbf{H}_{\chi^2} - \alpha \mathbf{K}_\mathrm{b}^\mathrm{t} \mathbf{H}_S(\mathbf{K}_\mathrm{b}\boldsymbol{v}) \mathbf{K}_\mathrm{b}. \tag{A12}$$

We note that we may return to a standard maximum-entropy algorithm operating in real space by simply setting K_b equal to the identity.

APPENDIX B: EQUIVALENCE OF METHODS

In this Appendix, we show that the 'hidden-space MEM' and 'hidden-space regularized MEM' introduced in Section 3 become equivalent if the regularization function is quadratic and the 'forward' and 'backward' transforms are inverses, i.e. $\mathbf{K}_\mathrm{f}\mathbf{K}_\mathrm{b} = \mathbf{I} = \mathbf{K}_\mathrm{f}\mathbf{K}_\mathrm{f}$. Note that this requires the hidden and visible spaces to have the same dimensionality.

For hidden-space MEM, the χ^2-statistic is given by

$$\chi^2(\boldsymbol{h}) = (\mathbf{R}\mathbf{K}_\mathrm{f}\boldsymbol{h} - \boldsymbol{d})^\mathrm{t} \mathbf{N}^{-1} (\mathbf{R}\mathbf{K}_\mathrm{f}\boldsymbol{h} - \boldsymbol{d}). \tag{B1}$$

The most general form of quadratic regularization is given by

$$S(\boldsymbol{h}) = -\boldsymbol{h}^\mathrm{t} \mathbf{C}^{-1} \boldsymbol{h}.$$

This is equivalent to assuming a multivariate Gaussian prior on the hidden-space variables $\boldsymbol{h}$, with mean zero and covariance matrix $\mathbf{C}$. In the special case of the quadratic approximation (20) to the positive/negative entropy with a model $\boldsymbol{m}$, the matrix $\mathbf{C}$ is defined as $C_{ij} = 4\, m_i \delta_{ij}$. The maximum entropy solution $\hat{\boldsymbol{h}}$ can be found by minimizing the function (3),

$$\begin{aligned} F(\boldsymbol{h}) &= \frac{1}{2}\chi^2(\boldsymbol{h}) - \alpha S(\boldsymbol{h}) \\ &= \frac{1}{2}(\mathbf{R}\mathbf{K}_\mathrm{f}\boldsymbol{h} - \boldsymbol{d})^\mathrm{t} \mathbf{N}^{-1} (\mathbf{R}\mathbf{K}_\mathrm{f}\boldsymbol{h} - \boldsymbol{d}) + \alpha \boldsymbol{h}^\mathrm{t} \mathbf{C}^{-1} \boldsymbol{h} \\ &= \frac{1}{2}\boldsymbol{d}^\mathrm{t} \mathbf{N}^{-1} \boldsymbol{d} - \boldsymbol{h}^\mathrm{t} \mathbf{K}_\mathrm{f}^\mathrm{t} \mathbf{R}^\mathrm{t} \mathbf{N}^{-1} \boldsymbol{d} \\ &\quad + \boldsymbol{h}^\mathrm{t} \left(\frac{1}{2}\mathbf{K}_\mathrm{f}^\mathrm{t} \mathbf{R}^\mathrm{t} \mathbf{N}^{-1} \mathbf{R} \mathbf{K}_\mathrm{f} + \alpha \mathbf{C}^{-1} \right) \boldsymbol{h}. \end{aligned}$$

Demanding that the gradient of F with respect to $\boldsymbol{h}$ vanishes at the maximum $\hat{\boldsymbol{h}}$, we obtain

$$\left(\frac{1}{2}\mathbf{K}_\mathrm{f}^\mathrm{t} \mathbf{R}^\mathrm{t} \mathbf{N}^{-1} \mathbf{R} \mathbf{K}_\mathrm{f} + \alpha \mathbf{C}^{-1} \right) \hat{\boldsymbol{h}} = \mathbf{K}_\mathrm{f}^\mathrm{t} \mathbf{R}^\mathrm{t} \mathbf{N}^{-1} \boldsymbol{d}. \tag{B2}$$

For hidden-space-regularized MEM, we have

$$\chi^2(\boldsymbol{v}) = (\mathbf{R}\boldsymbol{v} - \boldsymbol{d})^\mathrm{t} \mathbf{N}^{-1} (\mathbf{R}\boldsymbol{v} - \boldsymbol{d}) \quad \text{and}$$

$$S(\mathbf{K}_\mathrm{b}\boldsymbol{v}) = -\boldsymbol{v} \mathbf{K}_\mathrm{b}^\mathrm{t} \mathbf{C}^{-1} \mathbf{K}_\mathrm{b} \boldsymbol{v}.$$

The maximum-entropy solution $\hat{\boldsymbol{v}}$is given by

$$\left(\frac{1}{2}\mathbf{R}^\mathrm{t} \mathbf{N}^{-1} \mathbf{R} + \alpha \mathbf{K}_\mathrm{b}^\mathrm{t} \mathbf{C}^{-1} \mathbf{K}_\mathrm{b} \right) \hat{\boldsymbol{v}} = \mathbf{R}^\mathrm{t} \mathbf{N}^{-1} \boldsymbol{d}. \tag{B3}$$

In the special case where $\mathbf{K}_\mathrm{f}\mathbf{K}_\mathrm{b} = \mathbf{I} = \mathbf{K}_\mathrm{f}\mathbf{K}_\mathrm{f}$, on multiplying (B3) by $\mathbf{K}_\mathrm{f}^\mathrm{t}$, we obtain

$$\left(\frac{1}{2}\mathbf{K}_\mathrm{f}^\mathrm{t} \mathbf{R}^\mathrm{t} \mathbf{N}^{-1} \mathbf{R} \mathbf{K}_\mathrm{f} + \alpha \mathbf{C}^{-1} \right) \mathbf{K}_\mathrm{f}^{-1} \hat{\boldsymbol{v}} = \mathbf{K}_\mathrm{f}^\mathrm{t} \mathbf{R}^\mathrm{t} \mathbf{N}^{-1} \boldsymbol{d}.$$

By comparison with (B2), we see that $\hat{\boldsymbol{h}} = \mathbf{K}_\mathrm{f}^{-1} \hat{\boldsymbol{v}}$. The solutions for hidden-space MEM and hidden-space-regularized MEM are thus identical.

This paper has been typeset from a TEX/LATEX file prepared by the author.