Annu. Rev. Astron. Astrophys. 2005. 43:139–94
doi: 10.1146/annurev.astro.43.112904.104850

First published online as a Review in Advance on June 16, 2005

DIGITAL IMAGE RECONSTRUCTION: Deblurring and Denoising

R.C. Puetter,[1,4] T.R. Gosnell,[2,4] and Amos Yahil[3,4]
[1]*Center for Astrophysics and Space Sciences, University of California, San Diego, La Jolla, CA 92093*
[2]*Los Alamos National Laboratory, Los Alamos, NM 87545*
[3]*Department of Physics and Astronomy, Stony Brook University, Stony Brook, NY 11794*
[4]*Pixon LLC, Stony Brook, NY 11790; email: Rick.Puetter@pixon.com, Tim.Gosnell@pixon.com, Amos.Yahil@pixon.com*

Key Words image processing, image restoration, maximum entropy, Pixon, regularization, wavelets

■ **Abstract** Digital image reconstruction is a robust means by which the underlying images hidden in blurry and noisy data can be revealed. The main challenge is sensitivity to measurement noise in the input data, which can be magnified strongly, resulting in large artifacts in the reconstructed image. The cure is to restrict the permitted images. This review summarizes image reconstruction methods in current use. Progressively more sophisticated image restrictions have been developed, including (*a*) filtering the input data, (*b*) regularization by global penalty functions, and (*c*) spatially adaptive methods that impose a variable degree of restriction across the image. The most reliable reconstruction is the most conservative one, which seeks the simplest underlying image consistent with the input data. Simplicity is context-dependent, but for most imaging applications, the simplest reconstructed image is the smoothest one. Imposing the maximum, spatially adaptive smoothing permitted by the data results in the best image reconstruction.

1. INTRODUCTION

Digital image processing of the type discussed in this review has been developed extensively and now routinely provides high-quality, robust reconstructions of blurry and noisy data collected by a wide variety of sensors. The field exists because it is impossible to build imaging instruments that produce arbitrarily sharp pictures uncorrupted by measurement noise. It is, however, possible mathematically to reconstruct the underlying image from the nonideal data obtained from real-world instruments, so that information present but hidden in the data is revealed with less blur and noise. The improvement from raw input data to reconstructed image can be quite dramatic.

Our choice of nomenclature is deliberate. Throughout this review, "data" refers to any measured quantity, from which an unknown "image" is estimated through the process of image reconstruction.[1] The term image denotes either the estimated solution or the true underlying image that gives rise to the observed data. The discussion usually makes clear which context applies; in cases of possible ambiguity we use "image model" to denote the estimated solution. Note that the data and the image need not be similar and may even have different dimensionality, e.g., tomographic reconstructions seek to determine a 3D image from projected 2D data.

Image reconstruction is difficult because substantial fluctuations in the image may be strongly blurred, yielding only minor variations in the measured data. This causes two major, related problems for image reconstruction. First, noise fluctuations may be mistaken for real signal. Overinterpretation of data is always problematic, but image reconstruction magnifies the effect to yield large image artifacts. The high wave numbers (spatial frequencies) of the image model are particularly susceptible to these artifacts, because they are suppressed more strongly by the blur and are therefore less noticeable in the data. In addition, it may be impossible to discriminate between competing image models if the differences in the data models obtained from them by blurring are well within the measurement noise. For example, two closely spaced point sources might be statistically indistinguishable from a single, unresolved point source. A definitive resolution of the image ambiguity can then only come with additional input data.

Image reconstruction tackles both these difficulties by making additional assumptions about the image. These assumptions may appeal to other knowledge about the imaged object, or there may be accepted procedures to favor a "reasonable" or a "conservative" image over a "less reasonable" or an "implausible" one. The key to stable image reconstruction is to restrict the permissible image models, either by disallowing unwanted solutions altogether, or by making it much less likely that they are selected by the reconstruction. Almost all modern image reconstructions restrict image models in one way or another. They differ only in what they restrict and how they enforce the restriction. The trick is not to throw out the baby with the bath water. The more restrictive the image reconstruction, the greater its stability, but also the more likely it is to eliminate correct solutions. The goal is therefore to describe the allowed solutions in a sufficiently general way that accounts for all possible images that may be encountered, and at the same time to be as strict as possible in selecting the preferred images.

[1]Historically, the problem of deblurring and denoising of imaging data was termed image restoration, a subtopic within a larger computational problem known as image reconstruction. Most contemporary workers now use the latter, more general term, and we adopt this terminology in this review. Also, some authors use "image" for what we call "data" and "object" for what we call " image." Readers familiar with that terminology need to make a mental translation when reading this review.

There are great arguments on how image restriction should be accomplished. After decades of development, the literature on the subject is still unusually editorial and even contentious in tone. At times it sounds as though image reconstruction is an art, a matter of taste and subjective preference, instead of an objective science. We take a different view. For us, the goal of image reconstruction is to come as close as possible to the true underlying image, pure and simple. And there are objective criteria by which the success of image reconstruction can be measured. First, it must be internally self-consistent. An image model predicts a data model, and the residuals—the differences between the data and the data model—should be statistically consistent with our understanding of the measurement noise. If we see structure in the residuals, or if their statistical distribution is inconsistent with noise statistics, there is something wrong with the image model. This leaves image ambiguity that cannot be statistically resolved by the available data. The image reconstruction can then only be validated externally by additional measurements, preferably by independent investigators. Simulations are also useful, because the true image used to create the simulated data is known and can be compared with the reconstructed image.

There are several excellent reviews of image reconstruction and numerical methods by other authors. These include: Calvetti, Reichel & Zhang (1999) on iterative methods; Hansen (1994) on regularization methods; Molina et al. (2001) and Starck, Pantin & Murtagh (2002) on image reconstruction in astronomy; Narayan & Nityananda (1986) on the maximum-entropy method; O'Sullivan, Blahut & Snyder (1998) on an information-theoretic view; Press et al. (2002) on the inverse problem and statistical and numerical methods in general; and van Kempen et al. (1997) on confocal microscopy. There are also a number of important regular conferences on image processing, notably those sponsored by the International Society for Optical Engineering (http://www.spie.org), the Optical Society of America (http://www.osa.org), and the Computer Society of the Institute of Electrical and Electronics Engineers (http://www.computer.org).

Our review begins with a discussion of the mathematical preliminaries in Section 2. Our account of image reconstruction methods then proceeds from the simple to the more elaborate. This path roughly follows the historical development, because the simple methods were used first, and more sophisticated ones were developed only when the simpler methods proved inadequate.

Simplest are the noniterative methods discussed in Section 3. They provide explicit, closed-form inverse operations by which data are converted to image models in one step. These methods include Fourier and small-kernel deconvolutions, possibly coupled with Wiener filtering, wavelet denoising, or quick Pixon smoothing. We show in two examples that they suffer from noise amplification to one degree or another, although good filtering can greatly reduce its severity.

The limitations of noniterative methods motivated the development of iterative methods that fit an image model to the data by using statistical tests to determine how well the model fits the data. Section 4 launches the statistical discussion by

introducing the concepts of merit function, maximum likelihood, goodness of fit, and error estimates.

Fitting methods fall into two broad categories, parametric and nonparametric. Section 5 is devoted to parametric methods, which are suitable for problems in which the image can be modeled by explicit, known source functions with a few adjustable parameters. "Clean" is an example of a parametric method used in radio astronomy. We also include a brief discussion of parametric error estimates.

Section 6 introduces simple nonparametric iterative schemes including the van Cittert, Landweber, Richardson-Lucy, and conjugate-gradient methods. These nonparametric methods replace the small number of source functions with a large number of unknown image values defined on a grid, thereby allowing a much larger pool of image models. But image freedom also results in image instability, which requires the introduction of image restriction. The two simplest forms of image restriction discussed in Section 6 are the early termination of the maximum-likelihood fit, before it reaches convergence, and the imposition of the requirement that the image be nonnegative. The combination of the two restrictions is surprisingly powerful in attenuating noise amplification and even increasing resolution, but some reconstruction artifacts remain.

To go beyond the general methods of Section 6 requires additional restrictions whose task is to smooth the image model and suppress the artifacts. Section 7 discusses the methods of linear (Tikhonov) regularization, total variation, and maximum entropy, which impose global image preference functions. These methods were originally motivated by two differing philosophies but ended up equivalent to each other. Both state image preference using a global function of the image and then optimize the preference function subject to data constraints.

Global image restriction can significantly improve image reconstruction, but, because the preference function is global, the result is often to underfit the data in some parts of the image and to overfit in other parts. Section 8 presents spatially adaptive methods of image restriction, including spatially variable entropy, wavelets, Markov random fields and Gibbs priors, atomic priors and massive inference, and the full Pixon method. Section 8 ends with several examples of both simulated and real data, which serve to illustrate the theoretical points made throughout the review.

We end with a summary in Section 9. Let us state our conclusion outright. The future, in our view, lies with the added flexibility enabled by spatially adaptive image restriction coupled with strong rules on how the image restriction is to be applied. On the one hand, the larger pool of permitted image models prevents underfitting the data. On the other hand, the stricter image selection avoids overfitting the data. Done correctly, we see the spatially adaptive methods providing the ultimate image reconstructions.

The topics presented in this review are by no means exhaustive of the field, but limited space prevents us from discussing several other interesting areas of image reconstruction. A major omission is superresolution, a term used both for subdiffraction resolution (Hunt 1995, Bertero & Boccacci 2003) and subpixel

resolution. Astronomers might be familiar with the "drizzle" technique developed at the Space Telescope Science Institute (Fruchter & Hook 2002) to obtain subpixel resolution using multiple, dithered data frames. There are also other approaches (Borman & Stevenson 1998; Elad & Feuer 1999; Park, Park & Kang 2003).

Other areas of image reconstruction left out include: (*a*) tomography (Natterer 1999); (*b*) vector quantization, a technique widely used in image compression and classification (Gersho & Gray 1992, Cosman et al. 1993, Hunt 1995, Sheppard et al. 2000); (*c*) the method of projection onto convex sets (Biemond, Lagendijk & Mersereau 1990); (*d*) the related methods of singular-value decomposition, principal components analysis, and independent components analysis (Raykov & Marcoulides 2000; Hyvärinen, Karhunen & Oja 2001; Press et al. 2002); and (*e*) artificial neural networks, used mainly in image classification but also useful in image reconstruction (Dávila & Hunt 2000; Egmont-Petersen, de Ridder & Handels 2002).

We also confine ourselves to physical blur—instrumental and/or atmospheric—that spreads the image over more than one pixel. Loss of resolution due to the finite pixel size can in principle be overcome by better optics (stronger magnification) or a finer focal plane array. In practice, the user is usually limited by the optics and focal plane array at hand. The main recourse then is to take multiple, dithered frames and use some of the superresolution techniques referenced above. Ironically, if the physical blur extends over more than one pixel, one can also recover some subpixel resolution by requiring that the image be everywhere nonnegative (Section 2.4). This technique does not work if the physical blur is less than one pixel wide.

Another area that we omit is the initial data reduction that needs to be performed before image reconstruction can begin. This includes nonuniformity corrections, elimination of bad pixels, background subtraction where appropriate, and the determination of the type and level of statistical noise in the data. Major astronomical observatories often have online manuals providing instructions for these operations, e.g., the direct imaging manual of Kitt Peak National Observatory (http://www.noao.edu/kpno/manuals/dim).

2. MATHEMATICAL PRELIMINARIES

2.1. Blur

The image formed in the focal plane is blurred by the imaging instrument and the atmosphere. It can be expressed as an integral over the true image, denoted symbolically by $\otimes$:

$$M(\mathbf{x}) = P \otimes I = \int P(\mathbf{x}, \mathbf{y}) I(\mathbf{y})\, d\mathbf{y}. \tag{1}$$

For a 2D image, the integration is over the 2D $\mathbf{y}$-space upon which the image is defined. In general, the imaging problem may be defined in a space with an

arbitrary number of dimensions, and the dimensionalities of **x** and **y** need not even be the same (e.g., in tomography). Wavelength and/or time might provide additional dimensions. There may also be multiple focal planes, or multiple exposures in the same focal plane, perhaps with some dithering.

The kernel of the integral, $P(\mathbf{x}, \mathbf{y})$, is called the point-spread function. It is the probability that a photon originating at position **y** in the image plane ends up at position **x** in the focal plane. Another way of looking at the point-spread function is as the image formed in the focal plane by a point source of unit flux at position **y**, hence, the name point-spread function.

2.2. Convolution

In general, the point-spread function varies independently with respect to both **x** and **y**, because the blur of a point source may depend on its location in the image. Optical systems that suffer strong geometric aberrations behave in this way. Often, however, the point-spread function can be accurately written as a function only of the displacement $\mathbf{x} - \mathbf{y}$, independent of location within the field of view. In that case, Equation 1 becomes a convolution integral:

$$M(\mathbf{x}) = P * I = \int P(\mathbf{x} - \mathbf{y}) I(\mathbf{y})\, d\mathbf{y}. \tag{2}$$

When necessary, we use the symbol $*$ to specify a convolution operation to make clear the distinction with the more general integral operation $\otimes$. Most of the image reconstruction methods described in this review, however, are general and are not restricted to convolving point-spread functions.

Convolutions have the added benefit that they translate into simple algebraic products in the Fourier space of wave vectors **k** (e.g., Press et al. 2002):

$$\tilde{M}(\mathbf{k}) = \tilde{P}(\mathbf{k})\tilde{I}(\mathbf{k}). \tag{3}$$

For a convolving point-spread function there is thus a direct **k**-by-**k** correspondence between the true image and the blurred image. The reason for this simplicity is that the Fourier spectral functions $\exp(\imath \mathbf{k} \cdot \mathbf{x})$ are eigenfunctions of the convolution operator, i.e., convolving them with the point-spread function returns the input function multiplied by the eigenvalue $\tilde{P}(\mathbf{k})$.

The separation of the blur into decoupled operations on each of the Fourier components of the image points to a simple method of image reconstruction. Equation 3 may be solved for the image in Fourier space:

$$\tilde{I}(\mathbf{k}) = \tilde{M}(\mathbf{k}) / \tilde{P}(\mathbf{k}), \tag{4}$$

and the image is the inverse Fourier transform of $\tilde{I}(\mathbf{k})$. In practice, this method is limited by noise (Section 3.1). We bring it up here as a way to think about image reconstruction and, in particular, about sampling and biasing (Sections 2.3 and 2.4). Even in the more general case in which image blur is not a precise convolution, it is still useful to conceptualize image reconstruction in terms of Fourier components,

because the coupling between different Fourier components is often limited to a small range of **k**.

2.3. Data Sampling

Imaging detectors do not actually measure the continuous blurred image. In modern digital detectors, the photons are collected and counted in a finite number of pixels with nonzero widths placed at discrete positions in the focal plane. They typically form an array of adjacent pixels. The finite pixel width results in further blur, turning the point-spread function into a point-response function. Assuming that all the pixels have the same response, the point-response function is a convolution of the pixel-responsivity function S and the point-spread function:

$$H(\mathbf{x}, \mathbf{y}) = S * P. \tag{5}$$

The point-response function is actually only evaluated at the discrete positions of the pixels, x_i, typically taken to be at the centers of the pixels. The data expected to be collected in pixel i—in the absence of noise (Section 2.7)—is then

$$M_i = \int H(\mathbf{x}_i, \mathbf{y}) I(\mathbf{y})\, d\mathbf{y} = (H \otimes I)_i. \tag{6}$$

We also refer to M_i as the data model when the image under discussion is an image model.

Image reconstruction actually only requires knowledge of the point-response function, which relates the image to the expected data. There is never any need to determine the point-spread function, because the continuous blurred image is not measured directly. But it is necessary to determine the point-response function with sufficient accuracy (Section 2.9). Approximating it by the point-spread function is often inadequate.

2.4. How to Overcome Aliasing Due to Discrete Sampling

Another benefit of the Fourier representation is its characterization of sampling and biasing. The sampling theorem tells us precisely what can and cannot be determined from a discretely sampled function (e.g., Press et al. 2002). Specifically, the sampled discrete values completely determine the continuous function, provided that the function is bandwidth-limited within the Nyquist cutoffs:

$$-\frac{1}{2\boldsymbol{\Delta}} = -\mathbf{k}_c \le \mathbf{k} \le \mathbf{k}_c = \frac{1}{2\Delta}. \tag{7}$$

(The Nyquist cutoffs are expressed in vector form, because the grid spacing $\boldsymbol{\Delta}$ need not be the same in different directions.) By the same token, if the continuous function is not bandwidth-limited and has significant Fourier components beyond the Nyquist cutoffs, then these components are aliased within the Nyquist cutoffs and cannot be distinguished from the Fourier components whose wave vectors really lie within the Nyquist cutoffs. This leaves an inherent ambiguity regarding

the nature of any continuous function that is sampled discretely, thereby limiting resolution.

The point-spread function is bandwidth-limited and therefore so is the blurred image. The bandwidth limit may be a strict cutoff, as in the case of diffraction, or a gradual one, as for atmospheric seeing (in which case the effective bandwidth limit depends on the signal-to-noise ratio). In any event, the blurred image is as bandwidth-limited as the point-spread function. It is therefore completely specified by any discrete sampling whose Nyquist cutoffs encompass the bandwidth limit of the point-spread function. The true image, however, is what it is and need not be bandwidth-limited. It may well contain components of interest beyond the bandwidth limit of the point-spread function.

The previous discussion about convolution (Section 2.2) suggests that the reconstructed image would be as bandwidth-limited as the data and little could be done to recover the high-**k** components of the image. This rash conclusion is incorrect. We usually have additional information about the image, which we can utilize. Almost all images must be nonnegative. (There are some important exceptions, e.g., complex images, for which positivity has no meaning.) Sometimes we also know something about the shapes of the sources, e.g., they may all be stellar point sources. Taking advantage of the additional information, we can determine the high-**k** image structure beyond the bandwidth limit of the data (Biraud 1969). For example, we can tell that a nonnegative image is concentrated toward one corner of the pixel because the point-spread function preferentially spreads flux to pixels around that corner. (Autoguiders take advantage of this feature to prevent image drift.)

We can see how this works for a nonnegative image by considering the Fourier reconstruction of an image blurred by a convolving point-spread function (Section 2.2). The reconstructed Fourier image $\tilde{I}(\mathbf{k})$ is determined by Equation 4 within the bandwidth limit of the data. But the image obtained from it by the inverse Fourier transformation may be partly negative. To remove the negative image values, we must extrapolate $\tilde{I}(\mathbf{k})$ beyond the bandwidth limit. We are free to do so, because the data do not constrain the high-**k** image components. Of course, we cannot extrapolate too far, or else aliasing will again cause ambiguity. But we can establish some **k** limit on the image by assuming that the image has no Fourier components beyond that limit. The point is that the image bandwidth limit needs to be higher than the data bandwidth limit, and the reconstructed image therefore has higher resolution than the data. The increased resolution is a direct result of the requirement of nonnegativity, without which we cannot extrapolate beyond the bandwidth limit of the data. In the above example of subpixel structure, we can deduce the concentration of the image toward the corner of the pixel because the image is restricted to be nonnegative. If it could also be negative, the same data could result from any number of images, as the sampling theorem tells us. Subpixel resolution is enabled by nonnegativity. On the other hand, if we also know that the image consists of point sources, we can constrain the high-**k** components of the image even further and obtain yet higher resolution.

Note that the pixel-responsivity function is not bandwidth-limited in and of itself because it has sharp edges. If the point-spread function is narrower than a

pixel, the data are not Nyquist sampled. The reconstructed image is then subject to additional aliasing, and it may not be possible to increase resolution beyond that set by the pixelation of the data. If the point-spread function is much narrower than the pixel, a point source could be anywhere inside the pixel. We cannot determine its position with greater accuracy, because the blur does not spill enough of its flux into neighboring pixels.

2.5. Background Subtraction

The use of nonnegativity as a major constraint in image reconstruction points to the importance of background subtraction. Requiring a background-subtracted image to be nonnegative is much more restrictive, because when an image sits on top of a significant background, negative fluctuations in the image can be absorbed by the background. The user is therefore well advised to subtract the background before commencing image reconstruction, if at all possible. Astronomical images usually lend themselves to background subtraction because a significant area of the image is filled exclusively by the background sky. It may be more difficult to subtract the background in a terrestrial image.

The best way to subtract background is by chopping and nodding, alternating between the target and a nearby blank field and recording only difference measurements. But the imaging instrument must be designed to do so. (See the descriptions of such devices at major astronomical observatories, e.g., on the thermal-region camera spectrograph of the Gemini South telescope, http://www.gemini.edu/sciops/instruments/miri/T-ReCSChopNod.html.) Absent such capability, the background can only be subtracted after the data are taken. Several methods have been proposed to subtract background (Bijaoui 1980; Infante 1987; Beard, McGillivray & Thanisch 1990; Almoznino, Loinger & Brosch 1993; Bertin & Arnouts 1996). The background may be a constant or a slowly varying function of position. It should in any event not vary significantly on scales over which the image is known to vary, or else the background subtraction may modify image structures. There are also several ways to clip sources when estimating the background (see the discussion by Bertin & Arnouts 1996).

2.6. Image Discretization

In general, an image can either be specified parametrically by known source functions (Section 5) or it can be represented nonparametrically on a discrete grid (Section 6), in which case the integral in Equation 1 is converted to a sum, yielding a set of linear equations:

$$M_i = \sum_j H_{ij} I_j, \tag{8}$$

in matrix notation $M = HI$. Here M is a vector containing n expected data values, I is a vector of m image values representing a discrete version of the image, and H is an $n \times m$ matrix representation of the point-response function. Note that each of what we here term vectors is in reality often a multidimensional array. In a typical

2D case, $n = n_x \times n_y$, $m = m_x \times m_y$, and the point-response function is an $(n_x \times n_y) \times (m_x \times m_y)$ array.

Note also for future reference that one often needs the transpose of the point-response function H^T, which is the $m \times n$ matrix obtained from H by transposing its rows and columns. H^T is the point-response function for an optical system in which the roles of the image plane and the focal plane are reversed (known in tomography as back projection). It is not to be confused with the inverse of the point-response function H^{-1}, an operator that exists only for square matrices, $n = m$. When applied to the expected data, H^{-1} provides the image that gave rise to the expected data through the original optical system at hand.

The discussion of sampling and aliasing (Section 2.4) shows that the image must, under some circumstances, be determined with better resolution than the data to ensure that it is nonnegative. This requires the image grid to be finer than the data grid, so the image is Nyquist sampled within the requisite image bandwidth. In that case, the number of image values is not equal to the number of data points, and the point-response function is not a square matrix. Conversely, if the image is known to be more bandwidth-limited than the data, or if the signal-to-noise ratio is low, so the high-**k** components are uncertain, we may choose a coarser discretization of the image than the data.

In the case of a convolution, the discretization is greatly simplified by using Equation 2 to give:

$$M_i = \sum_j H_{i-j} I_j. \tag{9}$$

Note that Equation 9 assumes that the data and image grids have the same spacing, and the point-response function takes a different form than in Equation 8. Like the expected data and the image, H is also an n-point vector, not an $n \times n$ matrix. (Recall that n refers to the total number of grid points, e.g., for a 2D array $n = n_x \times n_y$.) Discrete convolutions of the form of Equation 9 have a discrete Fourier analog of Equation 3 and can be computed efficiently by fast Fourier transform techniques (e.g., Press et al. 2002). When the image grid needs to be more finely spaced than the data, the convolution is performed on the image grid and then sampled at the positions of the pixels on the coarser grid.

2.7. Noise

A major additional factor limiting image reconstruction is noise due to measurement errors. The measured data actually consist of the expected data M_i plus measurement errors:

$$D_i = M_i + N_i = (H \otimes I)_i + N_i = \int H(\mathbf{x}_i, \mathbf{y}) I(\mathbf{y})\, d\mathbf{y} + N_i. \tag{10}$$

The discrete form of Equation 10 is obtained in analogy with Equation 8.

Measurement errors fall into two categories. Systematic errors are recurring errors caused by erroneous measurement processes or failure to take into account

physical effects that modify the measurements. In addition, there are random, irreproducible errors that vary from one measurement to the next. Because we do not know and cannot predict what a random error will be in any given measurement, we can at best deal with random errors statistically, assuming that they are random realizations of some parent statistical distribution. In imaging, the most commonly encountered parent statistical distributions are the Gaussian, or normal, distribution and the Poisson distribution (Section 4.2).

To be explicit, consider a trial solution to Equation 10, $\hat{I}(\mathbf{y})$, and compute the residuals

$$R_i = D_i - M_i = D_i - \int H(\mathbf{x}_i, \mathbf{y})\hat{I}(\mathbf{y})\, d\mathbf{y}. \tag{11}$$

The image model is an acceptable solution of the inverse problem if the residuals are consistent with the parent statistical distribution of the noise. The data model is then our estimate of the reproducible signal in the measurements, and the residuals are our estimate of the irreproducible noise. There is something wrong with the image model if the residuals show systematic structure or if their statistical distribution differs significantly from the parent statistical distribution. Examples would be if its mean is not zero, or if the distribution is skewed, too broad, or too narrow. After the fit is completed, it is therefore imperative to apply diagnostic tests to rule out problems with the fit. Some of the most useful diagnostic tools are goodness of fit, analysis of the statistical distribution of the residuals and their spatial correlations, and parameter error estimation (Sections 4 and 5).

2.8. Instability of Image Reconstruction

Image reconstruction is unfortunately an ill-posed problem. Mathematicians consider a problem to be well posed if its solution (*a*) exists, (*b*) is unique, and (*c*) is continuous under infinitesimal changes of the input. The problem is ill posed if it violates any of the three conditions. The concept goes back to Hadamard (1902, 1923). Scientists and engineers are usually less concerned with existence and uniqueness and worry more about the stability of the solution.

In image reconstruction, the main challenge is to prevent measurement errors in the input data from being amplified to unacceptable artifacts in the reconstructed image. Stated as a discrete set of linear equations, the ill-posed nature of image reconstruction can be quantified by the condition number of the point-response-function matrix. The condition number of a square matrix is defined as the ratio between its largest and smallest (in magnitude) eigenvalues (e.g., Press et al. 2002).[2] A singular matrix has an infinite condition number and no unique solution. An ill-posed problem has a large condition number, and the solution is sensitive to small changes in the input data.

[2]If the number of data and image points is not equal, then the point-response function is not square, and its condition number is not strictly defined. We can then use the square root of the condition number of $H^T H$, where H^T is the transpose of H.

How large is the condition number? A realistic point-response function can blur the image over a few pixels. In this case, the highest **k** components of the image are strongly suppressed by the point-response function. In other words, the high-**k** components correspond to eigenfunctions of the point-response function with very small eigenvalues. Hence, there is no escape from a large condition number. Equations 8 or 9 can therefore not be solved in their present, unrestricted forms. Either the equations must be modified, or the solutions must be projected away from the subspace spanned by the eigenfunctions with small eigenvalues. The bulk of this review is devoted to methods of image restriction (Sections 3–8).

2.9. Accuracy of the Point-Response Function

Image reconstruction is further compromised if the point-response function is not determined accurately enough. The signal-to-noise ratio determines how accurately it needs to be determined. The goal is that the wings of bright sources will not be confused with weak sources nearby. The higher the signal-to-noise ratio, the greater the care needed in determining the point-response function. The residual errors caused by the imprecision of the point-response function should be well below the noise.

The first requirement is that the profile of the point-response function correspond to the real physical blur. If the point-response function assumed in the reconstruction is narrower than the true point-response function, the reconstruction cannot remove all the blur. The image model then has less than optimal resolution, but artifacts should not be generated. On the other hand, if the assumed point-response function is broader than the true one, then the image model looks sharper than it really is. In fact, if the scene has narrow sources or sharp edges, it may not be possible to reconstruct the image correctly. Artifacts in the form of "ringing" around sharp objects are then seen in the image model.

Second, the point-response function must be defined on the image grid, which may need to be finer than the data grid to ensure a nonnegative image (Section 2.4). A point-response function measured from a single exposure of one point source is then inadequate because it is only appropriate for sources that are similarly placed within their pixels as the source used to measure the point-response function. Either multiple frames need to be taken displaced by noninteger pixel widths, or the point-response function has to be determined from multiple point sources spanning different intrapixel positions. In either case, the point-response function is determined on an image grid that is finer than the data grid.

2.10. Numerical Considerations

We conclude the mathematical preliminaries by noting that the full matrix containing the point-response function is usually prohibitively large. A modern 1024×1024 detector array yields a data set of 10^6 elements, and H contains 10^{12} elements. Clearly, one must avoid schemes that require the use of the entire point-response function matrix. Fortunately, they do not all need to be stored in computer memory,

nor do all need to be used in the matrix multiplication of Equation 8. The number of nonnegligible elements is often a small fraction of the total, and sparse matrix storage can be used (e.g., Press et al. 2002). The point-response function may also exhibit symmetries, such as in the case of convolution (Equation 9), which enables more efficient storage and computation. Alternatively, because H always appears as a matrix multiplication operator, one can write functions that compute the multiplication on the fly without ever storing the matrix values in memory. Such computations can take advantage of specialized techniques, such as fast Fourier transforms (Section 3.1) or small-kernel deconvolutions (Section 3.2).

3. NONITERATIVE IMAGE RECONSTRUCTION

A noniterative method for solving the inverse problem is one that derives a solution through an explicit numerical manipulation applied directly to the measured data in one step. The advantages of the noniterative methods are primarily ease of implementation and fast computation. Unfortunately, noise amplification is hard to control.

3.1. Fourier Deconvolution

Fourier deconvolution is one of the oldest and numerically fastest methods of image deconvolution. If the noise can be neglected, then the image can be determined using a discrete variant of the Fourier deconvolution (Equation 4), which can be computed efficiently using fast Fourier transforms (e.g., Press et al. 2002). The technique is used in speckle image reconstruction (Jones & Wykes 1989; Ghez, Neugebauer & Matthews 1993), Fourier-transform spectroscopy (Abrams et al. 1994, Prasad & Bernath 1994, Serabyn & Weisstein 1995), and the determination of galaxy redshifts, velocity dispersions, and line profiles (Simkin 1974, Sargent et al. 1977, Bender 1990).

Unfortunately, the Fourier deconvolution technique breaks down when the noise may not be neglected. Noise often has significant contribution from high $\mathbf{k}$, e.g., white noise has equal contributions from all $\mathbf{k}$. But $\tilde{H}(\mathbf{k})$, which appears in the denominator of Equation 4, falls off rapidly with $\mathbf{k}$. The result is that high-$\mathbf{k}$ noise in the data is significantly amplified by the deconvolution and creates image artifacts. The wider the point-response function, the faster $\tilde{H}(\mathbf{k})$ falls off at high $\mathbf{k}$ and the greater the noise amplification. Even for a point-response function extending over only a few pixels, the artifacts can be so severe that the image is completely lost in them.

3.2. Small-Kernel Deconvolution

Fast Fourier transforms perform convolutions very efficiently when used on standard desktop computers but they require the full data frame to be collected before the computation can begin. This is a great disadvantage when processing raster video in pipeline fashion as it comes in, because the time to collect an entire

data frame often exceeds the computation time. Pipeline convolution of raster data streams is more efficiently performed by massively parallel summation techniques, even when the kernel covers as much as a few percent of the area of the frame. In hardware terms, a field-programmable gate array (FPGA) or an application-specific integrated circuit (ASIC) can be much more efficient than a digital signal processor (DSP) or a microprocessor unit (MPU). FPGAs or ASICs available commercially can be built to perform small-kernel convolutions faster than the rate at which raster video can straightforwardly feed them, which is currently up to ∼150 megapixels per second.

Pipeline techniques can be used in image reconstruction by writing deconvolutions as convolutions by the inverse H^{-1} of the point-response function

$$I = H^{-1} * D, \tag{12}$$

which is equivalent to the Fourier deconvolution (Equation 4). But H^{-1} extends over the entire array, even if H is a small kernel (spans only a few pixels). Not to be thwarted, one then seeks an approximate inverse kernel $G \approx H^{-1}$, which can be designed to span only ∼3 full widths at half maximum of H.

Moreover, G can also be designed to suppress the high-$\mathbf{k}$ components of H^{-1} and to limit ringing caused by sharp discontinuities in the data, thereby reducing the artifacts created by straight Fourier methods.

3.3. Wiener Filter

In Section 3.1, we saw that deconvolution of the data results in strong amplification of high-$\mathbf{k}$ noise. The problem is that the signal decreases rapidly at high $\mathbf{k}$, while the noise is usually flat (white) and does not decay with $\mathbf{k}$. In other words, the high-$\mathbf{k}$ components of the data have poor signal-to-noise ratios.

The standard way to improve $\mathbf{k}$-dependent signal-to-noise ratio is linear filtering, which has a long history in the field of signal processing and has been applied in many areas of science and engineering. The Fourier transform of the data $\tilde{D}(\mathbf{k})$ is multiplied by a $\mathbf{k}$-dependent filter Φ ($\mathbf{k}$), and the product is transformed back to provide filtered data. Linear filtering is a particularly useful tool in deconvolution, because the filtering can be combined with the Fourier deconvolution (Equation 4) to yield the filtered deconvolution,

$$\tilde{I}(\mathbf{k}) = \Phi(\mathbf{k}) \frac{\tilde{D}(\mathbf{k})}{\tilde{H}(\mathbf{k})}. \tag{13}$$

It can be shown (e.g., Press et al. 2002) that the optimal filter, which minimizes the difference (in the least squares sense) between the filtered noisy data and the true signal, is the Wiener filter, expressed in Fourier space as

$$\Phi(\mathbf{k}) = \frac{\langle |\tilde{D}_0(\mathbf{k})|^2 \rangle}{\langle |\tilde{D}_0(\mathbf{k})|^2 \rangle + \langle |\tilde{N}(\mathbf{k})|^2 \rangle} = \frac{\langle |\tilde{D}_0(\mathbf{k})|^2 \rangle}{\langle |\tilde{D}(\mathbf{k})|^2 \rangle}. \tag{14}$$

Here $\langle |\tilde{\mathrm{N}}(\mathbf{k})|^2 \rangle$ and $\langle |\tilde{D}_0(\mathbf{k})|^2 \rangle$ are the expected power spectra (also known as spectral densities) of the noise and the true signal, respectively. Their sum, which

appears in the denominator of Equation 14, is the power spectrum of the noisy data $\langle|\tilde{D}(\mathbf{k})|^2\rangle$, because the signal and the noise are—by definition—statistically independent, so their power spectra add up in quadrature.

The greatest difficulty in determining $\Phi(\mathbf{k})$ (Equation 14) comes in estimating $\langle|\tilde{D}_0(\mathbf{k})|^2\rangle$. In many cases, however, the noise is white and easily estimated at high $\mathbf{k}$, where the signal is negligible. In practice, it is necessary to average over many $\mathbf{k}$ values, because the statistical fluctuation of any individual Fourier component is large. The power spectrum of the signal is then determined from the difference $\langle|\tilde{D}_0(\mathbf{k})|^2\rangle = \langle|\tilde{D}(\mathbf{k})|^2\rangle - \langle|\tilde{N}(\mathbf{k})|^2\rangle$. Again, averaging is needed to reduce the statistical fluctuations.

A disadvantage of the Wiener filter is that it is completely deterministic and does not leave the user with a tuning parameter. It is therefore useful to introduce an ad hoc parameter β into Equation 14 to allow the user to adjust the aggressiveness of the filter.

$$\Phi(\mathbf{k}) = \frac{\langle|\tilde{D}_0(\mathbf{k})|^2\rangle}{\langle|\tilde{D}_0(\mathbf{k})|^2\rangle + \beta\langle|\tilde{N}(\mathbf{k})|^2\rangle}. \quad (15)$$

Standard Wiener filtering is obtained with $\beta = 1$. Higher values result in more aggressive filtering, whereas lower values yield a smaller degree of filtering.

3.4. Wavelets

We saw in Section 3.1 and 3.3 that the Fourier transform is a very convenient way to perform deconvolution, because convolutions are simple products in Fourier space. The disadvantage of the Fourier spectral functions is that they span the whole image and cannot be localized. One might wish to suppress high $\mathbf{k}$ more in one part of the image than in another, but that is not possible in the Fourier representation. The alternative is to use other, more localized spectral functions. These functions are no longer eigenfunctions of a convolving point-response function, so the image reconstruction is not as simple as in the Fourier case, but they might still retain the Fourier characteristics, at least approximately. How are we to choose those functions? On the one hand we wish more localization. On the other hand, we want to characterize the spectral functions by the rate of spatial oscillation, because we know that we need to suppress the high-$\mathbf{k}$ components. Of course, the nature of the Fourier transform is such that there are no functions that are perfectly narrow in both image space and Fourier space (the uncertainty principle). The goal is to find a useful compromise.

The functions to emerge from the quest for oscillatory spectral functions with local support have been wavelets. The most frequently used wavelets are those belonging to a class discovered by Daubechies (1988). In addition to striking a balance between $\mathbf{x}$ and $\mathbf{k}$ support, they satisfy the following three conditions: (*a*) they form an orthonormal set, which allows easy transformation between the spatial and spectral domains, (*b*) they are translation invariant, i.e., the same function can be used in different parts of the image, and (*c*) they are scale invariant, i.e., they form a hierarchy in which functions with larger wavelengths are scaled-up

versions of functions with shorter wavelengths. These three requirements have the important practical consequence that the wavelet transform of n data points and its inverse can each be computed hierarchically in $O\ (n\ \log_2\ n)$ operations, just like the fast Fourier transform (e.g., Press et al. 2002).

A more recent development has been a shift to nonorthogonal wavelets known as à trous (with holes) wavelets (Holschneider et al. 1989; Shensa 1992; Bijaoui, Starck & Murtagh 1994; Starck, Murtagh & Bijaoui 1998). The wavelet basis functions are redundant, with more wavelets than there are data points. Their advantage is that the wavelet transform consists of a series of convolutions, so each wavelet coefficient is a Fourier filter. Of course, each à trous wavelet, being localized in space, corresponds to a range of k, so the wavelets are not eigenfunctions of the point-response function. Also the à trous wavelet noise spectrum needs to be computed carefully, because the à trous wavelets are redundant and nonorthogonal. Spatially uncorrelated noise, which is white in Fourier space, is not white in à trous wavelet space. (It is white for orthonormal wavelets, such as Daubechies wavelets.)

Wavelet filtering is similar to Fourier filtering and involves the following: Wavelet-transform the data to the spectral domain, attenuate or truncate wavelet coefficients, and transform back to data space. The wavelet filtering can be as simple as truncating all coefficients smaller than $m\sigma$, where σ is the standard deviation of the noise. Alternatively, soft thresholding reduces the absolute values of the wavelet coefficients (Donoho & Johnstone 1994, Donoho 1995a). A further refinement is to threshold high **k** more strongly (Donoho 1995b) or to modify the high-**k** wavelets to reduce their noise amplification (Kalifa, Mallat & Rouge 2003). Yet another possibility is to apply a wavelet filter analogous to the Wiener filter (Equation 14).

Once the data have been filtered, deconvolution can proceed by the Fourier method or by small-kernel deconvolution. Of course, the deconvolution cannot be performed in wavelet space, because the wavelets, including the à trous wavelets, are not eigenfunctions of the point-response function. Wavelet filtering can also be combined with iterative image reconstruction (Section 8.2).

3.5. Quick Pixon

The Pixon method is another way to obtain spatially adaptive noise suppression. We defer the comprehensive discussion of the Pixon method and its motivation to Sections 8.5 and 8.6. Briefly, it is an iterative image restriction technique that smoothes the image model in a spatially adaptive way. A faster variant is the quick Pixon method, which applies the same adaptive Pixon smoothing to the data instead of to image models. This smoothing can be performed once on the input data, following which the data can be deconvolved using the Fourier method or small-kernel deconvolution.

The quick Pixon method, though not quite as powerful as the full Pixon method, nevertheless often results in reconstructed images that are nearly as good as those of the full Pixon method. The advantage of the quick Pixon method is its speed.

Because the method is noniterative and consists primarily of convolutions and deconvolutions, the computation can be performed in pipeline fashion using small-kernel convolutions. This allows one to build special-purpose hardware to process raster video in real time at the maximum available video rates.

3.6. Discussion

The performance of Wiener deconvolution can be assessed from the reconstructed images shown in Figure 1. For this example a 128 $\times$ 128 synthetic truth image, shown in panel (*b*) of the figure, is blurred by a Gaussian point-response function with a full width at half maximum of 4 pixels. Constant Gaussian noise is added to this blurred image, so that the brightest pixels of all of the synthetic sources yield a peak signal-to-noise ratio per pixel of 50. The resulting input data are shown in panel (*a*) of the figure.

Next, the *central column* of panels shows a Wiener reconstruction and associated residuals when less aggressive filtering is chosen by setting $\beta = 0.1$ (Equation 15). This yields greater recovered resolution and good, spectrally white residuals but at the expense of large noise-related artifacts that appear in the reconstructed image. In fact, noise amplification makes these artifacts so large as to risk confusion with real sources in the image. This illustrates the major difficulty of image ambiguity, which we emphasized from the start. The reconstructed image in panel (*c*) results in reasonable residuals, similar to those that would be obtained from the truth image in panel (*b*), because blurring by the point-response function suppresses the differences between the two images, and differences in the data model fall below the measurement noise. We reject panel (*c*) compared with panel (*b*) not because it fits the data less well, but because we know on the basis of other knowledge (experience) that it is a less plausible image. In an effort to improve the reconstruction, we might choose more aggressive filtering with $\beta = 10$, as in the Wiener reconstruction that appears in the *right-hand column*. Here the image artifacts are less troublesome but the resolution is poorer and the residuals now show significant correlation with the signal.

The improvement in resolution brought about by a selection of reconstructions is shown in Table 1, which lists the full widths at half maximum of two sources from Figure 1 with good signal-to-noise ratios. Shown are the widths of the sources in the truth image, the data, and the reconstructions. The Wiener reconstructions improve resolution by $\sim$1 pixel. This may be compared with other reconstructions not shown in Figure 1. The quick Pixon method (Section 3.5) and the nonnegative least-squares fit (Section 6.2) reduce the width by $\sim$1.5 pixels, whereas the full Pixon method (Section 8.6) reduces the true widths by $\sim$2.5 pixels, restoring the true widths of the sources. Bear in mind also that widths add in quadrature. A more appropriate assessment of the resolution boost is therefore made by considering the reduction in the squares of the widths in Table 1.

Figure 2 shows Wiener, wavelet, and quick Pixon reconstructions of simulated data obtained from a real image of New York City by blurring it using a Gaussian point-response function with full width at half maximum of four pixels and adding

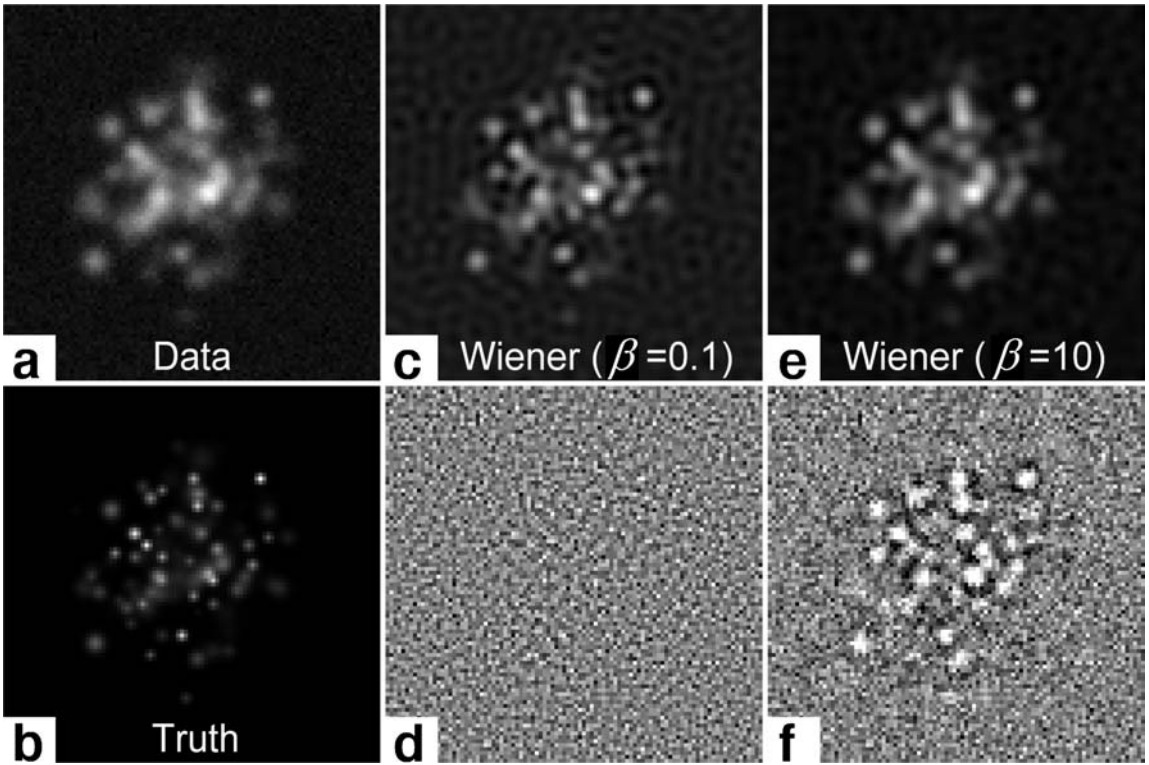

Figure 1 Wiener reconstructions of a synthetic image: (*a*) data, (*b*) truth image, and (*c*) weak filtering ($\beta = 0.1$), overfitting the data, with (*d*) residuals ($\chi^2/n = 0.89$), and (*e*) strong filtering ($\beta = 10$), underfitting the data, with (*f*) residuals ($\chi^2/n = 1.15$).

TABLE 1 Resolution improvement of image reconstruction techniques measured by the full widths at half maximum (in pixels) of two bright sources in Figure 1

Source	Truth	Data	Wiener $\beta = 0.1$	Wiener $\beta = 10$	Quick Pixon	Nonnegative least-squares	Full Pixon
1	1.83	4.40	2.91	3.50	2.67	2.70	1.90
2	1.81	4.39	3.17	3.85	3.11	2.77	1.78

Gaussian noise so the peak signal-to-noise ratio per pixel is 50. The *top panels* show, from left to right, a standard Wiener deconvolution with $\beta = 1$, a wavelet reconstruction with Wiener-like filtering with $\beta = 2$, and a quick Pixon reconstruction. The wavelet and quick Pixon deconvolutions are performed by a small kernel of 15×15 pixels (Section 3.2). The truth image and the data are not shown here for lack of space, but are shown in Figure 3.

The Wiener reconstruction shows excellent residuals but the worst image artifacts. The wavelet reconstruction shows weaker artifacts, but the residuals are poor, particularly at sharp edges. One can try to change β, but this only makes matters worse. The choice of $\beta = 2$ is our best compromise between more artifacts at lower threshold and poorer residuals at higher threshold. The quick Pixon reconstruction fares best. The residuals are tolerable, although somewhat worse than those of the wavelet reconstruction. The main advantage of the quick Pixon reconstruction is the low artifact level. For that reason, that image presents the best overall visual acuity.

The need to find a good tradeoff between resolution and artifacts is universal for noniterative image reconstructions and invites the question whether better techniques are available that simultaneously yield high resolution, minimal image artifacts, and residuals consistent with random noise. The search for such techniques has led to the development of iterative methods of solution as discussed in the next several sections.

4. ITERATIVE IMAGE RECONSTRUCTION

4.1. Statistics in Image Reconstruction

We saw in Section 3 that even though the noniterative methods take into account the statistical properties of the noise (with the exception of direct Fourier deconvolution), the requirement that image reconstruction be completed in one step prevents full use of the statistical information. Iterative methods are more flexible and can go a step further, allowing us to fit image models to the data. They thus infer an explanation of the data based upon the relative merits of possible solutions. More precisely, we consider a defined set of potential models of the image. Then, with the help of statistical information, we choose amongst these models the one that is the most statistically consistent with the data.

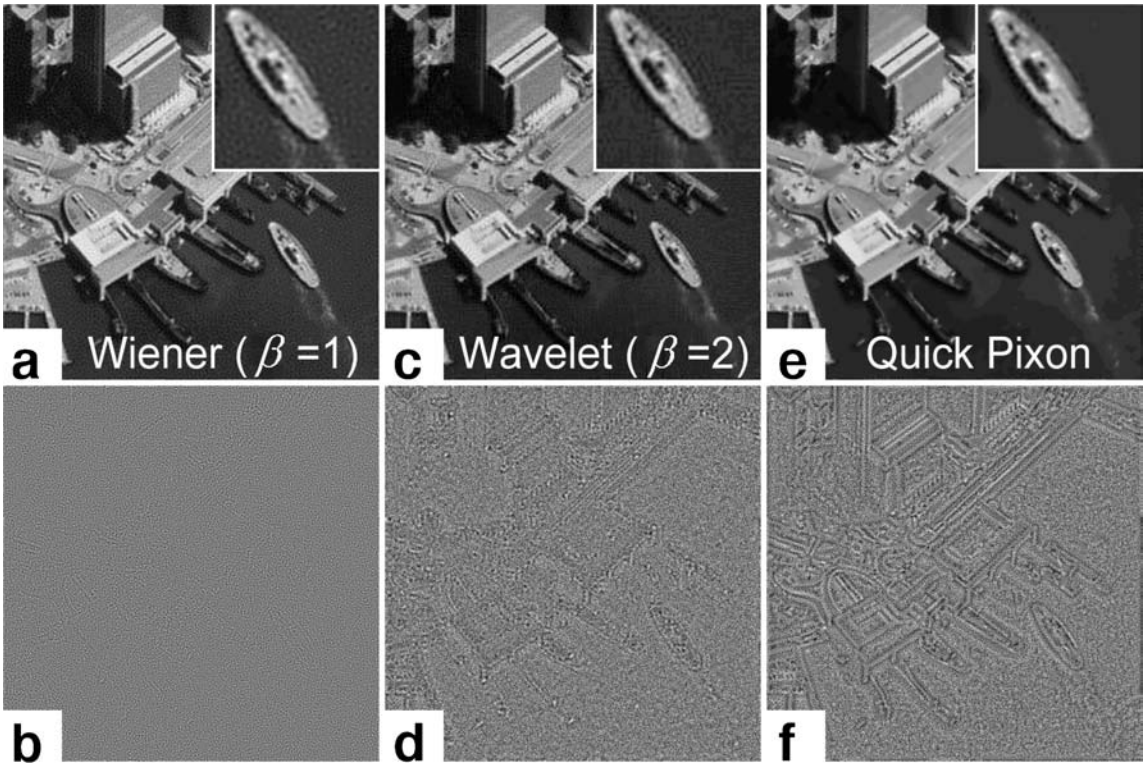

Figure 2 Variety of noniterative image reconstructions: (*a*) Wiener ($\beta = 1$) with (*b*) residuals, (*c*) wavelet ($\beta = 2$) with (*d*) residuals, and (*e*) quick Pixon with (*f*) residuals. The data and the truth image are shown in Figure 3.

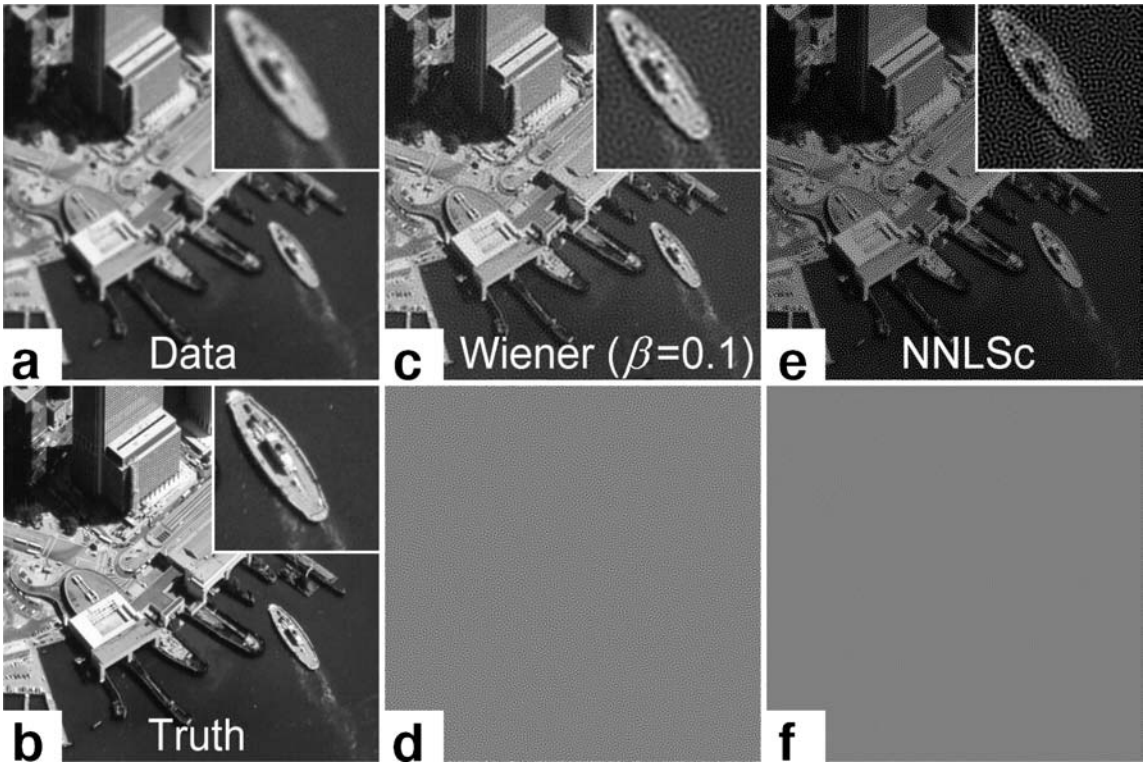

Figure 3 Converged nonnegative least-squares fit compared with weak Wiener filtering: (*a*) data, (*b*) truth image, (*c*) Wiener filter ($\beta = 0.1$) with (*d*) residuals ($\chi^2/n = 0.88$), and (*e*) converged nonnegative least-squares fit with (*f*) residuals ($\chi^2/n = 0.76$).

Consistency is obtained by finding the image model for which the residuals form a statistically acceptable random sample of the parent statistical distribution of the noise. The data model is then our estimate of the reproducible signal in the measurements, and the residuals are our estimate of the irreproducible statistical noise. Note that the residuals need not all have an identical parent statistical distribution, e.g., the standard deviation of the residuals may vary from one pixel to the next. But there must be a well-defined statistical law that governs all of them, and we have to know it, at least approximately, in order to fit the data.

There are three components of data fitting (e.g., Press et al. 2002). First, there must be a fitting procedure to find the image model. This is done by minimizing a merit function, often subject to additional constraints. Second, there must be tests of goodness of fit—preferably multiple tests—that determine whether the residuals obtained are consistent with the parent statistical distribution. Third, one would like to estimate the remaining errors in the image model.

To clarify what each of those components of data fitting is, consider the familiar example of a linear regression. We might determine the regression coefficients by finding the values that minimize a merit function consisting of the sum of the squares of the residuals. Then we check for goodness of fit in a variety of ways. One method is to consider the minimum value of the same sum of squares of the residuals that we used for our merit function. But this time we ask a different question, not what values of the coefficients minimize it, but whether the minimum sum of squares found is consistent with our estimate of the noise level. We also want to insure that the residuals are randomly distributed. Nonrandom features might indicate that the linear fit is insufficient and that we should add parabolic or other high-order terms to our fitting functions. In addition, we want to check that the distribution (histogram) of residual values follows the parent statistical distribution of the noise within expected statistical fluctuations. We suspect the fit if the mean of the residuals is significantly nonzero, or if their distribution is skewed or has unexpectedly strong or weak tails. Finally, once we find a satisfactory fit, we wish to know the uncertainty in the derived parameters, i.e., the scatter of values that we would find by performing linear regressions of multiple, independent data sets.

The same procedures are used in image reconstruction and are geared to the parent statistical distribution of the noise, because the goal is to produce residuals that are statistically consistent with that distribution. The merit function is usually the log-likelihood function described in Section 4.2, to which are added a host of image restrictions (Sections 6–8). Goodness of fit is diagnosed by the χ^2 statistic and by considering the statistical distribution and spatial correlations of the residuals. We check for spatially uncorrelated residuals with zero mean, standard deviation corresponding to the noise level of the data, and no unexpected skewness or tail distributions.

The precision of the image and error estimates are much harder to obtain. Visual inspection can be deceiving. In a good reconstruction, the image shows fewer fluctuations than the data. Conversely, a poor reconstruction may create significant

artifacts, whose amplitudes may exceed the noise level of the data. In neither case does the noise magically change. What is happening is that the image intensities are strongly correlated. The differences between the image intensities in neighboring pixels may be smaller or bigger than in the data, but that is because they have correlated errors. To compute error propagation in a reconstructed image analytically is next to impossible, so Monte Carlo simulations are the only realistic way to assess the errors of measurements made on the reconstructed image. An alternative is to fit the desired image features parametrically, because parametric fits have a built-in mechanism for error estimates, even in the presence of a nonparametric background (Section 5.2).

4.2. Maximum Likelihood

A given image model I results in a data model M (Equations 8 or 9). The parent statistical distribution of the noise in turn determines the probability of the data given the data model $p(D|M)$. This is then the conditional probability of the data given the image $p(D|I)$. The most common parent statistical distributions are the Gaussian (or normal) distribution and the Poisson distribution. The noise in different pixels is statistically independent, and the joint probability of all the pixels is the product of the probabilities of the individual pixels. The Gaussian probability is

$$p(D|I) = \prod_i \left(2\pi\sigma_i^2\right)^{-1/2} e^{-(D_i - M_i)^2/2\sigma_i^2}, \tag{16}$$

and the discrete Poisson distribution is

$$p(D|I) = \prod_i e^{-M_i} M_i^{D_i}/D_i!. \tag{17}$$

If there are correlations between pixels, $p(D|I)$ is a more complicated function.

In practice, it is more convenient to work with the log-likelihood function, a logarithmic quantity derived from the likelihood function:

$$\Lambda = -2\ln\left[p(D|I)\right] = -2\sum_i \ln\left[p(D_i|I)\right], \tag{18}$$

where the second equality in Equation 18 applies to statistically independent data. The factor of two is added for convenience, to equate the log-likelihood function with χ^2 for Gaussian noise and to facilitate parametric error estimation (Section 5).

The goal of data fitting is to find the best estimate $\hat{I}$ of I such that $p(D|\hat{I})$ is consistent with the parent statistical distribution. The maximum-likelihood method selects the image model by maximizing the likelihood function or, equivalently, minimizing the log-likelihood function (Equation 18). This method is known in statistics to provide the best estimates for a broad range of parametric fits in the limit in which the number of estimated parameters is much smaller than the number of data points (e.g., Stuart, Ord & Arnold 1998). We consider such parametric fits

first in Section 5. Most image reconstructions, however, are nonparametric, i.e., the "parameters" are image values on a grid, and their number is comparable to the number of data points. For these methods, maximum likelihood is not a good way to estimate the image and can lead to significant artifacts and biases. Nevertheless, it continues to be used in image reconstruction, but with additional image restrictions designed to prevent the artifacts. A major part of this review is devoted to nonparametric methods (Sections 6–8).

5. PARAMETRIC IMAGE RECONSTRUCTION

5.1. Simple Parametric Modeling

Parametric fits are always superior to other methods, provided that the image can be correctly modeled with known functions that depend upon a few adjustable parameters. One of the simplest parametric methods is a least-squares fit minimizing χ^2, the sum of the residuals weighted by their inverse variances:

$$\chi^2 = \sum_i \frac{R_i^2}{\sigma_i^2} = \sum_i \frac{(D_i - M_i)^2}{\sigma_i^2}. \tag{19}$$

For a Gaussian parent statistical distribution (Equation 16), the log-likelihood function, after dropping constants, is actually χ^2, so the χ^2 fit is also a maximum-likelihood solution.

For a Poisson distribution, the log-likelihood function, after dropping constants, is

$$\Lambda = 2 \sum_i (M_i + D_i \ln M_i), \tag{20}$$

a logarithmic function, whose minimization is a nonlinear process. The log-likelihood function also cannot be used for goodness-of-fit tests. One can write χ^2-like merit functions, but parameter estimation based on these statistics is usually biased by about a count per pixel, which can be a significant fraction of the flux at low counts. This bias is removed by Mighell (1999), who adds correction terms to both the numerator and denominator,

$$\chi_\gamma^2 = \sum_i \frac{[D_i + \min(D_i, 1) - M_i]^2}{D_i + 1}, \tag{21}$$

and shows that parameter estimation using this statistic is indeed unbiased.

5.2. Error Estimation

Fitting χ^2 has two additional advantages: The minimum χ^2 is a measure of goodness of fit, and the variation of the χ^2 around its minimum value can be used to estimate the errors of the parameters (e.g., Press et al. 2002). Here we wish to

emphasize the distinction between "interesting" and "uninteresting" parameters, and the role they play in image error estimation.

A convenient way to estimate the errors of a fit with p parameters is to draw a confidence limit in the p-dimensional parameter space, a hypersurface surrounding the fitted values on which there is a constant value of χ^2. If $\Delta\chi^2 = \chi^2 - \chi^2_{\min}$ is the difference between the value of χ^2 on the hypersurface and the minimum value found by fitting the data, then the tail probability α that the parameters would be found outside this hypersurface by chance is approximately given by a χ^2 distribution with p degrees of freedom (Press et al. 2002).

$$\alpha \approx P(\Delta\chi^2, p). \tag{22}$$

Equation 22 is approximate because, strictly speaking, it applies only to a linear fit with Gaussian noise, for which χ^2 is a quadratic function of the parameters and the hypersurface is an ellipsoid. It is common practice, however, to adopt Equation 22 as the confidence limit even when the errors are not Gaussian, or the fit is nonlinear, and the hypersurface deviates from ellipsoidal shape.

Parametric fits often contain a combination of q "interesting" parameters and $r = p - q$ "uninteresting" (sometimes called "nuisance") parameters. To obtain a confidence limit for only the interesting parameters, without any limits on the uninteresting parameters, one determines the q-dimensional hypersurface for which

$$\alpha \approx P(\Delta\chi^2, q). \tag{23}$$

The only proviso is that in computing $\Delta\chi^2$ for any set of interesting parameters q, χ^2 is optimized with respect to all the uninteresting parameters (Avni 1976, Press et al. 2002). A special case is that of a single interesting parameter ($q = 1$). The points at which $\Delta\chi^2 = m^2$ are then the $m\sigma$ error limits of the parameter. In particular, the 1σ limit is found where $\Delta\chi^2 = 1$.

Unfortunately, the errors of the nonparametric fits (Sections 6–8) cannot be estimated in this way, because of the difficulties in assigning a meaning to χ^2 in nonparametric fits (Section 6.1). This leaves Monte Carlo simulations as the only way to assess errors in the general nonparametric case. There is, however, the hybrid case of a combined parametric and nonparametric fit, in which the errors of the nonparametric part of the fit are of no interest. For example, we might wish to measure the positions and fluxes of stars in the presence of a background nebula, which is not interesting in its own right but affects the accuracy of our astrometry and photometry. In that case, we can perform a parametric fit of the interesting astrometric and photometric parameters, optimizing the background nonparametrically as "uninteresting" parameters.

5.3. Clean

Parameter errors are also important in models in which the total number of parameters is not fixed. The issue here is to determine when the fit is good enough and

additional parameters do not significantly improve it. The implicit assumption is that a potentially large list of parameters is ordered by importance according to some criterion, and the fit should not only determine the values of the important parameters but also decide the cutoff point beyond which the remaining, less important parameters may be discarded. For example, we may wish to fit the data to a series of point sources, starting with the brightest and continuing with progressively weaker sources, until the addition of yet another source is no longer statistically significant.

Clean, an iterative method that was originally developed for radio-synthesis imaging (Högbom 1974), is an example of parametric image reconstruction with a built-in cutoff mechanism. Multiple point sources are fitted to the data one at a time, starting with the brightest sources and progressing to weaker sources, a process described as "cleaning the image." In its simplest form, the Clean algorithm consists of four steps. Start with a zero image. Second, add to your image a new point source at the location of the largest residual. Third, fit the data for the positions and fluxes of all the point sources introduced into your image so far. (The image consists of a bunch of point sources. The data model is a sum of point-spread functions, each scaled by the flux of the point source at its center.) Fourth, return to the second step if the residuals are not statistically consistent with random noise.

Clean has enabled synthesis imaging of complicated fields of radio sources even with limited coverage of the Fourier (u,v) plane (see Thompson, Moran & Swenson 2001). There are many variants of the method. Clark (1980) selects many components and subtracts them in a single operation, rather than separately, thereby substantially reducing the computational effort by limiting the number of convolutions between image space and the (u,v) plane. Cornwell (1983) introduces a regularization term (Section 7.1) to reduce artifacts in reconstructions of extended sources due to the incomplete coverage of the (u,v) plane. Steer, Dewdney, & Ito (1984) propose a number of additional changes to prevent the formation of stripes or corrugations during the processing of images with extended features and arithmetic rounding problems. Phase closure (Pearson & Readhead 1984) provides additional constraints. But the problem remains that Clean fits extended objects a pixel at a time. This has led to methods that use multiple scales (Wakker & Schwarz 1988, Bhatnagar & Cornwell 2004).

6. NONPARAMETRIC IMAGE RECONSTRUCTION

Despite their great power in performing high-quality and robust image reconstructions, the use of parametric methods is severely restricted by the requirement that explicit functions be identified with which to model the image. In short, significant prior knowledge of image features is required. In this section we relax this restriction and introduce nonparametric methods for the estimation of image models. A general feature of such models is that the number of model values to be determined can be comparable to or exceed the number of data points. In the simplest case,

a nonparametric method accomplishes this by defining an image model on a grid of pixels equal in size to that of the data. The method must then by some means determine image values for all pixels in the image grid. In the worst case, each image value may be individually and independently adjustable.

Clearly, the step from parametric to nonparametric modeling is a drastic one that yields a combinatorial explosion of possible image models. In fact, nonparametric methods draw from a pool of possible image models that is much too general. Recalling from Section 2.8 our assertion that the inverse problem is ill conditioned, such generality proves especially challenging when the signal-to-noise ratio is low. One expects that because the space of potential solutions is so large, reconstruction artifacts will abound.

The result is that iterative nonparametric methods that enforce no restrictions on image models are often no better at controlling noise than the noniterative methods presented in Section 3. Obtaining both image generality and good noise control thus requires inclusion of additional constraints for limiting the permissible image models. In this and subsequent sections, we present a series of nonparametric methods that differ only in the means by which these constraints are designed and enforced and how the solution is found.

The iterative methods usually use the log-likelihood function as their merit function, despite its inadequacy for nonparametric fits, but they restrict its minimization in different ways. Some stop the fitting procedure before the merit function is fully minimized, some impose restrictions on permitted image values, some create a new merit function by adding to the log-likelihood function an explicit penalty function, which steers the solution away from unwanted image models, and some do more than one of these things. In this section we first present two constraint methods, early termination of the fit and enforcement of nonnegative image values, and then discuss a few iterative schemes to fit the log-likelihood function, known in the statistics literature as expectation-maximization methods (Dempster, Laird & Rubin 1977). In Section 7 we discuss global image restriction by means of a global penalty function, which serves to regularize the solution, allowing it to converge to a reasonable solution. Finally, Section 8 is devoted to spatially adaptive methods to restrict the image.

6.1. Early Termination of the Fit

Carried to completion, a nonparametric maximum-likelihood fit can result in zero residuals. For example, if the image and the data are defined on the same grid, then a nonnegative point-response function is a nonsingular, square matrix, which has an inverse. The maximum-likelihood solution is therefore the one for which the residuals are identically zero, as in Fourier deconvolution (Section 3.1). This solution, however, is far from optimal if the noise is expected to have a finite standard deviation. A set of zero residuals is hardly a statistically acceptable sample of the parent statistical distribution. The problem is that the maximum-likelihood method was designed for problems in which the number of unknown parameters

is much smaller than the number of data points, and we use it to solve a problem in which they are comparable or even equal.

One way to avoid letting the residuals become too small in an iterative fit is to terminate the fit before this happens. A fit might be stopped when a goodness-of-fit measure, such as the χ^2, falls below a designated value. But what is that value for a χ^2 statistic? The expectation value of χ^2 is the number of degrees of freedom, equal to the difference between the number of data points and the number of parameters, but what is the number of parameters? In the above example, the number of data points is equal to the number of image points, so the number of degrees of freedom is technically zero, and we should let the fit run to completion. But that is not what we would like to do.

We take the opposite point of view, placing a higher premium on avoiding noise amplification and spurious artifacts than on seeking a "perfect" fit that interprets statistical noise as real signal. If the image model were the true image the correct stopping point would be when χ^2 equals the number of data points n. We prefer to go even further and conservatively stop the fit earlier, when χ^2 reaches $n + \sqrt{2n}$, a point higher by one standard deviation of the χ^2 for n degrees of freedom.

There might be some concern about an iterative method that is not carried out to convergence. First, the result may depend on the initial image. In practice, this is rarely a problem. We normally set the initial image to be identically zero and find adequate fits. Second, the stopping criterion is a global condition. The solution might, in fact, overfit the data in some areas and underfit in other areas. This does happen and is one of the main reasons for adopting spatially adaptive reconstruction methods that limit the image locally and not globally (Section 8). Fortunately, it is easy to identify uneven fits by displaying the residuals.

6.2. Nonnegative Least-Squares

A simple constraint that greatly increases the performance of a maximum-likelihood method is to disallow negative image values. When applied to a least-squares fit, this procedure is known as a nonnegative least-squares fit. Nonnegativity is certainly a necessary restriction for almost all images. (There are exceptions, e.g., image reconstruction in the complex Fourier space.) But forcing the image to be nonnegative also strongly suppresses artifacts. A qualitative argument that supports this idea is that if the image contains both large positive and large negative fluctuations on length scales smaller than the width of the point-response function, then these fluctuations mutually cancel upon convolution with the point-response function. Restricting the image to nonnegative values thus also reduces the magnitude of the positive fluctuations. As a result, artifacts are significantly reduced.

The requirement that the image be nonnegative also increases resolution (Section 2.4). The degree of possible subpixel resolution depends on the signal-to-noise ratio and the width of the point-response function. Half-pixel resolution, and even quarter-pixel resolution, can often be obtained. When the structure of the source is known, e.g., a star is known to be a point source, it is possible to pinpoint

its position even better, often to a tenth of a pixel. Indeed, it may not be possible to find a good image reconstruction with an image model defined on the same grid as the data. It is not only feasible to extract subpixel information, it may be necessary to do so.

Procedures that impose nonnegativity include changes of variable and simply setting negative values to zero after each iteration. In our own work with iterative schemes that minimize the log-likelihood function we have found that (*a*) a change of variable can cause very slow convergence of image values residing near zero and (*b*) setting negative values to zero after each iteration does not hurt convergence and may actually speed it up (Section 6.6).

6.3. van Cittert

Having considered a couple of ways to restrict images, we next turn to iterative computational methods to find the image models. Instead of solving for all the unknown variables at once, one uses the approximate solution found in a previous iteration in order to compute the next iteration. In the statistics literature, such an iterative method is called an expectation-maximization method, because it alternates between substituting expectation values (the previous solution) for some of the variables in the likelihood function and maximizing the likelihood function with respect to the remaining unknowns. Performed correctly, expectation-maximization methods are guaranteed to converge (Dempster, Laird & Rubin 1977), but convergence can be slow.

The van Cittert (1931) method is one of the earliest and simplest iterative methods for image reconstruction problems in which the data and image are defined on the same grid. The iteration begins with the zeroth-order image $I^{(0)} \equiv 0$ at all grid points and iterates from there according to

$$I^{(k+1)} = I^{(k)} + \alpha(D - H \otimes I^{(k)}) = \alpha D + Q \otimes I^{(k)}, \tag{24}$$

where $Q = \mathbf{1} - \alpha H$, and $\mathbf{1}$ is the identity kernel. The iterations are designed to converge to the deconvolved image. Successive substitutions into Equation 24 yield

$$I^{(k)} = \alpha \sum_{j=0}^{k-1} Q^j \otimes D = H^{-1} \otimes (\mathbf{1} - Q^k) \otimes D \underset{k\to\infty}{\longrightarrow} H^{-1} \otimes D, \tag{25}$$

where Q^j denotes a j-fold convolution of the function Q with itself, the second equality represents the sum of the geometric series, and the limit $k \to \infty$ applies as long as $Q^k \otimes D \to 0$ in that limit.

The limiting solution has zero residuals, just as in the case of the Fourier deconvolution discussed in Section 3.1. (But note that the van Cittert method is not limited to deconvolutions.) If carried far enough, the van Cittert method therefore exhibits noise amplification just as do the Fourier-based methods, and the iteration must be terminated prior to convergence. The art of applying the van Cittert method

is in choosing a value of the parameter α and establishing a stopping criterion, so that the computation time, noise amplification, and degree of recovered resolution are acceptable. Although the convergence of the van Cittert iterations can be slow, solutions can be obtained especially quickly when the point-spread function is centrally peaked and relatively narrow (Lagendijk & Biemond 1991).

Numerous modifications to the technique include disallowing any negative image values, setting upper bounds to the image values, and more sophisticated methods that apply noise filters at select iterations (Agard 1984; Biemond, Lagendijk & Mersereau 1990; Wallace, Schaefer & Swedlow 2001). Such a modified version of the method has been commercially implemented for applications in 3D deconvolution in light microscopy (Wallace, Schaefer & Swedlow 2001). Other implementations use wavelet-based filtering at each iteration, removing statistically insignificant features from the solution (Section 8.2).

6.4. Landweber

Another iterative scheme (Landweber 1951) is:

$$I^{(k+1)} = I^{(k)} + \alpha H^T \otimes \frac{R}{\sigma^2}, \tag{26}$$

where the superscript T denotes the transpose operation, and α is a small positive parameter. This method is designed to minimize the sum of the squares of the residuals by insuring that the next change in the image, $\Delta I = I^{(k+1)} - I^{(k)}$, is in the direction of the negative of the gradient (negradient) of χ^2 with respect to I. The choice of α, however, is arbitrary and depends on the image. If it is too large, the iteration can overshoot the minimum along the negradient direction and even result in worse residuals. Indeed, workers using the method have found that it often initially produces a good solution but thereafter begins to diverge (Bertero & Boccacci 1998; Calvetti, Reichel & Zhang 1999).

In practice, users of the Landweber method often modify the procedure to avoid negative image values, which yields the projective Landweber method (Eicke 1992). Other simple constraints can be imposed using projection operators in either the spatial or spectral domains (Bertero & Boccacci 2000). Another variation is to modify α during the iteration (Liang & Xu 2003).

6.5. Richardson-Lucy

The Richardson-Lucy method (Richardson 1972, Lucy 1974, Shepp & Vardi 1982) was developed specifically for data comprising discrete, countable events that follow a Poisson distribution. The nonlinear log-likelihood function (Equation 20) is minimized iteratively using multiplicative corrections:

$$I^{(k+1)} = \left[H^T \otimes \left(\frac{D}{M^{(k)}} \right) \right] I^{(k)}. \tag{27}$$

The square brackets on the right-hand side of Equation 27 enclose the factor by which the previous $I^{(k)}$ is multiplied (not convolved) to give the new $I^{(k+1)}$. It results from a back projection operation, in which the ratio between the data, D, and the data model of the previous iteration, $M^{(k)} = H \otimes I^{(k)}$, is operated upon by H^T, the transpose (not the inverse) of the point-response function.

Lucy (1974) shows that the algorithm is flux conserving, maintains image non-negativity, and decreases the log-likelihood function in each iteration, at least if one takes only part of the step indicated by Equation 27. But the method yields noise-related artifacts when the signal-to-noise ratio is low (van Kempen et al. 1997).

Improvement can be achieved in a number of ways. Snyder & Miller (1991) first exaggerate deblurring by obtaining the maximum-likelihood solution for a point-response function that is deliberately broadened by convolving it with an extra sieve function. The solution, which is too sharp and may contain ringing, is then broadened by the same sieve function. Another approach is to modify the log-likelihood function by adding a penalty function along the lines discussed in Section 7 (Joshi & Miller 1993, Conchello & McNally 1996), modifying Equation 27 according to the general expectation-maximization procedure of Dempster, Laird & Rubin (1977).

6.6. Conjugate-Gradient

The iterative schemes described in Sections 6.3–6.5 are all designed to converge to the maximum-likelihood solution (and are stopped early to avoid overfitting the data), but their convergence is slow. Modern minimization techniques converge much faster by utilizing the Hessian matrix of second-order partial derivatives of the merit function with respect to the variables. Unfortunately, the Hessian matrix is too big to be computed for typical image reconstruction problems. One is therefore left with minimization schemes that collect and use the information contained in the Hessian matrix without ever computing the entire matrix.

An excellent example of such a technique is the conjugate-gradient method (e.g., Press et al. 2002). The method starts from some initial image $I^{(0)}$, where it computes the negative gradient (negradient) $g^{(0)}$ of the log-likelihood function with respect to the image and sets the initial conjugate-gradient direction $h^{(0)} = g^{(0)}$. It then constructs a sequence of negradients $g^{(k)}$ and conjugate-gradient directions $h^{(k)}$ as follows. First, it locates the minimum of the log-likelihood function along the conjugate-gradient direction $h^{(k)}$. Second, at the position of the minimum it computes the next negradient $g^{(k+1)}$. Third, it sets the new conjugate-gradient direction to a linear combination of the old conjugate-gradient direction and the new negradient

$$h^{(k+1)} = g^{(k+1)} + \gamma_k h^{(k)}. \tag{28}$$

The coefficient γ_k is chosen to optimize convergence. We generally prefer the one devised by Polak & Ribiere (1969):

$$\gamma_k = \frac{\sum_j \left(g_j^{(k+1)} - g_j^{(k)}\right) g_j^{(k+1)}}{\sum_j \left(g_j^{(k)}\right)^2}, \tag{29}$$

where the sums are over all the image points.

The stopping criterion for the conjugate-gradient minimization is similar to that of the slower methods. There is some evidence that the first iterations of the conjugate-gradient method introduce mainly low-**k** components into the solution, and components with higher **k** are added mainly in later iterations (Hansen 1994). Stopping the iterations in time therefore also provides a smoother solution and helps to reduce high-**k** noise amplification.

We have found that the most effective way to impose nonnegative solutions is to modify the conjugate-gradient method as follows. At each iteration of the conjugate gradient minimization, first compute the negradient. Second, check the negradient components of all the pixels whose image values are zero and set the negradient components to zero if they are negative, i.e., pointing toward negative image values. Third, compute the conjugate-gradient direction in the usual way. Fourth, find the minimum along the conjugate gradient direction without regard to the sign of the image. Fifth, truncate all negative image values to zero, thereby jumping to a new solution. Sixth, go back to the first step and continue with the next conjugate gradient iteration as though no truncation took place.

This procedure may seem ad hoc and liable to disrupt convergence, but the converse is true. The procedure, in fact, belongs to a class of iterative schemes called projections onto convex sets, which are guaranteed to converge (Biemond, Lagendijk & Mersereau 1990; Press et al. 2002). Occasionally, the truncation leads to an increase instead of a decrease in the value of the merit function. We have found that this is actually an advantage, because it enables the minimization to escape from local minima. For many minimizations we reach the stopping point in about 10 iterations. If the conjugate-gradient algorithm requires more iterations, it is a good idea to stop the conjugate-gradient iteration every 5–10 iterations and start it anew at that point, i.e., to set the conjugate-gradient direction in the direction of the negradient $h^{(j+1)} = g^{(j+1)}$.

Finally, we comment that, for a quadratic log-likelihood function, it is possible to solve for the position of the minimum along the conjugate-gradient direction analytically and proceed there in one step (before truncation for negative image values). For nonlinear log-likelihood functions it is necessary to search iteratively for the minimum, which requires that the log-likelihood function (but not its gradient) be computed several times along the conjugate-gradient direction. Convergence may also be accelerated by using preconditioners, replacing the negradients with vectors that point more closely in the direction of the function minimum. For a thorough discussion of these issues see Press et al. (2002). (For the linear case they actually present the biconjugate-gradient method; the conjugate-gradient method is a special case, which can be programmed more efficiently.)

7. GLOBAL IMAGE RESTRICTION

In Section 6 we introduced two ways to control noise-related artifacts in image reconstruction: early termination of iterative fits and enforcement of image nonnegativity. As we show in Section 8.7, even when both are employed, one is still unable simultaneously to suppress the artifacts and fit the data in an adequate manner. In short, the class of allowed solutions defined by nonparametric maximum-likelihood methods is still too large despite the benefits of these methods of image restriction. The remainder of this review considers additional constraints that can and should be brought to bear on image reconstruction. This section considers global image restrictions. Section 8 is devoted to the more powerful, spatially adaptive image restrictions.

7.1. Duality Between Regularization and Bayesian Methods

Two main approaches have been developed to impose global constraints on the solutions of ill-posed problems in general and image reconstruction in particular. One approach is to steer the solution away from unwanted images by modifying the merit function, adding a regularization term to the log-likelihood function to give:

$$\Lambda' = \Lambda + \lambda B(I). \tag{30}$$

Here $B(I)$ is a penalty function that increases with the degree of undesirability of the solution, and λ is a penalty normalization parameter that controls the relative strength of the penalty function with respect to the log-likelihood function. (We show in Section 7.2 that λ plays the role of a Lagrange multiplier.)

The other approach is to assign to each image model an a priori probability $p(I)$, also called a prior, and to maximize the product $p(D|I)p(I)$ of the likelihood function and the prior. This approach is motivated by the desire to maximize the conditional probability of the image given the data $p(I|D)$, known as the image a posteriori probability. Bayes' (1763) theorem is used to relate these quantities:

$$p(I|D) = \frac{p(D|I)p(I)}{p(D)} \propto p(D|I)p(I). \tag{31}$$

The data are fixed for any image reconstruction, so $p(D)$ is a constant and maximizing $p(I|D)$ amounts to maximizing the product $p(D|I)p(I)$. The image reconstruction is called a Bayesian method, and the solution is called the maximum a posteriori image. Expressed logarithmically, we obtain an expression similar to Equation 30:

$$\Lambda' = \Lambda - 2\ln[p(I)]. \tag{32}$$

One might imagine that the Bayesian approach is more restrictive, because the regularization term has an arbitrary penalty function with an adjustable normalization, whereas the prior image probability could have theoretical underpinning and

be completely specified in advance, without adjustable parameters. In reality, the choice of the prior is just as arbitrary, reflecting the preference of the practitioner for particular types of images. Moreover, even when the probabilities are set in some axiomatic way, as in the maximum-entropy method (Section 7.5), an adjustable parameter is again introduced, changing Equation 32 to a form equivalent to Equation 30:

$$\Lambda' = \Lambda - \lambda S(I). \tag{33}$$

Operationally, therefore, there is no difference between regularization and Bayesian methods. They both add a term to the log-likelihood function and minimize the modified merit function. The extra term can be positive and viewed as a penalty function or negative and viewed as a preference function. It amounts to the same thing.

Finally, we note in passing that in some of the Bayesian literature (e.g., Hoeting et al. 1999) the authors recommend using the average of the a posteriori image instead of the maximum:

$$\langle I \rangle = \frac{\int p(I|D) I \, dI}{\int p(I|D) \, dI}. \tag{34}$$

In practice, however, the evaluation of the average image from Equation 34 is computationally very costly. Furthermore, the effort may not be justified, because the a posteriori probability is sharply peaked, so the difference between the average and the mode is likely to be small.

7.2. Penalty Normalization as a Lagrange Multiplier

An additional benefit of regularization is that the fit of a properly regularized problem can be carried out to convergence. This may suggest that there is no longer any need for a stopping criterion. This is illusory. Although it is nice to have a converging fit, it must also produce residuals that are statistically consistent with the parent statistical distribution. That is, their χ^2 should be approximately equal to the number of data points (Morozov 1966). This is achieved by adjusting the penalty normalization parameter λ in Equation 30 or 33. In fact, one can think of global image restriction as a formulation of image preference subject to data constraint. We seek the best image, given our preference function, subject to one or more constraints imposed by the data. Viewed in this way, λ is a Lagrange multiplier adjusted to enforce the data constraint. (It makes little difference if the Lagrange multiplier multiplies the constraint or the preference function.) A subtle point is whether the data constraint should be a log-likelihood function or a goodness-of-fit function. The two are identical for Gaussian noise, of course, but they do differ for other types of noise. A Bayesian purist might opt for a log-likelihood function, given that the aim is to maximize the a posteriori probability. But a goodness of fit works just as well, e.g., Equation 21 for Poisson noise.

The use of χ^2 as a goodness-of-fit stopping criterion assumes advance knowledge of the standard deviations σ of the noise. When the noise level is not known in advance, it is possible to estimate it directly from the data. If the data model is sufficiently smooth, at least in parts of the image, it is possible to estimate σ from the standard deviation of the data in neighboring pixels. Care must be exercised to insure that the regions are indeed smooth (or σ will be systematically overestimated) and comprise enough pixels (so that the statistical error in the determination of σ is manageable).

Two other methods have been proposed to set the Lagrange multiplier λ. Generalized cross validation (Wahba 1977; Golub, Heath & Wahba 1979; Galatsanos & Katsaggelos 1992; Golub & von Matt 1997) finds λ by bootstrapping, repeatedly removing random data points from the fit and measuring the effect on the derived image. The L-curve method (Miller 1970; Lawson & Hanson 1974; Hansen 1992, 1994; Engl & Grever 1994) evaluates the sum of the squares of the residuals as a function of λ, which gives an L-shaped curve, hence the name of the method. The preferred value of λ is at the knee of the L, where the curvature is highest.

7.3. Linear (Tikhonov) Regularization

The simplest penalty function is quadratic in the image. The advantage of the quadratic penalty function is that its gradient with respect to the image is linear, as is the gradient of the χ^2. The optimization of a merit function consisting of the sum of a χ^2 and a quadratic penalty function is then a linear problem. The method is often called Tikhonov (1963) regularization, although it seems to have been independently suggested by a number of authors (see Press et al. 2002, who also present a succinct discussion of the method).

The penalty function for linear regularization is the sum of the squares of a linear mapping of the image:

$$B(I) = \sum_i \left(\sum_j F_{ij} I_j \right)^2, \tag{35}$$

which is designed to penalize highly variable images. It is often a finite difference approximating first-order or second-order derivatives. As with all regularization methods, the strength of the penalty function is controlled by the Lagrange multiplier λ (Equation 30), which is adjusted so that χ^2 is approximately equal to the number of data points.

The solution of the linear regularization problem simplifies significantly when the blur is a convolution and F is also chosen to be a convolution. By analogy with the Fourier method (Section 2.4), the Fourier transform of the gradient of the merit function is a linear equation in $\tilde{I}(\mathbf{k})$, whose solution is

$$\tilde{I}(\mathbf{k}) = \frac{\tilde{H}(\mathbf{k})^* \tilde{D}(\mathbf{k})}{|\tilde{H}(\mathbf{k})|^2 + \lambda |\tilde{F}(\mathbf{k})|^2}. \tag{36}$$

Note that the complex-conjugate, $\tilde{H}(\mathbf{k})^*$, which appears in the numerator of Equation 36, is the Fourier transform of the transpose of the point-response function H^T. In the absence of a regularization term $\lambda = 0$, and Equation 36 reduces to Equation 4, as derived with the Fourier method. The regularization term in the denominator of Equation 36 serves to suppress the high-$\mathbf{k}$ components, as $\tilde{F}(\mathbf{k})$ is designed to peak at high $\mathbf{k}$. One can think of Equation 36 as a generalization of the Wiener filter (Equation 14) to allow more elaborate filtering.

Image reconstruction using linear regularization has been applied often in the field of microscopy (e.g., van Kempen et al. 1997). Recent studies have enforced nonnegative images either by a change of variables (Carrington et al. 1995; Verveer, Gemkow & Jovin 1999) or by clipping negative values at each step of a conjugate-gradient iteration (Lagendijk & Biemond 1991, Vandervoort & Strasters 1995).

7.4. Total Variation

Regularization schemes whose penalty functions are smooth functions of the image tend to perform poorly when the underlying truth image contains sharp edges or steep gradients. A penalty function that overcomes this problem is the total variation (Rudin, Osher & Fatemi 1992; Vogel & Oman 1998):

$$B(I) = \sum_i |\nabla I|_i. \tag{37}$$

Equation 37 can be generalized by considering other functions of ∇I. Charbonnier et al. (1997) discuss the types of functions that would be useful and suggest a few possibilities.

In the form of Equation 37, the total variation has the property that it applies the same penalty to a step edge as it does to a smooth transition over the same range of image amplitudes. The penalty increases only when the image model develops oscillations. This is a serious limitation, which would cause us to discard this penalty function unless it is known ahead of time that the image contains many sharp edges. If this is not the case, especially if the signal-to-noise ratio is low, the user risks introducing significant artifacts.

7.5. Maximum Entropy

The maximum-entropy method is an attempt to provide an objective image preference in analogy with the principles that underlie statistical physics (Jaynes 1957a, 1957b). It assumes that the image is made up of a very large number of quanta, each with intensity q, and that there is an equal probability that any quantum lands in any image pixel, as if tossed at random by monkeys (Gull & Daniell 1978). The probability of obtaining a particular set $(n_1, n_2, \ldots, n_L)$ of pixel occupation numbers, with $n_j = I_j/q$, is then proportional to the degeneracy $N!/n_1!n_2!\ldots n_L!$. In the asymptotic limit of large occupation numbers, the factorials can be approximated by the Stirling formula (e.g., Press et al. 2002), and the logarithm of the prior

becomes:

$$S(I) = \ln [p(I)] \approx -\sum_i n_i \ln (n_i) = -\sum_i (I_i/q) \ln (I_i/q), \qquad (38)$$

where we have dropped an overall normalization constant of $p(I)$ (additive constant after taking the logarithm). The preferred image is then the one that maximizes the entropy. Equation 38 is analogous to the spatial distribution probability of particles of an ideal gas, whose logarithm is the Boltzmann entropy. An imaging entropy of the form $I \ln(I)$ was originally proposed by Frieden (1972).

There are three fundamental problems with this approach: (*a*) there are competing forms of the entropy, even for the same image, (*b*) the maximum-entropy image, for any entropy scheme, is not the preferred image, and (*c*) the entropy depends on the quantum q, which is set arbitrarily and is not related to any physical quanta making up the image. Because of these problems, the maximum-entropy method has steered away from its precepts of statistical physics. Let us deal with these issues one by one.

First, the functional form of the image entropy is not unique. One might view the electric field in Fourier space, with the spatial image intensity as its power spectrum, to be the fundamental carrier of image information. In that case, the entropy is (Ponsonby 1973, Ables 1974, Wernecke & D'Addario 1977):

$$S(I) = \sum_i \ln (n_i) = \sum_i \ln (I_i/q). \qquad (39)$$

The same expression is obtained from photon Bose-Einstein statistics (Narayan & Nityananda 1986). The use of entropy of the form $\ln(I)$ in imaging actually predates the use of Equation 38 (Burg 1967).

Second, the image with the highest entropy is a flat image with constant intensity, which is not the preferred image. The purpose of image reconstruction is to find the true underlying image that has been degraded by blurring and noise, but the target image is not flat. The flat image also has the unfortunate property of invariance under random scrambling of the pixels. Surely, any real image would be terribly degraded under such scrambling and the scrambled image should not be considered equally preferable to the real image. The spatial distribution of the residuals, once normalized by the standard deviations of the pixels, should be invariant under random scrambling of the pixels, but not the image. So, perhaps one should define the entropy based on the residuals and not on the image. We return to this point in Section 8.5.

Third, the quanta in the maximum-entropy method cannot be the photons, as might be supposed based on the analogy with statistical physics. Maximization of the a posteriori probability entails a balance between the log-likelihood function and the entropy, so the deviations of the optimal solution are within expected statistical fluctuations from both the maximum-likelihood solution and the maximum-entropy solution. But this is not possible, because the maximum-entropy solution is flat, whereas the true image is not, and the difference between the two is highly

significant statistically. This brings us back to the second problem: the underlying image is not the flat maximum-entropy image.

In practice, users of the maximum-entropy method simply multiply the entropy by an unknown factor λ, and adjust its value to obtain a reasonably good fit to the data. The maximum-entropy minimization function thus takes the form of Equation 33. For entropy of the form of Equation 38, this corresponds, approximately, to setting the quantization to a high level, far above that of individual photons. In the case of entropy of the form of Equation 39, a change of quantization actually does not help, because it only affects the entropy additively, so the multiplicative factor is totally arbitrary. In either case, multiplying the entropy by a factor λ corresponds to raising the prior probability to the power of λ, a procedure that is alien to the Bayesian approach.

The maximum-entropy method has been applied extensively, particularly in radio astronomy (see the review by Narayan & Nityananda 1986). But, as they emphasize, the success of the maximum-entropy method is not due to the Bayesian precepts that led to it, and from which the method has veered away. It really results from the characteristics of the entropy function used, particularly infinite slope at $I = 0$, which steers the solution away from zero and negative values, and a negative second derivative, which makes the negative of the entropy (negentropy) a suitable penalty function. The function $\sqrt{I}$, with no theoretical basis, would be just as good and represents an intermediate case between Equations 38 and 39.

The real problem of the maximum entropy method is one that it shares with all the methods of global image restriction, namely that the same restriction is applied everywhere in the image. This one-criterion-fits-all approach often leads to underfits in parts of the image and overfits in other parts. Instead, we should allow variable image restriction that adapts itself to image conditions. This is the latest development in image processing, to which we now turn.

8. SPATIALLY ADAPTIVE IMAGE RESTRICTION

Here we explore another class of image restrictions, which applies different image restrictions across the image. These techniques are more flexible, as they can adapt themselves to different image conditions, e.g., greater smoothness or variation in signal-to-noise ratio across the image. But they can also more easily lead to confusion between signal and noise, thereby resulting in stronger artifacts. The proof of the pudding is in the images produced. We present a comparison of the global and local methods in Section 8.7.

8.1. Spatially Variable Maximum Entropy

As we saw, the maximum-entropy functionals in Equations 38 and 39 are particularly ill-suited to adapt to image content because they are maximized by flat images. Recognizing this limitation, Skilling (1989) proposes a modified entropy

functional, whose maximum occurs at a preassigned reference image J. The probabilities of pixel occupancy are no longer equal, so Equation 38 is replaced by the Poisson log-likelihood function (Equation 20) to give

$$S(I) = \sum_j [I_j - J_j - I_j \ln(I_j/J_j)]. \quad (40)$$

If the reference image is only determined to within an unknown normalization constant, then the total fluxes of the image and the reference image are set equal to each other, in which case the first two terms on the right-hand side of Equation 40 cancel, and the equation simplifies to:

$$S(I) = -\sum_j I_j \ln(I_j/J_j), \quad (41)$$

a form known as the Kullback relative information, Kullback-Leibler divergence, or cross entropy (Kullback & Leibler 1951, Gray 1990).

A spatially variable entropy can increase the quality of a maximum-entropy reconstruction. Equation 40 has been used to introduce reference images of various types (Gull 1989; Charter 1990; Weir 1992; Bontekoe, Koper & Kester 1994; Bridle et al. 1998; Jones et al. 1998, 1999; Marshall et al. 2002; Strong 2003). A commercial version called MEMSYS is available from Maximum Entropy Data Consultants Ltd. and is compared with other image reconstructions in Section 8.7. Variations of Equation 41 have been applied in gravitational lensing (Seitz, Schneider & Bartelmann 1998) and medical imaging (Byrne 1993, 1998, and references therein). Additional applications have been used in conjunction with wavelets (Section 8.2).

8.2. Wavelets

We saw in Section 3.4 that wavelets can provide spatially adaptive denoising of data prior to noniterative deconvolution. The reader is referred to that section for a discussion of the motivation for wavelet filtering and its characteristics. Most of the astronomical applications, however, have actually used iterative methods, in which wavelet filtering is applied repeatedly during the iterations (see the review by Starck, Pantin & Murtagh 2002). The basic idea is to filter the residuals between iterations, setting the insignificant ones to zero, and leaving only significant structures. The decision as to which wavelets are significant and which are not can be made initially (Starck & Murtagh 1994; Starck, Murtagh & Gastaud 1998) or can be updated in each iteration (Murtagh, Starck & Bijaoui 1995; Starck, Murtagh & Bijaoui 1995). Wavelet-based denoising can be applied in combination with Clean or the van Cittert or Richardson-Lucy methods (Wakker & Schwarz 1988; Starck, Pantin & Murtagh 2002).

Wavelets can also be used in the maximum-entropy method, writing the entropy in terms of wavelet coefficients instead of pixel values (Pantin & Starck 1996; Starck, Murtagh & Gastaud 1998; Starck & Murtagh 1999; Starck et al. 2001; Figueiredo & Nowak 2003; Maisinger, Hobson & Lasenby 2004). This has been

applied to the analysis of the cosmic microwave background radiation (Hobson, Jones & Lasenby 1999; Sanz et al. 1999; Tenorio et al. 1999).

A disadvantage of images defined by wavelet basis functions is that the basis functions have both positive and negative image values. The negative values therefore must be filled in by other basis functions, if the image is to be nonnegative. This is less efficient, i.e., requires more basis functions, than representing a nonnegative image by nonnegative basis functions. We argue in Section 8.5 that minimum complexity, i.e., minimizing the number of basis functions used to characterize the image, is the key to spatially adaptive image restriction. Wavelet basis functions are at a disadvantage in this respect, because of the complicated way in which they enforce nonnegativity.

8.3. Markov Random Fields and Gibbs Priors

Recall that the maximum-entropy prior (Section 7.5) assumes that the pixels are statistically independent; the combined prior of all the pixels is the product of the priors of the individual pixels. Equivalently, the total entropy of all the pixels is the sum of the entropies of the individual pixels. This continues to hold even when a reference image is introduced (Section 8.1). Pixels have varying priors but continue to be statistically independent. An alternative is to introduce pixel correlations into the prior, i.e., the probability of an image value at a pixel would depend on the image values in neighboring pixels. These conditional probabilities are called Markov random fields.

The starting point of a Markov random field is a neighborhood system that identifies for each pixel j a neighborhood C_j, called its clique, such that the probability of obtaining I_j is simply a conditional probability on the image values in C_j. The prior can be written in the form of an exponential of a potential function V of the clique members (Besag 1974, 1986; Geman & Geman 1984):

$$p(I) \propto \exp\left[-\sum_j \sum_{k \in C_j} V_{C_j}(I_k)\right]. \tag{42}$$

This form is reminiscent of the Gibbs function of statistical physics (which describes the interactions between particles), so the prior in Equation 42 is called a Gibbs prior.

The point for image reconstruction is that the potential terms V in Equation 42 depend on the cliques, which introduce spatial correlations. The main application so far has been in medical imaging (Shepp & Vardi 1982, Hebert & Leahy 1989, Green 1990) in which the goal is to delineate more clearly body organs by locating different cliques inside and outside the organs. Normally the organs are delineated in advance, perhaps with the aid of other images (Gindi et al. 1991). Adaptive delineation is also possible (Figueiredo & Leitao 1994, Higdon et al. 1997). The method holds promise for pattern recognition in general (Lu & Jiang 2001).

8.4. Atomic Priors and Massive Inference

The entropy prior (Section 7.5), even in its spatially variable form (Section 8.1), suffers from two inherent theoretical difficulties. First, the prior of a sum of two images, e.g., the sum of two polarization states, is not the convolution of the priors of the two images, as probability theory requires. Second, the prior does not converge to a continuum limit as the image pixels become infinitesimally small. In order to overcome these difficulties, Sibisi & Skilling (1997) and Skilling (1998, 2003) propose to construct the image from "atoms" scattered randomly over the field of view, each carrying a flux, which itself is a Poisson random variable. The positions of the atoms are kept to machine precision, so pixelation is not an issue, and the Poisson distribution of the individual atomic fluxes guarantees that the prior of a sum of fluxes is correctly given by the convolution of the priors of the individual fluxes.

The difficulty of the scheme is to find the positions and fluxes of the atoms. This is done by a Markov chain Monte Carlo simulation. A few atoms are first injected randomly. Additional atoms are then sampled from the Markov random field, and the process is repeated until a smooth image is obtained. The method, called massive inference, is therefore spatially adaptive by construction and can be quite powerful. It is also very computationally intensive. The only published applications that we have been able to find are to 1D time series and spectra (Skilling 1998; Ochs et al. 1999; Ebbels, Lindon & Nicholson 2001).

8.5. Ockham's Razor and Minimum Complexity

Is there a general principle that can guide us in designing image restriction? So far we have considered several specific methods. Some are parametric fits that specify explicit functional forms. Others are nonparametric methods that restrict the image model in one way or another, either globally or with a degree of spatial adaptation. A common characteristic of all these methods is that they establish correlations between image values at different locations. The character of these correlations depends on the reconstruction method, but a common thread is that image restriction and image correlations go hand in hand. The stronger the image restriction, the stronger are the correlations, i.e., they extend over larger separations. A restatement of the image reconstruction problem might therefore be: "Find the most strongly correlated image that fits the data." The trick comes in designing image restriction to express the correct kind of correlations while remaining general enough to include all possible images that may be encountered.

Another way of considering the problem is in terms of the information content of the data. As we know, the data consist of reproducible information in the form of signal due to an underlying image and irreproducible information due to noise. We are only interested in the reproducible information, which is what we really mean by information content. The most conservative and reliable way to divide the information into its reproducible and irreproducible parts is to maximize the information associated with the noise and to minimize the signal information, i.e.,

to look for the least informative characterization of the image. It follows that, if information content can be measured by an entropy, then we should strive to maximize the entropy of the noise, i.e., the residuals, and not the entropy of the image, as is done by the maximum-entropy method (Section 7.5).

The idea of seeking the minimalist explanation of observed data goes back to the English theologian William of Ockham (ca. 1285–1349), who advocated parsimony of postulates, stating, "*Pluralitas non est ponenda sine necessitate*," which translates as, "Plurality should not be posited without necessity." This principle, known today as Ockham's razor, has become a cornerstone of the scientific method.

It is straightforward to apply Ockham's razor to parametric fits: One accepts a parameter only if it is statistically significant, thereby restricting the number of parameters to the minimum required by the data. That is common scientific practice, and is what the Clean method does in the area of image reconstruction (Section 5.3). It is more difficult to apply Ockham's razor to nonparametric methods, because it is not clear what "parsimony of postulates" means in that case. There have been attempts to define complexity by equating it with the algorithmic information content, the length of the program needed by a "universal computer" (Turing 1936, 1937) to print the reproducible information content and stop (Solomonoff 1964, Kolmogorov 1965, Chaitin 1966). Unfortunately, defined in such a general and abstract way, it is not possible to find the minimum complexity in any manageable time, because the set of possible models (images in our case) is combinatorially large. In computer science, the problem is said to be NP-hard (Cook 1971), i.e., it is intrinsically harder than those that can be solved by a nondeterministic universal computer in polynomial time in the size of the problem.

To avoid the abstract generality and the combinatorial explosion of possible image models, we need to restrict the set of image models from which we seek the minimally complex solution. Another way of saying this is that we need an image language to describe image information content. A parametric model is an extremely specific way to characterize image information content but it is also very restrictive. At the other extreme, a nonparametric representation of an image by means of independent values at each point of a large grid is too loose, and only serves to introduce artifacts. We need to design an intermediate language that can describe the image in terms of the shapes, sizes, and positions of image structures, so we can describe complex images more compactly with a minimal number of components. In analogy with our daily use of language, we need a rich vocabulary that embraces all the possible information that we may wish to impart, and then use the minimum number of words required in any instance.

8.6. Full Pixon

Smoothing reduces the number of image components. The easiest way to minimize image complexity is therefore to require the maximum, spatially adaptive image smoothness permitted by the data. This is the approach taken by the Pixon method

(Piña & Puetter 1993, Puetter & Yahil 1999). Given a trial nonnegative image ϕ, called a pseudoimage, consider the image obtained by smoothing it with a nonnegative, spatially variable kernel K:

$$I_j = \sum_k K_{jk}\phi_k = (K \otimes \phi)_j \,. \tag{43}$$

The goal of Pixon image reconstruction is to find the smoothest possible image by selecting the broadest possible nonnegative K for which a nonnegative ϕ can be found, such that the image given by Equation 43 fits the data adequately. It does so by optimizing K and ϕ in turn.

It is straightforward to find ϕ given K. Expressing the data model in terms of the pseudoimage, we have:

$$M = H \otimes I = H \otimes (K \otimes \phi) = (H \otimes K) \otimes \phi. \tag{44}$$

So, one simply replaces H by $H \otimes K$ and solves for ϕ using a nonnegative least-squares fit with a stopping criterion (Section 6.2). Then I is given by Equation 43. This is reminiscent of the method of sieves (Section 6.5). The difference is that the Pixon method allows K_{jk} to vary from one grid position k to the next.

To determine K, the Pixon method uses a finite set of kernels, called Pixon kernels, which are rich enough to allow all images of interest to be expressed in the form of Equation 43, but not too extensive to include unwanted images. The design of the Pixon kernels depends on the type of images at hand. For most applications, a set of circularly (spherically) symmetric kernels, whose widths form a geometric series, work very well. The exact functional form of the kernels is not too important. The important point is that the Pixon kernels should span the sizes and general shapes of the expected image features, so that a pseudoimage can be smoothed with the Pixon kernels and yield those features.

A set of trial images is constructed by convolving ϕ separately with each of the kernels in turn. The final image is then obtained by selecting for each grid point the trial image from which the image value is taken. The aim is to select at each grid point the trial image with the broadest kernel function while still fitting the data adequately. The image made of the indices of the trial images selected at each grid point is called the Pixon map. Details of the determination of the Pixon map may vary from application to application. In any event, the Pixon map should be smooth on the same scales used to smooth the pseudoimage, and this prevents discontinuities in the final image. A further refinement is to allow "intermediate kernels" by interpolating between the trial images. This smoothes the image further and/or allows the use of fewer kernel functions. The geometric spacing of the widths is designed for optimal characterization of multiscale image structures.

A Pixon image reconstruction thus proceeds alternately between determining ϕ and K. The starting point is a determination of ϕ with some initial K. For example, the initial K might be a delta function, in which case the first image is the nonnegative least-squares solution. Another possibility is to start the fit with

kernels that are deliberately too broad—resulting in a poor fit to the data—and to reduce the kernel widths gradually during the iterations until the data are fit satisfactorily, a process called annealing. For most images, Pixon reconstruction can proceed in a total of only three steps: (*a*) find a nonnegative least-squares image (delta-function *K*), (*b*) determine the Pixon map for the nonnegative least-squares image, and (*c*) update the pseudoimage using the Pixon map just determined. Annealing is used for images with a wide spectrum of features on all scales, so the large scales are fit first followed by smaller scales.

It is important to emphasize that, because the Pixon method deliberately seeks to find the smoothest image, it characterizes image features in the broadest possible way. This bias is deliberate and is intended to prevent narrow artifacts from masquerading as real sources. Sometimes, however, external information tells us that some sources are narrow. Then we must change the language used to describe the image, i.e., we must select different kernels, eliminating broad ones. In the limit of a field of point sources, the Pixon method becomes a Clean reconstruction (Section 5.3), using the signal-to-noise ratio to eliminate weak sources. If there is diffuse emission in addition to the point sources, we need to restore the broad kernels, but it may be possible to eliminate intermediate-size kernels if it is known that the diffuse emission is smooth enough.

Pixon image reconstruction has been applied in astronomy (e.g., Metcalf et al. 1996; Dixon et al. 1997; Figer et al. 1999; Gehrz et al. 2001; Young, Puetter & Yahil 2004), microscopy (Kirkmeyer et al. 2003, Shibata et al. 2004), medical imaging (Vija et al. 2005, Wesolowski et al. 2005), and defense and security.

8.7. Discussion

Before concluding, we undertake a more comprehensive discussion of the various merits and shortcomings of the principal image reconstruction concepts presented in this review. Until now, the bulk of our discussion has focused on theoretical accounts for how a certain approach might improve over the results obtained with another; here we present supporting examples to illustrate how the methods might compare in practice.

It is, however, very difficult for a single practitioner to make absolutely fair comparisons between the various methods. One simple reason is that different techniques may be appropriate for different data sets. More problematic is the fact that different researchers acquire different competencies with the various methods—especially as regards the more technically complex ones—so that a purported superiority may reflect only the skill level or biases of the user. Historically, the fairest comparisons of different image reconstruction techniques have issued from organized "shootouts," in which experts in the different methodologies process the same raw data and then mutually evaluate the results. Unfortunately, producing such events requires considerable effort and is undertaken only rarely. To our knowledge, Bontekoe et al. (1991) organized the only shootout among some of the major methods discussed in this review. Like most authors, we have

not made such an effort and refrain from attempting definitive comparisons. We do, however, take advantage of our expertise in the use of the Pixon method to underscore what we believe to be some of the most important emerging issues in high-performance image processing.

Our first comparison appears in Figure 3, in which we reconsider the image of New York City shown in Figure 2. Reproduced are the Wiener reconstruction with $\beta = 0.1$ and a nonnegative least-squares reconstruction carried to convergence, together with the truth image and the blurred and noisy data. Both reconstructions appear to have good residuals, but are chock-full of artifacts. (The artifact level is somewhat more severe for the nonnegative least-squares solution.) In fact, both reconstructions overfit the data, with χ^2/n values of 0.88 and 0.76, for the Wiener and nonnegative least-squares reconstructions, respectively. (The differences between these values and the expected value of unity are significant at the 40σ and 80σ levels, respectively.) Figure 3 reemphasizes the danger of overfitting the data. It is hard to see this in the residual plots in Figure 3, but spectral analysis of the residuals shows that the reconstructions primarily overfit the low-**k** data components, whereas the high-**k** components are treated as noise and are not reconstructed. The reconstructions amplify the artifacts because of overfitting more strongly with increasing **k** (Section 2.8). The result is that the images are dominated by the artifacts with the highest overfitted **k**, making the artifacts particularly noticeable.

Figure 4 presents a nonnegative least-squares reconstruction of the same New York City image shown in Figures 2 and 3, this time with early termination at $\chi^2 = n + \sqrt{2n}$. Also shown are the quick Pixon reconstruction already displayed in Figure 2 and a new full Pixon reconstruction. The nonnegative least-squares solution shows better residuals than the quick Pixon reconstruction but significantly stronger artifacts. The full Pixon reconstruction, by contrast, shows both better residuals and appears to be artifact free.

All three reconstructions have reasonable χ^2, but that is only a single global measure, which can hide variations across the image. The artifacts of the nonnegative least-squares fit show that it overfits the data in parts of the image and underfits them in other parts. The quick Pixon reconstruction does a better job avoiding artifacts but it does not fit the data well enough, and signal structures can be seen in the residuals. Only the full Pixon reconstruction manages to fit the data well enough to leave reasonable residuals while avoiding artifacts. (The residuals are not perfect, and the reconstruction might perhaps benefit by adding more kernel functions.)

An illustration of image reconstruction in the astronomical arena is presented in Figure 5. The raw data, in the form of 60 μm scans collected by the *Infrared Astronomical Satellite* (IRAS), were corrected for nonuniformity and coadded before being presented to leading experts in a number of reconstruction techniques (Bontekoe et al. 1991). The point-response function was also provided. The coadded data are shown in (*b*) with the point-response function on the same scale in an insert. The competing reconstructions are shown as contour diagrams on the right: (*c*) the high-resolution (HIRES) method of NASA's Infrared Processing and Analysis Center, based on the maximum-correlation method (Rice 1993), (*d*) the

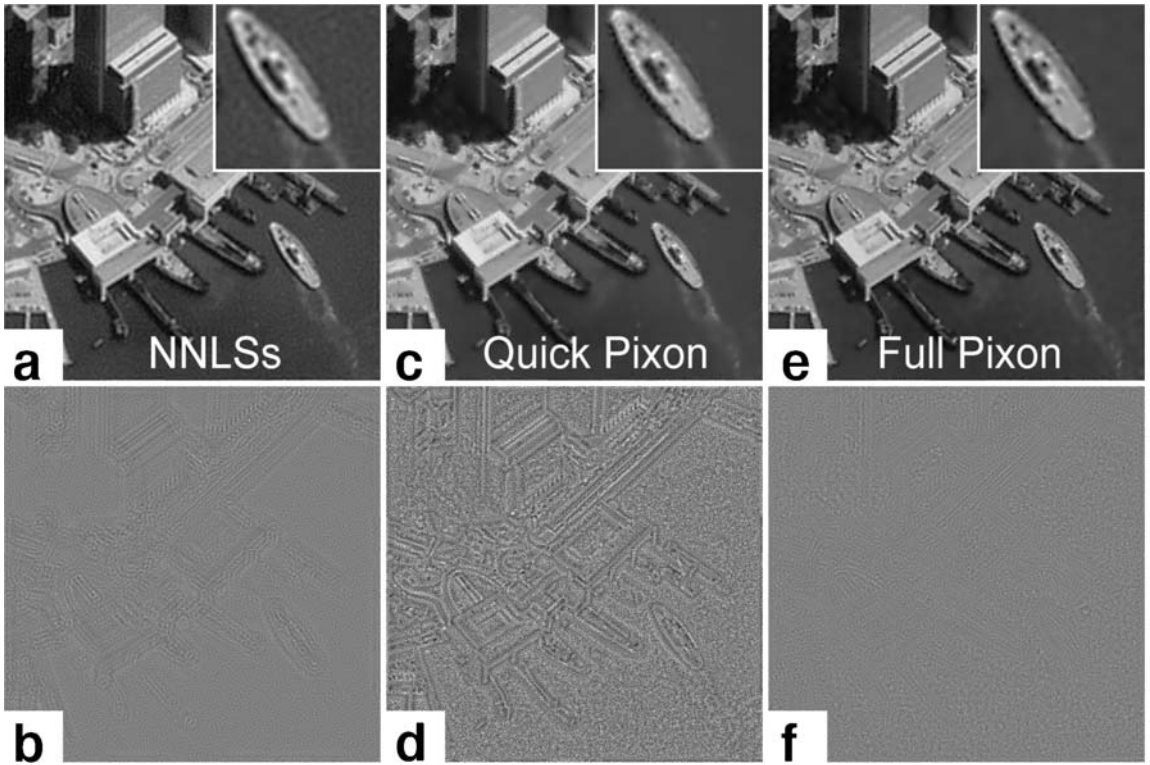

Figure 4 Variety of iterative image reconstructions: (*a*) stopped nonnegative least-squares fit with (*b*) residuals, (*c*) quick Pixon with (*d*) residuals, and (*e*) full Pixon with (*f*) residuals. The data and the "truth" image are shown in Figure 3.

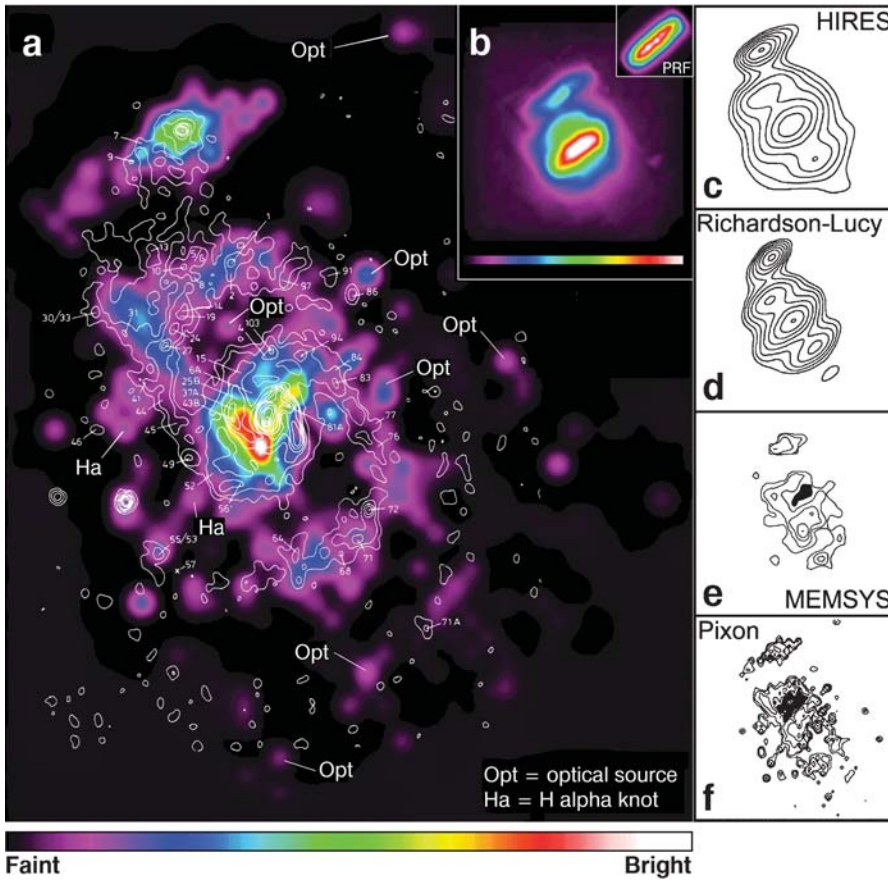

Figure 5 Variety of image reconstructions of 60 μm scans of the galaxy pair M51/NGC5195 taken by the *Infrared Astronomical Satellite* (Bontekoe et al. 1991): (*a*) false color image of the Pixon reconstruction in (*f*) overlaid with 5 GHz radio continuum contours (van der Hulst et al. 1988), (*b*) coadded input data with the point-response function on the same scale in an insert, (*c*) NASA high-resolution reconstruction, (*d*) Richardson-Lucy reconstruction, (*e*) maximum-entropy reconstruction, and (*f*) Pixon reconstruction. Objects identified in optical images are also marked in (*a*): (Opt) stars, (Ha) *Hα* emission knots. The black patches in (*e*) and (*f*) represent zero intensity. The scales of panels (*a*), (*b*) and (*c*)–(*f*) are unequal and in the ratios 1.0 : 0.18 : 0.28.

Richardson-Lucy method, (*e*) a commercial version of the maximum-entropy method (MEMSYS 3), and (*f*) the Pixon method. That the fine features appearing in the Pixon reconstruction actually exist is demonstrated in (*a*), where we plot contours of radio continuum intensity on top of the Pixon image. Also shown are features visible in optical images. Note that the scales of panels (*a*), (*b*) and (*c*)–(*f*) are unequal and in the ratios 1.0 : 0.18 : 0.28.

The Richardson-Lucy and HIRES reconstructions clearly fail to recover much more than the gross shape of the galaxy pair. In particular, they fail to recover the hole in the galactic emission that appears just north of the nucleus (solid black portions seen in the Pixon and maximum-entropy images). The maximum-entropy reconstruction begins to resolve structure in the galaxy's spiral arms, but the Pixon result is clearly superior. It actually recovers sources 200 times fainter than those visible in the maximum-entropy reconstruction. The linear spatial resolution is also better by a factor of $\sim$3. The stark difference between the Pixon and other reconstructions can be attributed to the maximal, spatially adaptive smoothing, which protects the reconstruction from getting lost in artifacts.

Figure 6 shows direct external validation of Pixon processing of 12 μm data taken by IRAS. Panel (*a*) shows a collage of scans kindly prepared by Dr. Romano at the Aerospace Corporation. For each point in each scan the collage includes the scan flux at the pixel corresponding to the center of the beam at the time the flux was measured. (The average flux is used when several scans overlap on the same pixel.) The main sources are all point-like stars (there is also some diffuse background emission), yet they are spread significantly by the point-response function, particularly in the cross-scan direction. Panel (*b*) shows a HIRES reconstruction of the image from the original scans (not the collage) performed at NASA's Infrared Processing and Analysis Center. Many stars are visible, but the point-response function has clearly not been optimally deconvolved, and blur persists in the cross-scan direction. Panel (*c*) shows the Pixon reconstruction performed on the collage in (*a*), which reveals many more sources than HIRES and little residual cross-scan blur. Finally, panel (*d*) shows an image of the same area of the sky, taken 12 years later by the *Midcourse Space Experiment* (MSX) satellite of the U.S. Air Force with a much improved imaging system, validating the Pixon reconstruction.

Finally, Figure 7 shows an example from nuclear medicine (Vija et al. 2005): a phantom with numerous dowels of varying diameters and heights containing radioactive material emitting gamma rays. The goal is to image the smallest possible dowels containing the least amount of radioactive material. The *top* panels show the raw counts, whereas the *bottom* panels show the results of the Pixon reconstructions. The acquisition times are varied to yield total counts ranging from 0.2 megacounts in the *left* panels to 0.8 megacounts in the *middle* panels to 6.4 megacounts in the *right* panels. This provides a range of signal-to-noise ratios, whereby the tradeoff between image fidelity and acquisition time or dose can be assessed. For these planar scintigraphic images, the blur is insignificant compared with the Poisson counting noise, i.e., the Pixon kernels used to smooth the image are everywhere wider than the point-response function, so only adaptive smoothing is performed and no deblurring is attempted. Parameters controlling Pixon processing are kept fixed for all acquisitions to strain the test of how adaptive and data driven the method is. A visual comparison of the Pixon reconstruction in panel (*b*) and the raw counts in panel (*c*) shows that the Pixon images improve image quality in a way that can be achieved by the raw images only by increasing the counts by an order of magnitude.

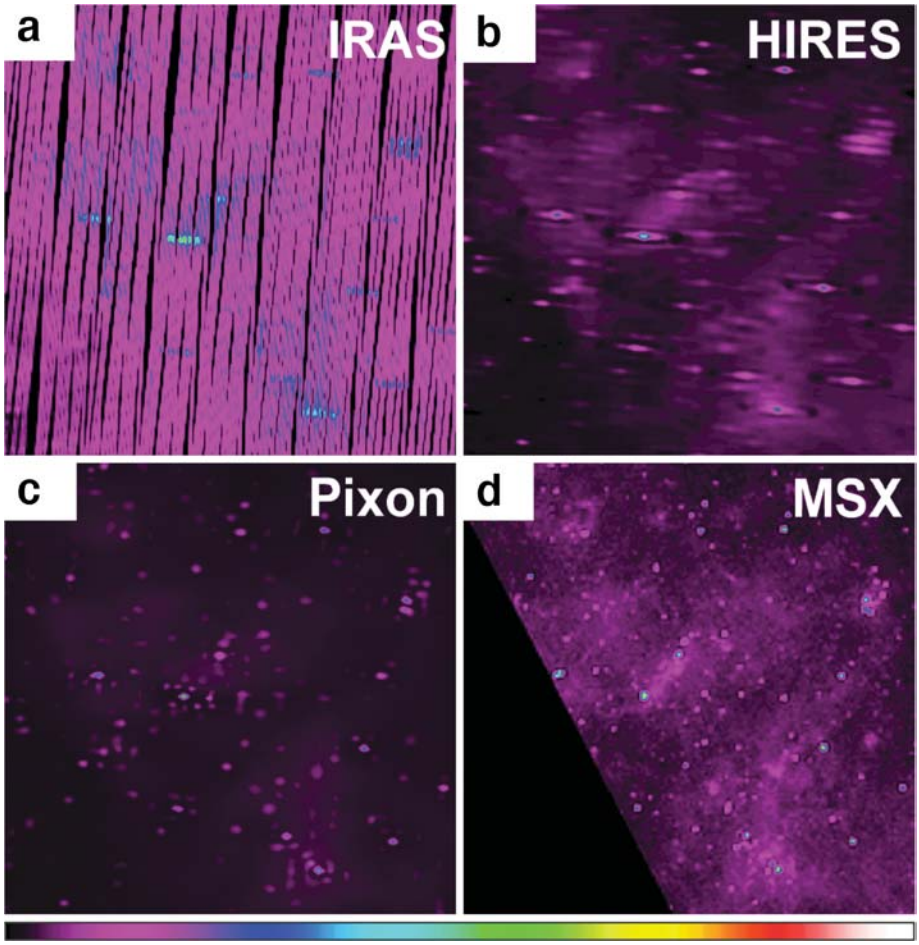

Figure 6 Externally validated comparison of Pixon and NASA reconstructions of 12 μm scans taken by the *Infrared Astronomical Satellite*: (*a*) collage of the scans, (*b*) high-resolution reconstruction by NASA's Infrared Processing and Analysis Center, (*c*) Pixon reconstruction, and (*d*) an image obtained 12 years later by the *Midcourse Space Experiment* satellite of the U.S. Air Force and processed by the Space Dynamics Laboratory (Logan, UT).

9. SUMMARY

The past few decades have seen the evolution of two unmistakable trends in high-performance image processing: (*a*) techniques for image restriction are essential for solving the inverse problem of image reconstruction, and (*b*) performance is significantly improved when the image-restriction strategy is allowed to adapt to varying conditions across the image.

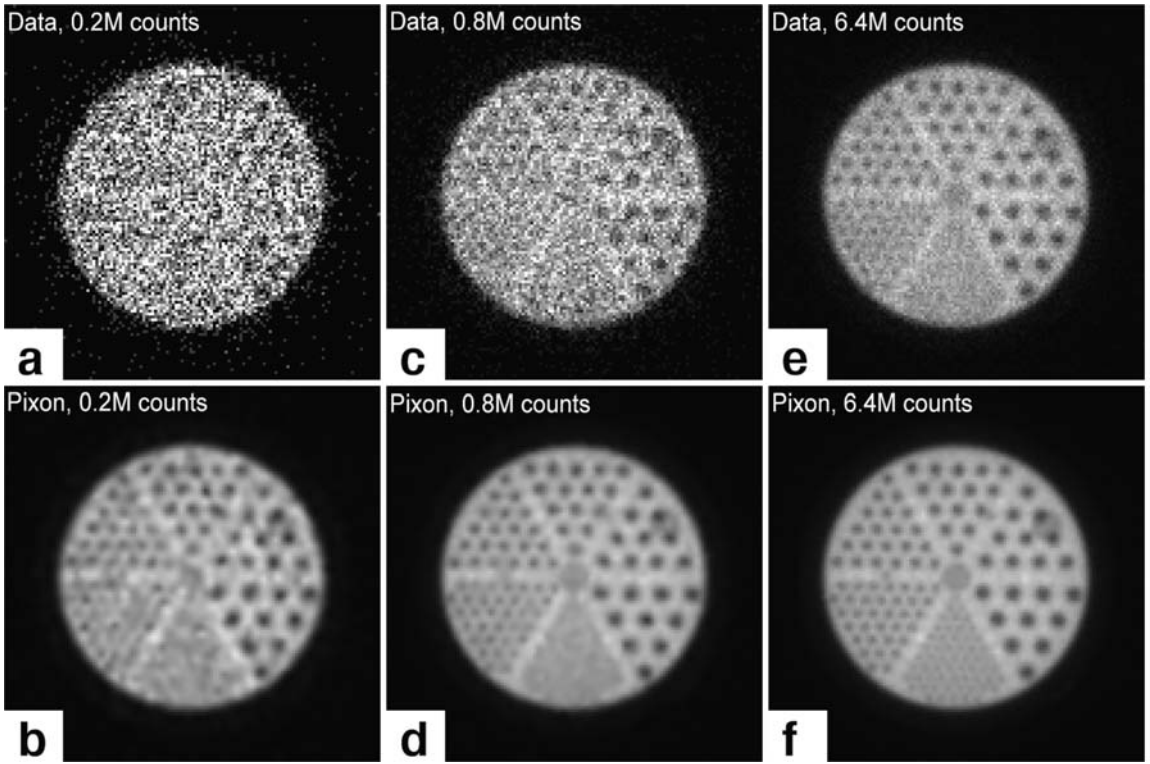

Figure 7 Order-of-magnitude noise suppression of planar scintigraphic γ-ray images of medical phantoms by Pixon reconstructions: (*a*) 0.2 megacounts data with (*b*) Pixon image, (*c*) 0.8 megacounts data with (*d*) Pixon image, and (*e*) 6.4 megacounts data with (*f*) Pixon image.

Several arguments can be given to justify the need for image restriction. First, an image model can often be found that overfits the data completely, leaving zero residuals. This image model is clearly wrong because the noise is not caused by the true underlying image but by irreproducible observational errors. There should be finite residuals that follow the parent statistical distribution of the noise.

Second, the maximum-likelihood fit—the method of choice for parametric fits in which the number of data points greatly exceeds the number of parameters—is powerless to prevent data overfitting in nonparametric fits, in which the number of parameters is comparable to the number of data points.

Third, the problem of image artifacts is exacerbated because a blurring point-response function has very small eigenvalues whose eigenfunctions are dominated by components with high wave vectors **k**. Upon inversion, these high-**k** components of the data are therefore greatly magnified. When they are due to noise, they result in large artifacts, sometimes to the point of swamping the true image.

The negative side of image restriction is that it may be either insufficiently restrictive, allowing image artifacts to get through, or too restrictive, resulting in a poor fit to the data. This limitation of image restriction has led to the development of increasingly sophisticated methods that try to find images with good fits to the data and as few artifacts as possible. The simplest noniterative methods discussed in Section 3 have given way to the more powerful iterative methods of Sections 6–8.

Even among the iterative methods we see different approaches and capabilities. The use of a global penalty function, or equivalently a preference function (Section 7), can significantly improve image reconstruction. The strength of the penalty term can be adjusted to give an acceptable value to a global goodness-of-fit statistic such as χ^2. But the effect of image restriction may be uneven because of variable conditions across the image. Parts of the image may be overfitted, while other parts are underfitted.

The limitation of global image restriction has naturally led to a desire to impose spatially adaptive image restriction. The danger here is that spatially adaptive image restriction may do a poorer job of separating signal from noise, so it must operate within strict guidelines designed to prevent arbitrary solutions. The guidelines are context-dependent. They boil down to our preconception of what the image should look like and they may legitimately vary from one application to the next.

The principle that should be common to all the applications is minimum complexity. Start with a rich language, whose vocabulary describes all the types, shapes, and sizes of components that you expect to encounter in the image. Then try to minimize the number of words that you actually use to describe the image. This is exactly what we do so effectively in our daily use of language. We have a large vocabulary at our disposal but we use only a small fraction of this vocabulary to convey any specific information.

There are objective measures of success. The ultimate ones are external validations by independent measurements. But we can also test our image reconstruction internally by analyzing the residuals. They should be consistent with a random

sample of the parent statistical distribution. We can and should test this in multiple ways, not just by means of a χ^2 test but also by searching for coherent structures in the residual map and by investigating the statistical distribution of the residuals. Simulations can also be run, in which the reconstructed image can be compared directly with the "truth" image.

If we succeed simultaneously to minimize the complexity of the image and to produce statistically acceptable residuals, the reconstructed image is the most reliable image possible, the best that one can deduce from the data at hand.

DISCLOSURE STATEMENT

R.C. Puetter and A. Yahil are founders and owners of Pixon LLC, and T.R. Gosnell was an employee of the company from October 2000 through December 2004.

ACKNOWLEDGMENT

The authors thank George Romano of the Aerospace Corporation for help in preparing the collage of scans from IRAS used in Figure 6.

The *Annual Review of Astronomy and Astrophysics* is online at http://astro.annualreviews.org

LITERATURE CITED

Ables JG. 1974. *Astron. Astrophys. Suppl.* 15: 383–93

Abrams MC, Davis SP, Rao MLP, Engleman R, Brault JW. 1994. *Astrophys. J. Suppl.* 93: 351–95

Agard DA. 1984. *Annu. Rev. Biophys. Bioeng.* 13:191–219

Almoznino E, Loinger F, Brosch N. 1993. *Mon. Not. Roy. Astron. Soc.* 265:641–48

Avni Y. 1976. *Astrophys. J.* 210:642–46

Bayes T. 1763. *Philos. Trans. R. Soc. London* 53:370–418

Beard SM, MacGillivray HT, Thanisch PF. 1990. *MNRAS* 247:311–21

Bender R. 1990. *Astron. Astrophys.* 229:441–51

Bertero M, Boccacci P. 1998. *Introduction to Inverse Problems in Imaging*. London: Inst. Phys. Publ.

Bertero M, Boccacci P. 2000. *Astron. Astrophys. Suppl.* 147:323–33

Bertero M, Boccacci P. 2003. *Micron* 34:265–73

Bertin E, Arnouts S. 1996. *Astron. Astrophys.* 117:393–404

Besag JE. 1974. *J. R. Stat. Soc. B* 36:192–236

Besag JE. 1986. *J. R. Stat. Soc. B* 48:259–302

Bhatnagar S, Cornwell TJ. 2004. *Astron. Astrophys.* 426:747–54

Biemond J, Lagendijk RL, Mersereau RM. 1990. *Proc. IEEE* 78:856–83

Bijaoui A. 1980. *Astron. Astrophys.* 84:81–84

Bijaoui A, Starck JL, Murtagh F. 1994. *Traitement du Signal* 11:229–43

Biraud Y. 1969. *Astron. Astrophys.* 1:124–27

Bontekoe TR, Kester DJM, Price SD, Dejonge ARW, Wesselius PR. 1991. *Astron. Astrophys.* 248:328–36

Bontekoe TR, Koper E, Kester DJM. 1994. *Astron. Astrophys.* 284:1037–53

Borman S, Stevenson RL. 1998. In *Proc. 1998 Midwest Symp. Circuits Syst.*, pp. 374–78. Notre Dame, IN: IEEE

Bridle SL, Hobson MP, Lasenby AN,

Saunders R. 1998. *Mon. Not. Roy. Astron. Soc* 229:895–903
Burg JP. 1967. *Annu. Meet. Int. Soc. Expl. Geophys.* Reprinted in 1978. *Modern Spectrum Analysis*, ed. DG Childers, pp. 34–41. New York: IEEE
Byrne C. 1993. *IEEE Trans. Image Process.* 2:96–103
Byrne C. 1998. *Inverse Problems* 14:1455–67
Calvetti D, Reichel L, Zhang Q. 1999. *Appl. Comput. Control, Signals Circuits* 1:313–67
Carrington WA, Lynch RM, Moore ED, Isenberg G, Fogarty KE, Fay FS. 1995. *Science* 268:1483–87
Chaitin GJ. 1966. *J. Assoc. Comput. Mach.* 13:547–69
Charbonnier P, BlancFeraud L, Aubert G, Barlaud M. 1997. *IEEE Trans. Image Process.* 6:298–311
Charter MK. 1990. In *Maximum Entropy and Bayesian Methods*, ed. PF Fougere, pp. 325–39. Dordrecht: Kluwer
Clark BG. 1980. *Astron. Astrophys.* 89:377–78
Conchello JA, McNally JG. 1996. *Proc. SPIE* 2655:199–208
Cook SA. 1971. *Proceedings of the 3rd Annual Symposium on the Theory of Computing*, pp. 151–58. New York: ACM
Cornwell TJ. 1983. *Astron. Astrophys.* 121: 281–85
Cosman PC, Oehler KL, Riskin EA, Gray RM. 1993. *Proc. IEEE* 81:1325–41
Daubechies I. 1988. *Commun. Pure Appl. Math.* 41:909–96
Dávila CA, Hunt BR. 2000. *Appl. Opt.* 39: 3473–85
Dempster AP, Laird NM, Rubin DB. 1977. *J. R. Stat. Soc. B* 39:1–38
Dixon DD, Tumer TO, Zych AD, Cheng LX, Johnson WN, et al. 1997. *Astrophys. J.* 484:891–99
Donoho DL. 1995a. *IEEE Trans. Inf. Theory* 41:613–27
Donoho DL. 1995b. *J. Appl. Comput. Harmon. Anal.* 2:101–26
Donoho DL, Johnstone IM. 1994. *C. R. Acad. Sci. I* 319:1317–22
Ebbels TMD, Lindon JC, Nicholson JK. 2001. *Applied Spectroscopy* 55:1214–24
Egmont-Petersen M, de Ridder D, Handels H. 2002. *Pattern Recognit.* 35:2279–301
Eicke B. 1992. *Num. Funct. Anal. Opt.* 13:413–29
Elad M, Feuer A. 1999. *IEEE Trans. Pattern Anal. Mach. Intell.* 21:817–34
Engl HW, Grever W. 1994. *Numer. Math.* 69:25–31
Figer DF, Morris M, Geballe TR, Rich RM, Serabyn E, et al. 1999. *Astrophys. J.* 525:759–71
Figueiredo MAT, Leitao JMN. 1994. *IEEE Trans. Image Process.* 3:789–801
Figueiredo MAT, Nowak RD. 2003. *IEEE Trans. Image. Process.* 12:906–16
Frieden BR. 1972. *J. Opt. Soc. Am.* 62:511–18
Fruchter AS, Hook RN. 2002. *Publ. Astron. Soc. Pac.* 114:144–52
Galatsanos NP, Katsaggelos AK. 1992. *IEEE Trans. Image Process.* 1:322–36
Gehrz RD, Smith N, Jones B, Puetter R, Yahil A. 2001. *Astrophys. J.* 559:395–401
Geman S, Geman D. 1984. *IEEE Trans. Pattern Anal. Mach. Intell.* 6:721–41
Gersho A, Gray RM. 1992. *Vector Quantization and Signal Compression*. Dordrecht: Kluwer
Ghez AM, Neugebauer G, Matthews K. 1993. *Astron. J.* 106:2005–23
Gindi G, Lee M, Rangarajan A, Zubal IG. 1991. In *Information Processing in Medical Imaging*, ed. ACF Colchester, DJ Hawkes, pp. 121–31. Berlin: Springer-Verlag
Golub GH, Heath M, Wahba G. 1979. *Technometrics* 21:215–23
Golub GH, von Matt U. 1997. *J. Comput. Graph. Stat.* 6:1–34
Gray RM. 1990. *Entropy and Information Theory*. Berlin: Springer-Verlag
Green PJ. 1990. *IEEE Trans. Med. Imaging* 9:84–93
Gull SF. 1989. In *Maximum Entropy and Bayesian Methods*, ed. J Skilling, pp. 53–71. Dordrecht: Kluwer
Gull SF, Daniell GJ. 1978. *Nature* 272:686–90
Hadamard J. 1902. *Bull. Princeton Univ.* 13:49–52 (In French)

Hadamard J. 1923. *Lectures on Cauchy's Problem in Linear Partial Differential Equations*. New Haven: Yale Press. Reprinted 1952. New York: Dover

Hansen PC. 1992. *SIAM Rev.* 34:561–80

Hansen PC. 1994. *Numer. Algorithms* 6:1–35

Hebert T, Leahy R. 1989. *IEEE Trans. Med. Imaging* 8:194–202

Higdon DM, Bowsher JE, Johnson VE, Turkington TG, Gilland DR, Jaszczak RJ. 1997. *IEEE Trans. Med. Imaging* 16:516–26

Hobson MP, Jones AW, Lasenby AN. 1999. *MNRAS* 309:125–40

Hoeting J, Madigan D, Raftery A, Volinsky C. 1999. *Stat. Sci.* 14:382–417

Högbom JA. 1974. *Astrophys. J. Suppl.* 15:417–26

Holschneider M, Kronland-Martinet R, Morlet J, Tchamitchian P. 1989. *Wavelets: Time-Frequency Methods and Phase Space*, ed. JM Combes, A Grossman, P Tchamitchian, pp. 286–97. Berlin: Springer-Verlag

Hunt BR. 1995. *Int. J. Imaging Syst. Technol.* 6:297–304

Hyvärinen A, Karhunen J, Oja E. 2001. *Independent Component Analysis*. New York: Wiley

Infante L. 1987. *Astron. Astrophys.* 183:177–184

Jaynes ET. 1957a. *Phys. Rev.* 106:620–30

Jaynes ET. 1957b. *Phys. Rev.* 108:171–90

Jones AW, Hancock S, Lasenby AN, Davies RD, Gutierrez CM, et al. 1998. *MNRAS* 294:582–594

Jones AW, Lasenby AN, Mukherjee P, Gutierrez CM, Davies RD, et al. 1999. *MNRAS* 310:105–09

Jones R, Wykes C. 1989. *Holographic and Speckle Interferometry: A Discussion of the Theory, Practice and Application of the Techniques, 2nd ed*. Cambridge, UK: Cambridge Univ

Joshi S, Miller MI. 1993. *J. Opt. Soc. Am. A* 10:1078–85

Kalifa J, Mallat S, Rouge B. 2003. *IEEE Trans. Image Process.* 12:446–57

Kirkmeyer BP, Puetter RC, Yahil A, Winey KI. 2003. *J. Polym. Sci.: Polym. Phys.* 41:319–26

Kolmogorov AN. 1965. *Problems Inf. Transm.* 1:4–7

Kullback S, Leibler RA. 1951. *Ann. Math. Stat.* 22:76–86

Lagendijk RL, Biemond J. 1991. *Iterative Identification and Restoration of Images*. Dordrecht: Kluwer

Lawson CL, Hanson RJ. 1974. *Solving Least Squares Problems*. Englewood Cliffs, NJ: Prentice-Hall

Liang L, Xu Y. 2003. *IEEE Signal Process. Lett.* 10:129–32

Lu Q, Jiang T. 2001. *Pattern Recognition* 34:2029–39

Lucy LB. 1974. *Astron. J.* 79:745–54

Maisinger K, Hobson MP, Lasenby AN. 2004. *MNRAS* 347:339–54

Marshall PJ, Hobson MP, Gull SF, Bridle SL. 2002. *MNRAS* 335:1037–48

Metcalf TR, Hudson HS, Kosugi T, Puetter RC. 1996. *Astrophys. J.* 466:585–94

Mighell KL. 1999. *Astrophys. J.* 518:380–93

Miller K. 1970. *SIAM J. Math. Anal.* 1:52–74

Molina R, Nunez J, Cortijo FJ, Mateos J. 2001. *IEEE Signal Process. Mag.* 18:11–29

Morozov VA. 1966. *Sov. Math.* 7:414–17

Murtagh F, Starck JL, Bijaoui A. 1995. *Astron. Astrophys. Suppl.* 112:179–89

Narayan R, Nityananda R. 1986. *Annu. Rev. Astron. Astrophys.* 24:127–70

Natterer F. 1999. *Acta Numer.* 8:107–41

Ochs MF, Stoyanova RS, Arias-Mendoza F, Brown TR. 1999. *J. Magnetic Resonance* 137:161–76

O'Sullivan JA, Blahut RE, Snyder DL. 1998. *IEEE Trans. Inf. Theory* 44:2094–123

Pantin E, Starck JL. 1996. *Astron. Astrophys. Suppl.* 112:179–89

Park SC, Park MK, Kang MG. 2003. *IEEE Signal Process. Mag.* 3:21–36

Pearson TJ, Readhead ACS. 1984. *Annu. Rev. Astron. Astrophys.* 22:97–130

Piña RK, Puetter RC. 1993. *Publ. Astron. Soc. Pac.* 105:630–37

Polak E, Ribiere G. 1969. *Revue Française Inf. Rech. Oper.* 16:35–43

Ponsonby JEB. 1973. *MNRAS* 163:369–80

Prasad CVV, Bernath PF. 1994. *Astrophys. J.* 426:812–21

Press WH, Teukolsky SA, Vetterling WY, Flannery BP. 2002. *Numerical Recipes in C++: The Art of Scientific Computing*. Cambridge, UK: Cambridge Univ.

Puetter RC, Yahil A. 1999. *Astron. Soc. Pac. Conf. Ser.* 172:307–16

Raykov T, Marcoulides GA. 2000. *A First Course in Structural Equation Modeling*. Mahwah, NJ: Lawrence Erlbaum

Rice W. 1993. *Astron. J.* 105:67–96

Richardson W. 1972. *J. Opt. Soc. Am.* 62:55–59

Rudin LI, Osher S, Fatemi E. 1992. *Physica D* 60:259–68

Sanz JL, Argüeso F, Cayón L, Martinez-González E, Barreiro RB, Troffolatti L. 1999. *MNRAS* 309:672–80

Sargent WLW, Schechter PL, Boksenberg A, Shortridge K. 1977. *Astrophys. J.* 212:326–334

Seitz S, Schneider P, Bartelmann M. 1998. *Astron. Astrophys.* 337:325–37

Serabyn E, Weisstein EW. 1995. *Astrophys. J.* 451:238–51

Shensa MJ. 1992. *IEEE Trans. Signal. Process.* 40:2464–82

Shepp LA, Vardi Y. 1982. *IEEE Trans. Med. Imaging* 1:113–22

Sheppard DG, Panchapakesan K, Bilgin A, Hunt BR, Marcellin MW. 2000. *IEEE Trans. Image Process.* 9:295–98

Shibata N, Pennycook SJ, Gosnell TR, Painter GS, Shelton WA, Becher PF. 2004. *Nature* 428:730–33

Sibisi S, Skilling J. 1997. *J. R. Stat. Soc.* 59:217–35

Simkin S. 1974. *Astron. Astrophys.* 31:129–36

Skilling J. 1989. In *Maximum Entropy and Bayesian Methods*, ed. J Skilling, pp. 45–52. Dordrecht: Kluwer

Skilling J. 1998. In *Maximum Entropy and Bayesian Methods*, ed. GJ Erickson, JT Rychert, pp. 1–14. Dordrecht: Kluwer

Skilling J. 2003. *BayeSys and MassInf*. Cambridge, UK: Maximum Entropy Data Consultants

Snyder DL, Miller MI. 1991. *Random Point Processes in Time and Space*. Berlin: Springer-Verlag

Solomonoff RJ. 1964. *Inf. Control* 7:1–22

Starck JL, Murtagh F. 1994. *Astron. Astrophys.* 288:342–48

Starck JF, Murtagh F, Bijaoui A. 1995. *Comput. Vis. Graph. Image Process.* 57:420–31

Starck JL, Murtagh F, Bijaoui A. 1998. *Image Processing and Data Analysis. The Multiscale Approach*. Cambridge, UK: Cambridge Univ.

Starck JL, Murtagh F, Gastaud R. 1998. *IEEE Trans. Circuits Syst. II – Analog Digit. Signal Process.* 45:1118–24

Starck JL. Murtagh F. 1999. *Signal Processing* 76:147–65

Starck JL. Murthagh F. Querre P. Bonnarel F. 2001. *Astron. Astrophys.* 368:730–46

Starck JL, Pantin E, Murtagh F. 2002. *Publ. Astron. Soc. Pac.* 114:1051–69

Steer DG, Dewdney PE, Ito MR. 1984. *Astron. Astrophys.* 137:159–65

Strong AW. 2003. *Astron. Astrophys*. 411: L127–29

Stuart A, Ord K, Arnold S. 1998. *Kendall's Advanced Theory of Statistics*. London: Arnold

Tenorio L, Jaffe AH, Hanany S, Lineweaver CH. 1999. *MNRAS* 310:823–34

Thompson AR, Moran JM, Swenson GW. 2001. *Interferometry and Synthesis in Radio Astronomy*. New York: Wiley

Tikhonov AN. 1963. *Soviet Math.* 4:1035–38

Turing AM. 1936. *Proc. London Maths. Soc. Ser. 2* 42:230–65

Turing AM. 1937. *Proc. London Maths. Soc. Ser. 2* 43:544–46

van Cittert PH. 1931. *Z. Phys.* 69:298–308 (In German)

van der Hulst JM, Kennicutt RC, Crane PC, Rots AH. 1988. *Astron. Astrophys.* 195:38–52

Vandervoort HTM, Strasters KC. 1995. *J. Microsc.* 178:165–81

van Kempen GMP, van Vliet LJ, Verveer PJ, van der Voort HTM. 1997. *J. Microsc.* 185:354–65

Verveer PJ, Gemkow MJ, Jovin TM. 1999. *J. Microsc.* 193:50–61

Vija AH, Gosnell TR, Yahil A, Hawman EG, Engdhal JC. 2005. *Med. Imaging, Proc. SPIE.* 5747:634–45

Vogel CR, Oman ME. 1998. *IEEE Trans. Image Process.* 7:813–24

Wahba G. 1977. *Siam J. Numer. Anal.* 14:651–67

Wakker BP, Schwarz UJ. 1988. *Astron. Astrophys.* 200:312–22

Wallace W, Schaefer LH, Swedlow JR. 2001. *BioTechniques* 31:1076–97

Weir N. 1992. *Astron. Soc. Pac. Conf. Ser.* 25:186–90

Wernecke SJ, D'Addario LR. 1977. *IEEE Trans. Comput.* 26:351–64

Wesolowski CA, Yahil A, Puetter RC, Babyn PS, Gilday DL, Khan MZ. 2005. *Comput. Med. Imaging Graph.* 29:65–81

Young EF, Puetter R, Yahil A. 2004. *Geophys. Res. Lett.* 31:L17–21