DIGITAL IMAGE RECONSTRUCTION: DEBLURRING AND DENOISING

R. C. Puetter [1], T. R. Gosnell [2], and Amos Yahil [3]
[1]*Center for Astrophysics and Space Sciences, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0424, and Pixon LLC, P.O. Box 312, East Setauket, NY 11733; email: Rick.Puetter@pixon.com*
[2]*Los Alamos National Laboratory, Los Alamos, NM 87545, and Pixon LLC, P.O. Box 312, East Setauket, NY 11733; email: Tim.Gosnell@pixon.com*
[3]*Department of Physics and Astronomy, Stony Brook University, Stony Brook, NY 11794-3800, and Pixon LLC, P.O. Box 312, East Setauket, NY 11733; email: 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 difficulty is overcome by restricting the permitted images. This review summarizes image reconstruction methods in current use. Progressively more sophisticated image restrictions have been developed, including (a) filtering of 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.

CONTENTS

1 INTRODUCTION

Digital image processing of the type discussed in this review has 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 output data of real-world instruments, so that information present but hidden in the data is revealed with less blur and noise.

Our choice of nomenclature is deliberate. Throughout this review, the term [**AU: House style strongly discourages the use of italics or quotation marks for emphasis. Most such deleted.** Deleted the rest]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 the term 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 a difficult problem because fluctuations in the image may be strongly blurred, in turn yielding only minor variations in the measured data. Hence, small variations in the data that result from measurement noise are perfectly compatible with large variations in the underlying image. Image reconstruction can thus lead to significant noise amplification, particularly at high spatial frequencies, and the ambiguity of potentially acceptable image reconstructions is thereby dramatically increased. The main challenge of image reconstruction is to overcome this equivocalness.

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 the baby out 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, therefore, is 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 eliminating wrong images.

There are great arguments on how this 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

[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 the term 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.

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. Second, an image reconstruction can be externally validated by repeat measurements, preferably by independent investigators. Simulations are also useful in this regard, as the truth image used to create the simulated data is known and can be compared with the reconstructed image.

Image processing, in general, has many other legitimate goals beyond accurate reconstruction. The dynamic range of the human visual system is limited and we might, for example, wish to adjust the image contrast to simultaneously discern features in shadow and in glare. We might also wish to doctor an image for artistic or cosmetic reasons, or we might be more interested in segmenting an image to facilitate interpretation than in viewing the actual picture. In all these cases, we sacrifice fidelity for other purposes. These adjustments, which often go under the name of image enhancement, are outside the scope of this review. In our view, they should in any case be applied after image reconstruction. The first task is to get the most accurate image. Then do with it whatever you like.

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 in Section 2 with the identification of image reconstruction as a noisy inverse problem. 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, which 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 analyzed by a variety of noniterative methods 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 and maximum likelihood.

Fitting methods fall into two broad categories, parametric and nonparametric. Section 5 is devoted to parametric methods, which are suitable for problems where 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.

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. The pragmatists simply add a regularization term to the unconstrained merit function to penalize unwanted images. The Bayesian approach is to assign an a priori probability function to the image, known as a prior, and to maximize the product of the prior and the likelihood function, known as the a posteriori probability. There is no operational difference between these two approaches. 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, and the full Pixon method. These adaptive methods are more flexible, because they allow image restriction to change from one position in the image to another. But they must also be careful not to allow spatial adaptation to be too loose or they will simply foster data overfits. 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. Among these are: (a) superresolution, a general term used for both subpixel resolution (Borman & Stevenson 1998, Elad & Feuer 1999, Fruchter & Hook 2002, Park, Park & Kang 2003) and subdiffraction resolution (Hunt 1995, Bertero & Boccacci 2003); (*b*) tomography (Natterer 1999); (*c*) vector quantization, a technique widely used in image compression and classification (Gersho & Gray 1992, Cosman et al. 1993, Hunt 1995, Sheppard et al. 2000); (*d*) the method of projection onto convex sets (Biemond, Lagendijk & Mersereau 1990); (*e*) 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 (*f*) 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 do not discuss two additional topics that can be of great practical importance. First, we confine ourselves to physical blur (instrumental and/or atmospheric) that spreads the image over more than one pixel. Loss of subpixel 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 in the last paragraph. 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 (see Section 6.2). This technique does not work if the physical blur is less than one pixel wide.

A second area we omit is the initial data reduction that is often necessary 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 THE NOISY INVERSE PROBLEM

2.1 Statement of the Problem

Digital image reconstruction is an example of a noisy inverse problem (e.g., Press et al. 2002). Given discrete, noisy data D_i—typically measured in pixels of one or more 2D detector arrays—we seek to estimate the underlying, continuous image $I(\mathbf{y})$ that gives rise to the observed data. For modern linear detectors, the relation between the image and the data is given by a linear integral equation (Fredholm integral equation of the first kind) with added noise:

$$D_i = M_i + N_i = \int H(\mathbf{x}_i, \mathbf{y}) I(\mathbf{y})\, d\mathbf{y} + N_i = (H \otimes I)_i + N_i \quad . \tag{1}$$

Here N_i is the measurement noise of the datum D_i, and M_i is the expected datum, which is a linear integral over the image $I(\mathbf{y})$. The kernel of the integral, $H(\mathbf{x}_i, \mathbf{y})$, is the instrumental response at the position $\mathbf{x}_i$ in which the datum D_i is measured, a known function of its arguments determined by the physical properties of the imaging instrument and possibly the atmosphere. The integral operation is denoted symbolically by $\otimes$. For a two-dimensional image, the domain of integration is the 2D $\mathbf{y}$-space upon which the image is defined. In general, however, the imaging problem may be defined on a space with an arbitrary number of dimensions, and the dimensionalities of $\mathbf{x}$ and $\mathbf{y}$ need not even be the same (e.g., in tomography).

An important additional interpretation of the function $H(\mathbf{x}, \mathbf{y})$ is immediately available upon substitution of a unit point-source image, $I(\mathbf{y}) = \delta(\mathbf{y} - \mathbf{y}')$, into Equation 1:

$$M_i = H(\mathbf{x}_i, \mathbf{y}') \quad . \tag{2}$$

Thus $H(\mathbf{x}_i, \mathbf{y})$ is equivalent to the expected data when a single point source at position $\mathbf{y}$ comprises the image; it generally expresses the details of the blurring relative to the true image. For this reason, $H(\mathbf{x}_i, \mathbf{y})$ is formally called the point-response function.[2] That the point-response function varies independently with respect to both $\mathbf{x}_i$ and $\mathbf{y}$ merely expresses the possibility that features in the data derived from a point-source image may depend upon where in the field of view the point source resides. Optical systems that suffer strong geometric aberrations would behave in this way.

A special case of Equation 1—one commonly encountered in image reconstruction—is convolution. In this case, the point-response function is only a function of the displacement $\mathbf{x}_i - \mathbf{y}$, independent of location within the field of view:

$$D_i = M_i + N_i = \int H\left(\mathbf{x}_i - \mathbf{y}\right) I\left(\mathbf{y}\right) d\mathbf{y} + N_i = \left(H * I\right)_i + N_i. \tag{3}$$

Deconvolution is the term used for the inverse problem of solving Equation 3, which distinguishes it from the more general image reconstruction problem of solving Equation 1. When necessary, we use the symbol $*$ to specify a convolution operation to clarify the distinction from the more general integral operation $\otimes$. Most of the image reconstruction methods described in this review, however, are general and are not restricted to deconvolutions.

The quality of the solution to a noisy inverse problem is mainly limited by the measurement errors. These 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, (e.g., Press et al. 2002).

To be explicit, consider a trial solution to Equation 1, $\hat{I}(\mathbf{y})$, and compute the residuals

$$R_i = D_i - M_i = D_i - \int H\left(\mathbf{x}_i, \mathbf{y}\right) \hat{I}\left(\mathbf{y}\right) d\mathbf{y}. \tag{4}$$

The image model can be considered an acceptable solution of the inverse problem if the residuals are statistically consistent with being a 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

[2] This terminology makes a distinction with the point-spread function, which characterizes only the point-to-point blurring before data sampling occurs on the detector. In fact, image reconstruction only requires knowledge of the point-response function; there is never any need to determine the point-spread function.

of the irreproducible noise.[3] 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, e.g., if its mean is not zero, or it is 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, parameter error estimation, and analysis of the statistical distribution of the residuals and their spatial correlations.

Unfortunately, the integral Equations 1 and 3 of image reconstruction are ill-posed problems. 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 less concerned with existence and uniqueness. Physical problems have solutions, and the solutions can be expressed uniquely with proper discretization (see Section 2.2) or by using functional forms. The main worry is the stability of the solutions, lest measurement errors in the input data be amplified to unacceptable artifacts in the solutions. Such noise amplification indeed plagues image reconstruction for any point-response function that spreads over even a few data pixels; it is the main challenge of image reconstruction. We return to this point in Section 3 but first we consider the discretization of the image.

2.2 Image Discretization

Technically, a continuous image model cannot be obtained from discrete data. This leaves two options. The image can either be characterized by continuous source functions with unknown parameters, in which case the solution is a parametric fit to the data (see Section 5) or it can be represented nonparametrically on a discrete grid (see Sections 6–8). In the latter case, the integral in Equation 1 is converted to a sum, yielding a set of linear equations:

$$D_i = M_i + N_i = \sum_j H_{ij} I_j + N_i \quad , \tag{5}$$

in matrix notation $D = M + N = HI + N$. In these expressions, D, M, and N are the data, data-model, and noise vectors, respectively, each containing n 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 are in reality often multidimensional arrays. 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.

In the case of a convolution, the discretization is greatly simplified by using Equation 3 to give:

$$D_i = M_i + N_i = \sum_j H_{i-j} I_j + N_i \quad . \tag{6}$$

This discretization is only possible if the data and image share the same n-point grid. The point-response function in Equation 6 takes a different form than in

[3] One might quibble that an image of a one-time event is not reproducible. What we mean by reproducibility in this case is that the event can, at least in principle, be imaged independently by a number of observers, and the signal is reproducible in the sense of being independently confirmed.

Equation 5. Like the data and the image, it 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 $n = n_x \times n_y$.) A subtle issue is how to deal with the edges of the image and data, where Equation 6 breaks down. The two common ways to handle that problem are to assume periodic boundary conditions for the image, or to extrapolate it beyond the range of the data and treat the extra data points as missing data. (Missing data are not confined to the edges; there may also be bad pixels or masked areas in the middle of the array.)

Note that the number of image values need not be equal to the number of data points, so the point-response function is not necessarily a square matrix. If the image changes slowly over the field of view, or the signal-to-noise ratio is low, we may choose a coarser discretization of the image than the data. Conversely, for data with high signal-to-noise ratio and a point-response function that spreads over several pixels, it is possible to determine the image at the subpixel level. (This does not violate the sampling theorem, because the image is required to be nonnegative; see Section 6.2 for details.) The data may also consist of several dithered frames. It is then again possible, with adequate signal-to-noise ratio, to determine the image on a finer grid than that of an individual frame, although the total number of data points in all the frames may be equal to or exceed the number of image points. Finally, note that it is not necessary to abandon the convenient discrete form of Equation 6 for convolutions when the number of data points is not equal to the number of image points. Instead, one breaks up the discrete operator into two. First one performs a discrete convolution, using Equation 6, and then one resamples the result using an expression equivalent to Equation 5 with the convolved image as input. The computation is greatly speeded up in this way, because convolutions can be performed using fast Fourier transforms (e.g., Press et al. 2002), and resampling is a simple local operation, which can be computed very efficiently.

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).[4] 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. How large a condition number can we tolerate? A realistic point-response function can blur the image over a few pixels, and its smallest eigenvalues decay fast as a function of the full width at half maximum of the point-response function. There is no escape from a large condition number. Equations 5 or 6 can therefore not be solved in their present, unrestricted forms. Either the equations need to be modified, or the solutions must be projected away from the subspace spanned by the eigenfunctions with small eigenvalues.

Image reconstruction is further compromised if the point-response function is not determined with sufficient accuracy. In particular, if the point-response function is determined from point sources whose positions are not known in advance, it may be necessary to take several dithered frames to determine their positions to subpixel accuracy. Otherwise, the measured point-response function can in-

[4] If the number of data and image points is not equal, 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.

troduce systematic errors in reconstructions that use it. Measured from a single source, it is only appropriate for sources that are similarly placed within their pixels as the source from which the point-response function was measured. If several sources are used to measure the point-response function, with random positions within their respective pixels, then the measured point-response function is blurred relative to the true point-response function because of the spread of subpixel positions. These problems arise for data with good signal-to-noise ratio. The higher the signal-to-noise ratio the greater the care with which the point-response function needs to be determined.

Finally, note that the full point-response function matrix 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 has to 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 5. 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 6, 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 or small-kernel deconvolutions (see Section 3).

3 NONITERATIVE METHODS

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, some noniterative methods do not control noise amplification very well, as we show below.

3.1 Fourier Deconvolution

Fourier deconvolution is one of the oldest and numerically fastest methods of image deconvolution. The technique uses the discrete form of the Fourier convolution theorem (e.g., Press et al. 2002) to express Equation 6 in Fourier space as:

$$\tilde{D}(\mathbf{k}) = \tilde{H}(\mathbf{k})\,\tilde{I}(\mathbf{k}) + \tilde{N}(\mathbf{k}) \quad , \tag{7}$$

where $\mathbf{k}$ is the spatial frequency (wave number) and the tilde indicates the Fourier transform. If the noise can be neglected, then the image I may be found by solving for $\tilde{I}$:

$$\tilde{I}(\mathbf{k}) = \frac{\tilde{D}(\mathbf{k})}{\tilde{H}(\mathbf{k})} \tag{8}$$

and computing the inverse Fourier transform of $\tilde{I}$. The technique is still used today in speckle image reconstruction (e.g., Jones 1983, Ghez, Neugebauer & Matthews 1993) and Fourier-transform spectroscopy (e.g., Abrams et al. 1994, Prasad & Bernath 1994, Serabyn & Weisstein 1995).

Unfortunately, the Fourier deconvolution technique breaks down when the noise is nonnegligible. Noise often has significant contribution from high spatial frequencies, e.g., white noise has equal contributions from all frequencies. But $\tilde{H}\left(\mathbf{k}\right)$, which appears in the denominator of Equation 8, falls off exponentially at high $\mathbf{k}$. The result is that high-frequency noise in the data is significantly amplified by the deconvolution to create image artifacts. In fact, the wider the point-response function the faster $\tilde{H}\left(\mathbf{k}\right)$ 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

Equation 8 is the Fourier representation of the deconvolution equation

$$I = H^{-1} * D \quad , \tag{9}$$

where H^{-1} is the inverse of the point-response function H. Instead of performing the deconvolution in Fourier space, it is possible to find an approximate inverse kernel $G \approx H^{-1}$, which can be designed to limit noise amplification. (The use of small inverse kernels is actually more general and not limited to pure deconvolutions.)

If G extends over only a few pixels, it is also possible to compute the convolution efficiently by direct summationor recursive direct summations, instead of using Fourier methods. This is particularly useful when processing raster video data in real time, because one can process the data in pipeline fashion as it comes in without waiting for the full frame to be loaded, as Fourier methods require. In hardware terms, this allows efficient use of a field-programmable gate array (FPGA) or an application-specific integrated circuit (ASIC), which are much more efficient than a digital signal processor (DSP) or a microprocessing unit (MPU).

The inverse kernel G can be set in several ways. Our own preference is to strike a balance between optimal deconvolution, noise amplification, and ringing caused by the finite extent of G. One way to do that is to minimize the sum:

$$S = \sum_i \left[\delta_i - \left(G * H\right)_i\right]^2 + \lambda \sum_i G_i^2 + \mu \sum_k \left[\left(G * H\right)_k + f \mathrm{Max}\left(G * H\right)\right]^4 \quad . \tag{10}$$

The first term in Equation 10 forces the deconvolved point-response function $G * H$ toward the target resolution function δ, which may be a delta function. The second term controls noise amplification with an adjustable parameter λ. The third term restricts negative overshoots of $G * H$, where μ is again an adjustable parameter, and the parameter f sets the negative cutoff point, below which ringing is penalized. The sum in the third term extends only over the points for which the term inside the square brackets is negative; the fourth power is used, instead of the second power used in the other two terms, to provide greater sensitivity to negative overshoots. In addition, G is normalized to unit sum, and this normalization is enforced at every iteration of the nonlinear minimization. In practical applications we normally set $\mu = 10^6$ and $f = 0.02$, and leave λ as a user-adjustable parameter to control the tradeoff between deblurring and noise amplification. For reasonable tradeoffs, the resolution improvement is typically about a factor of two.

3.3 Wiener Filter

In Section 3.1, we showed that if noise in the data is spatially uncorrelated—the most common case—then the artifacts in a resulting image generated by Fourier deconvolution are concentrated at high spatial frequencies. In other words, a white noise spectrum in the data translates to a blue artifact spectrum in the image. Wiener deconvolution is a modification to direct Fourier deconvolution that seeks to reduce the amount of noise in the reconstructed image by suppressing the high-frequency components of the data (low-pass filter of the data).

The idea of Wiener deconvolution is to apply a linear filter to the data that optimally controls the magnitude of the high-frequency content. Thus, in the frequency domain, Equation 8 is modified to yield a filtered Fourier transform of the image,

$$\tilde{I}\left(\mathbf{k}\right) = \Phi\left(\mathbf{k}\right) \frac{\tilde{D}\left(\mathbf{k}\right)}{\tilde{H}\left(\mathbf{k}\right)} \quad . \tag{11}$$

The filter function $\Phi(\mathbf{k})$ is chosen to minimize the difference between the true image and the deconvolved image in the least-squares sense (e.g., Press et al. 2002), yielding:

$$\Phi\left(\mathbf{k}\right) = \frac{\langle|\tilde{D}_0\left(\mathbf{k}\right)|^2\rangle}{\langle|\tilde{D}_0\left(\mathbf{k}\right)|^2\rangle + \langle|\tilde{N}\left(\mathbf{k}\right)|^2\rangle} \quad , \tag{12}$$

where $\langle|\tilde{N}\left(\mathbf{k}\right)|^2\rangle$ and $\langle|\tilde{D}_0\left(\mathbf{k}\right)|^2\rangle$ are the expected power spectra (also known as spectral densities) of the noise and the data model, respectively.

The noise is often spectrally white and may therefore be estimated by simple inspection of the Fourier-transform of the noisy data at high spatial frequencies, where the data model is likely to be small. (In practice, it is necessary to average over many frequencies, because the statistical fluctuation of any individual Fourier component is large.) The greatest difficulty comes in estimating the power spectrum of the data model, although the problem is made easier when the noise is white and a good estimate of the noise has been made.

Note that the Wiener filter does not attempt to recover the noise-free power spectrum of the data. In fact, the filtered data suppresses it:

$$\langle|\Phi\left(\mathbf{k}\right)\tilde{D}\left(\mathbf{k}\right)|^2\rangle = \frac{\langle|\tilde{D}_0\left(\mathbf{k}\right)|^2\rangle^2}{\langle|\tilde{D}_0\left(\mathbf{k}\right)|^2\rangle + \langle|\tilde{N}\left(\mathbf{k}\right)|^2\rangle} \quad . \tag{13}$$

This suggests that the Wiener formula might be modified to allow a tradeoff parameter to balance noise and signal suppression. We often introduce a free parameter β into Equation 12 to adjust the aggressiveness of the filter:

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

3.4 Wavelets

We saw in Sections 3.1 and 3.3 that the Fourier transform is a very convenient way to perform a deconvolution, because convolutions are simple products in Fourier space. The reason for this simplicity is that the sinusoidal Fourier spectral functions are eigenfunctions of the point-response function, so the point-response function operating on a Fourier spectral function simply returns that function multiplied by its eigenvalue. Similarly, when applying the inverse point-response function, the Fourier spectral function is divided by the eigenvalue. When the eigenvalue is small, however, the result of applying the inverse point-response function is large amplification, including amplification of the noise. To reduce noise amplification, Wiener filtering seeks to suppress the high-frequency components prior to deconvolution (see Section 3.3).

The disadvantage of the Fourier spectral functions is that they span the whole image and cannot be localized. One might wish to suppress high frequencies 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 convolution 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 spatial frequencies, because we know that we need to suppress the high-frequencies. 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 ones most used are those proposed by Daubechies (1988). In addition to striking a balance between spatial and frequency 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 (a) 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 and its inverse can each be computed hierarchically in $O\left(n \log_2 n\right)$ operations, just like the fast Fourier transform (e.g., Press et al. 2002).

Wavelet filtering is similar to Fourier filtering: 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 $m\sigma$. Alternatively, soft thresholding reduces the absolute values of the wavelet coefficients (Donoho & Johnstone 1994, Donoho 1995a). A further refinement is to threshold high frequencies more strongly (Donoho 1995b) or to modify the high-frequency 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 12 or 14.

Once the data have been filtered, deconvolution can proceed by the Fourier method or by small-kernel deconvolution. (Note that deconvolution cannot be performed in wavelet space, because the wavelets are not eigenfunctions of the point-response function.)

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 Section 8. 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. Special-purpose, pipeline hardware composed of current commercial-off-the-shelf components can achieve processing speeds of up to 300 megapixels per second.

3.6 Discussion

The performance of Wiener deconvolution can be assessed from the reconstructed images shown in Figure 1. For this example a 128 × 128 synthetic truth image, shown in the *lower left corner* 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 of 50. The resulting input data are shown in the *upper left corner* 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 14). 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. To avoid this effect, more aggressive filtering with $\beta = 10$ is chosen for 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 still show some signal.

The computation of the Wiener filter from Equation 14 requires an estimate of power spectra of the data model and the noise. Determining the white noise is straightforward. For the data model we use the Fourier transform of the data model obtained by convolving the truth image with the point-response function. The truth image is, of course, not known in a real reconstruction. Our purpose in using truth information is to present the absolute best that a Wiener reconstruction can achieve and to point out the remaining problems of resolution and artifacts, whatever tradeoff is chosen between them.

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 Gaussian noise so the peak signal-to-noise ratio 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. The truth image and the data are not shown here for lack of space, but are shown in Figure 3, Section 8. 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 the thresholding level, 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. Artifacts are minimal, the residuals are tolerable, and the 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 begs the question whether better techniques are available that simultaneously yield high resolution, minimal image artifacts, and residuals indistinguishable from 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 STATISTICAL METHODS

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.

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 data 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— which determine whether the residuals obtained are statistically consistent with being a random sample of 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. In addition, we 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. 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 being a random sample of that distribution. The merit function is usually the log-likelihood function described in the next subsection, to which are added a host of image restrictions (see Sections 6–8). Goodness of fit is diagnosed primarily by the χ^2 statistic, but also by considering the statistical distribution and spatial correlations of the residuals. Parameter error estimation is usually based on the behavior of χ^2 near its minimum value. See Press et al. (2002) or standard statistics references for the latter two topics.

4.2 Maximum Likelihood

The likelihood function of the data is defined as the joint conditional probability $p(D|I)$ of observing the data set D if the underlying image is indeed the image model I. In practice, it is more convenient to work with the log-likelihood function, a logarithmic quantity derived from the likelihood function:

$$\Lambda = -2 \ln \left[p \left(D|I \right) \right] = -2 \sum_i \ln \left[p \left(D_i|I \right) \right] \quad , \tag{15}$$

where the second equality in Equation 15 applies to statistically independent data, for which the joint probability of the entire data set is just the product of the probabilities of the individual data points. The factor of two is added for convenience to facilitate parameter error estimation and to equate the log-likelihood function with χ^2 for Gaussian noise.

Clearly, for the true underlying image, $p(D|I)$ is really the parent statistical distribution of the noise. The goal of data fitting is to find the best estimate $\hat{I}$ of I such that $p(D|\hat{I})$ is statistically 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 15. This method is known in statistics to provide the best estimates for a broad range of parametric fits 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, in and of itself, is not a good way to estimate the image; it was not designed for such problems in the first place. Nevertheless, it continues to be used in image reconstruction, albeit with additional image restrictions. The bulk of this review is devoted to nonparametric methods (see Sections 6-8).

5 PARAMETRIC METHODS

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 method is a least-squares fit minimizing the χ^2, the sum of the residuals weighted by their inverse variances:

$$\chi^2 = \sum_i \frac{R_i^2}{\sigma_i^2} = \sum_i \frac{(D_i - M_i)^2}{\sigma_i^2} \quad . \tag{16}$$

For a Gaussian parent statistical distribution, the log-likelihood function, after dropping constants, is actually the χ^2, so the χ^2 fit is also a maximum-likelihood solution. For a Poisson distribution, Mighell (1999) proposes to replace the logarithmic log-likelihood function with a useful χ^2-like merit function that provides unbiased estimates of the data model, even at low counts:

$$\chi_\gamma^2 = \sum_i \frac{[D_i + \min(D_i, 1) - M_i]^2}{D_i + 1} \quad . \tag{17}$$

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).

5.2 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.

This is what the Clean method does. Based on parametric techniques, it is an iterative image-reconstruction method that was originally developed for radio-synthesis imaging (Högbom 1974). 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 single bright point source as an initial guess for the image and perform a parametric fit. Second, add a new point source with appropriate amplitude at the location of the largest residual. Third, recalculate estimates of all point-source parameters (positions and amplitudes) introduced so far. Fourth, return to the second step if 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 plane (see Thompson, Moran & Swenson 2001), but the method is cumbersome for extended sources of unknown structure, which must be constructed a pixel at a time (Cornwell 1983). This has led to methods that employ phase closure (Pearson & Readhead 1984) or use multiple scales (Wakker & Schwarz 1988, Bhatnagar & Cornwell 2004, Cornwell & Holdaway, submitted to Astron. Astrophys.).

6 NONPARAMETRIC METHODS

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 even 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 understood to 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 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 potential 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 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 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 one for which the residuals are identically zero, as in Fourier deconvolution (see 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 are using 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? Its expectation value 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, which only ends up interpreting statistical noise as real, reproducible signal. If the image model were the true image, with no adjustment, 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 a notch earlier, when the 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 only 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. Fortunately, it is easy to identify uneven fits by displaying the residuals and/or their statistical distribution function.

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 it 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.

An additional benefit of the nonnegative restriction is potential subpixel resolution (Biraud 1969). The sampling theorem (e.g., Press et al. 2002) is not violated, because it applies only to unrestricted functions while we force the image to be nonnegative. With a point-response function extending over several pixels and a nonnegative image, the data do contain information on subpixel structure, which the reconstruction can extract. The possible degree of 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 as soon as they occur during the 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 does not hurt convergence and may actually speed it up. The latter is an example of projection onto convex sets (Biemond et al. 1990, Press et al. 2002).

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. The van Cittert (1931) method is one of the earliest and simplest iterative methods for 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

$$I^{(k+1)} = I^{(k)} + \alpha \left(D - H \otimes I^{(k)} \right) = \alpha D + Q \otimes I^{(k)} \quad , \tag{18}$$

where $Q = \mathbf{1} - \alpha H$; $\mathbf{1}$ is the identity kernel. Successive substitutions into Equation 18 yield

$$I^{(k)} = \alpha \sum_{j=0}^{k-1} Q^j \otimes D = H^{-1} \otimes \left(\mathbf{1} - Q^k \right) \otimes D \underset{k \to \infty}{\longrightarrow} H^{-1} \otimes D \quad , \tag{19}$$

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 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 et al.

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 et al. 2001) . Other implementations use wavelet-based filtering at each iteration, removing statistically insignificant features from the iterant (see Section 8.2).

6.4 Landweber

Another iterative scheme (Landweber 1951) is:

$$I^{(k+1)} = I^{(k)} + \alpha H^T \otimes \frac{R}{\sigma^2} \quad , \tag{20}$$

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 et al. 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).

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 log-likelihood function is minimized iteratively using multiplicative corrections:

$$I^{(k+1)} = \left[H^T \otimes \left(\frac{D}{M^{(k)}} \right) \right] I^{(k)} \quad . \tag{21}$$

The square brackets on the right-hand side of Equation 21 enclose the factor by which the previous iterant $I^{(k)}$ is multiplied (not convolved) to give the new iterant $I^{(k+1)}$. 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 of the point-response function.

Lucy (1974) shows that the algorithm is flux conserving, maintains image nonnegativity, and decreases the log-likelihood function in each iteration, at least if one takes only part of the step indicated by Equation 21. 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 broadened by convolving it with an extra sieve function. This 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 21 according to the general expectation-maximization procedure of Dempster et al. (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) of the log-likelihood function with respect to the image $g^{(0)}$ 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

$$h^{(k+1)} = g^{(k+1)} + \gamma_k h^{(k)} \quad . \tag{22}$$

The coefficient γ_k is chosen to optimize convergence. We generally prefer the one devised by Polak and Ribiere (see Press et al. 2002):

$$\gamma_k = \frac{\sum_j \left(g_j^{(k+1)} - g_j^{(k)} \right) g_j^{(k+1)}}{\sum_j \left(g_j^{(k)} \right)^2} \quad , \tag{23}$$

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 low spatial frequencies into the solution, and higher frequencies are added mainly in later iterations (Hansen 1994). Stopping the iterations on time therefore also provides a smoother solution, helping to reduce noise amplification at high spatial frequencies.

We have found that the most effective way to impose nonnegative solutions is by modifying the conjugate-gradient method as follows. For each conjugate-gradient direction find the minimum without regard to the sign of the image. Then simply set the negative image values to zero and compute the new conjugate-gradient direction from the new point as though no truncation took place. (We also set the new negradient components to zero if they point from zero image values to negative values.) 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 process 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. See Press et al. (2002) for details. (For the linear case they actually present the biconjugate-gradient method; the conjugate-gradient method is a special case that 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 8.6 below, 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:

$$\Lambda' = \Lambda + \lambda B\left(I\right) \quad . \tag{24}$$

Here $B\left(I\right)$ is a penalty function that increases with the degree of undesirability of the solution, and λ is the 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 each image model an a priori probability $p\left(I\right)$, also called a prior, and to maximize the product $p\left(D|I\right)p\left(I\right)$ 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 $p\left(I|D\right)$, known as the image a posteriori probability. Bayes' (1763) theorem is used to relate these quantities:

$$p\left(I|D\right) = \frac{p\left(D|I\right)p\left(I\right)}{p\left(D\right)} \propto p\left(D|I\right)p\left(I\right) \quad . \tag{25}$$

The data are fixed for any image reconstruction, so $p\left(D\right)$ is a constant and maximizing $p\left(I|D\right)$ amounts to maximizing the product $p\left(D|I\right)p\left(I\right)$. The image reconstruction is called a Bayesian method, and the solution is called the maximum a posteriori (MAP) image. Expressed logarithmically, we obtain an expression similar to Equation 24:

$$\Lambda' = \Lambda - 2\ln\left[p\left(I\right)\right] \quad . \tag{26}$$

One might imagine that the Bayesian approach is more restrictive, as 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 (see Section 7.5), an adjustable parameter is again introduced, changing Equation 26 to a form equivalent to Equation 24:

$$\Lambda' = \Lambda - \lambda S\left(I\right) \quad . \tag{27}$$

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 the authors recommend using the average of the a posteriori image instead of the maximum (e.g., Hoeting et al. 1999):

$$\langle I\rangle = \frac{\int p\left(I|D\right) I dI}{\int p\left(I|D\right) dI} . \tag{28}$$

In practice, however, the evaluation of the average image from Equation 28 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 Parameter 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 seemingly indicates that there is no longer any need for a stopping criterion, but that is illusory. Although it is nice to have a converging fit, it also must produce residuals that are statistically consistent with the parent statistical distribution, e.g., 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 24 or 27. 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 can be considered a Lagrange multiplier adjusted to enforce the data constraint. (It makes no 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, as the starting point was to maximize the a posteriori probability, but a goodness of fit works just as well, e.g., the χ^2_γ function of Mighell (1999) for Poisson noise (see Equation 17).

The use of χ^2 as a goodness-of-fit stopping criterion assumes advance knowledge of the standard deviations of the noise σ. When the noise level is not known in advance, it is possible to estimate it directly from the data. If the data are 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 can be systematically overestimated, and that they 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 & Wahba1979; 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;) simply 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 curve has its highest curvature.

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:

$$B(I) = \sum_i \left(\sum_j F_{ij} I_j \right)^2 \quad , \tag{29}$$

which is designed to penalize high spatial frequencies. 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 24), which is adjusted so that the χ^2 is approximately equal to the number of data points.

The solution of the linear regularization problem simplifies significantly when the point-response function is a convolution operator and F is also chosen to be a convolution. By analogy with the Fourier method (see Section 3.1), the Fourier transform of the gradient of the merit function is a linear equation in $\tilde{I}(\mathbf{k})$, whose solution is

$$\tilde{I}(\mathbf{k}) = \frac{\tilde{H}(\mathbf{k})^* \tilde{D}(\mathbf{k})}{|\tilde{H}(\mathbf{k})|^2 + \lambda |\tilde{F}(\mathbf{k})|^2} . \tag{30}$$

Note that the complex-conjugate, $\tilde{H}(\mathbf{k})^*$, appears in the numerator of Equation 30, which 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 30 reduces to Equation 8, as derived with the Fourier method. The regularization term in the denominator of Equation 30 serves to suppress the high spatial frequencies, as $\tilde{F}(\mathbf{k})$ is designed to peak at high frequencies.

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 of 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):

$$B(I) = \sum_i |\nabla I|_i \quad . \tag{31}$$

Equation 31 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 31, the total variation has the property of applying 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 preponderant sharp edges[**AU: This sentence sounds like it's missing something..? ".the data resultS IN an image"—Please clarify.** See modification]. If this is not the case, especially if the signal-to-noise ratio is low, the user risks introducing sizable 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 many 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 logarithm of the prior becomes:

$$S(I) = \ln[p(I)] = -\sum_i n_i \ln(n_i) = -\sum_i (I_i/q)\ln(I_i/q) \quad , \tag{32}$$

where we have dropped an overall normalization constant of $p(I)$ (additive constant after taking the logarithm). Equation 32 is analogous to the spatial distribution probability of particles of an ideal gas, whose logarithm is the Boltzmann entropy. The preferred image is then the one that maximizes the 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 anyway, 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):

$$S\left(I\right) = \sum_i \ln\left(n_i\right) = \sum_i \ln\left(I_i/q\right) \quad . \tag{33}$$

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 32 (Burg 1967).

Second, a gray maximum-entropy image cannot be 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 gray. If it were, we would not be interested in it in the first place. The gray 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 the image. We return to this point in Section 8.4.

Third, it is not clear how the image should be quantized. The problem is avoided in statistical physics, because the quanta are physical particles, but for the maximum-entropy method, the quanta should not be the photons. If they were, the fitted image would be identical to the gray maximum-entropy image to within Poisson counting statistics, which it clearly should not be. We are back to the second problem. The underlying image is not the gray 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 27. For entropy of the form of Equation 32, 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 33, 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.

In the end, the success of the maximum-entropy method has not been due to the Bayesian precepts that led to it, and from which the method has veered away, but to the specific functional form used. As Narayan & Nityananda (1986) emphasize, the exact form of the entropy is less important than its general functional characteristics, 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 32 and 33.

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-size-fits-all approach often leads to underfits in parts of the image and overfits in other parts. Instead, image restriction should allow variable image restriction that adapts itself to image conditions. This is the latest development in image processing, to which we turn next in Section 8.

8 SPATIALLY ADAPTIVE IMAGE RESTRICTION

Here we explore another class of image restrictions, which apply 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 a change in signal-to-noise ratio. But they also may not be too loose, or the extra flexibility may turn out to be no more than a way to permit noise amplification. The proof of the pudding is in the images produced. We present a comparison of the global and local methods in Section 8.6.

8.1 Spatially Variable Maximum Entropy

As we saw, the maximum-entropy functionals in Equations 32 and 33 are particularly ill-suited to adapt to image content, because they are invariant under random scrambling of the pixels. Recognizing this limitation, Skilling (1989) proposes inclusion of a reference image to obtain a modified entropy functional,

$$S\left(I\right) = \sum_j \left[I_j - J_j - I_j \ln\left(I_j / J_j\right)\right] \quad , \tag{34}$$

where J is the reference image.

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 Equation 34 simplifies to:

$$S\left(I\right) = -\sum_j I_j \ln\left(I_j / J_j\right) \quad , \tag{35}$$

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 34 has been used to introduce image correlations of various types (Gull 1989; Charter 1990; Weir 1992; Bontekoe, Koper & Kester 1994). Variants of the Kullback-Leibler cross entropy (Equation 35) have been applied to medical imaging (Byrne 1993, 1998, and references therein).

8.2 Wavelets

We saw in Section 3.4 that wavelets can provide spatially adaptive denoising of the data prior to noniterative deconvolution. Wavelet filtering can be extended to iterative methods by denoising repeatedly during the iterations. 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 has been applied in combination with the van Cittert, Richardson-Lucy, Clean, and maximum-entropy methods (see the review by Starck et al. 2002).

8.3 Markov Random Fields and Gibbs Priors

Markov random field theory provides a convenient and consistent way to model spatially dependent entities such as image pixels. This is achieved by characterizing mutual influences among the pixels using conditional Markov-random-field distributions. The starting point is a neighborhood system that identifies for each pixel j a neighborhood C_j, called a clique, such that the probability of obtaining I_j is simply a conditional probability on the image values in C_j. This conditional probability can be expressed in the form of a Gibbs function familiar from statistical physics (Besag 1974, 1986; Geman & Geman 1984):

$$p\left(I\right) \propto \exp\left[-\sum_j \sum_{k \in C_j} V_{C_j}\left(I_k\right)\right] \quad . \tag{36}$$

In Equation 36 the potential terms V depend on the cliques. This allows the introduction of a spatially dependent prior. The main application 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 images obtained using other techniques (Gindi et al. 1991). Adaptive delineation is also possible (Figueiredo & Leitao 1994, Higdon et al. 1997).

8.4 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 image reconstruction method used, 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. 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. We alluded to this idea in our discussion of the maximum-entropy method in Section 7.5, where we commented that the goal should be to maximize the entropy of the residuals, not of the image.

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 (see Section 5.2). Applying Ockham's razor to nonparametric methods is more difficult 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 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.

To avoid the abstract generality, we must be more specific about what we mean by information content, i.e., we need precise language to describe it. A parametric model is one very specific way to characterize the information content of an image, 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. Image values of neighboring grid points are often strongly correlated and assigning them independent information content is wrong and only serves to introduce artifacts. We need to design a language that can describe the image in terms of a hierarchy of structures, so we can describe complex images more compactly. In analogy with our daily use of language, we need a rich vocabulary to describe all the types, shapes, and sizes of components that we expect to encounter in the image, and then we need to minimize the number of words that we actually use to describe the image.

8.5 Full Pixon

One way to minimize image complexity is to identify simplicity with smoothness. This is the language used by the Pixon method to define image information content (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:

$$I_j = \sum_k K_{jk} \Phi_k = (K \otimes \Phi)_j \quad . \tag{37}$$

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 37 fits the data adequately.

How does one find the best combination of K and Φ? The Pixon method iteratively optimizes K and Φ in turn. Given K, it is straightforward to find Φ. Expressing the data model in terms of the pseudoimage, we have:

$$M = H \otimes I = H \otimes (K \otimes \Phi) = (H \otimes K) \otimes \Phi \quad . \tag{38}$$

So, one simply replaces H by $H \otimes K$ and solves for Φ using a nonnegative least squares fit with a stopping criterion (see Section 6.2). Then I is given by Equation 37. This is reminiscent of the method of sieves mentioned in Section 6.5, the difference being that the Pixon method allows K_{jk} to vary from one grid position k to the next.

But how should we choose K and what do we mean by making it as broad as possible? After all, we could choose a delta-function pseudoimage and K equal to any I we wish, in which case there would be no image restriction. That is clearly not the intent. What the Pixon method does is to restrict itself to a family of kernels, called Pixon kernels, which are rich enough to allow us to express all images of interest in the form of Equation 37, yet not too broad to include unwanted images such as in the aforementioned example.

The design of the Pixon kernels depends on the type of images at hand. For most applications, circularly (spherically) symmetric kernels of variable width work very well, e.g.,

$$K_{jk} \propto \begin{cases} 1 - (\mathbf{y}_j - \mathbf{y}_k)^2 / a_k^2 & : \quad |\mathbf{y}_j - \mathbf{y}_k| < a_k \\ 0 & : \quad \text{otherwise} \end{cases} \quad , \tag{39}$$

where a_k is the width of the kernel at grid position $\mathbf{y}_k$. The goal of the Pixon image reconstruction is then to find the largest a_k for each k for which there is an adequate fit to the data with nonnegative Φ. The exact functional form of Equation 39 is not too important; we could just as well use a Gaussian function instead of the inverted paraboloid. On the other hand, if the images are known to contain elongated structures, then the Pixon kernels might fruitfully be enlarged to include elliptical kernels. 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.

The set of a_k for all k specifies K and is called the Pixon map. The Pixon map is optimized by finding the largest possible width a_k for each k. This is done by seeking the largest a_k by which I, obtained in a previous iteration, can be further smoothed around $\mathbf{y}_k$ while still adequately fitting the data. One consideration in deciding whether the fit continues to be adequate is the effect that the image change ΔI_k has on χ^2. The idea is that the combined operation of the Pixon kernel and the point-response function acting on ΔI_k, $H \otimes K \otimes \Delta I_k$, affects a limited number of data pixels for each input ΔI_k, which we term the data footprint of image point k; $\Delta\chi^2$ need only be measured in that footprint. Another consideration is the significance of the new I_k, which can be measured by the signal-to-noise ratio in the footprint of k. Details of the ways in which $\Delta\chi^2$ and the signal-to-noise ratio are measured can vary. The trick is to avoid evaluating each a_k separately, as that is computationally prohibitive, but when the a_k are determined simultaneously, great care must be exercised to avoid cross-talk between the width tests of the different a_k. In any event, the user is provided with tuning parameters that determine how hard smoothing is driven.

In practice, it is useful to restrict the kernels to a finite set whose widths form a geometric series. A set of trial images is constructed by convolving Φ separately with each of these kernels to form a set of trial images. The final image is then obtained by selecting for each grid point the trial image from which the image value is taken. A further refinement is to allow intermediate values of a_k and to compute the image value by interpolation between the trial images. The geometric spacing of the widths is designed for optimal characterization of multiscale image structures.

Pixon image reconstructions thus proceed 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, and only then are the smaller scales fit.

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 narrower artifacts from masquerading as real sources. Sometimes, however, external information exists such that the sources are narrow. Then we must change the language used to describe the image, i.e., select different kernels, eliminating broad ones. In the limit of a field of point sources, the Pixon method becomes a Clean reconstruction(see Section 5.2), using the signal-to-noise ratio to eliminate weak sources.

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 in defense and security.

8.6 Discussion

In this final section 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 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 (e.g., Bontekoe et al. 1991). Unfortunately, producing such events requires considerable effort and is undertaken only rarely. Like most authors, we have not made such efforts and refrain from attempting definitive comparisons. We do, however, take advantage of our expertise in 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 are overfits to the data, with reduced χ^2 (χ^2/n) values of 0.88 and 0.76, for the Wiener and nonnegative least-squares reconstructions, respectively. Figure 3 emphasizes the point that residuals that superficially look good may, in fact, hide overfits. (The values of the reduced χ^2 are significantly below unity. The data contain $n = 512^2 = 262{,}144$ pixels, so the standard deviation of the reduced χ^2 is 0.003.)

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.

Further illustration of the advantage of the Pixon method in the astronomical arena is presented in Figure 5. Reconstructions of the M51/NGC5195 galaxy pair are shown using (*b*) simple coadding, (*c*) the high-resolution (HIRES) method of NASA's Infrared Processing and Analysis Center, based on the maximum-correlation method (Rice 1993), (*d*) the Richardson-Lucy method, (*e*) a commercial version of the maximum-entropy method (MEMSYS), 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.

The source of the raw data is 60μm scans collected by the *Infrared Astronomical Satellite*. This is the only direct comparison of the chosen methods for which each entry was contributed by a leading expert in each technique (Bontekoe et al. 1991). 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. The sensitivity obtained in the Pixon reconstruction is, in fact, a factor of 200 higher than that of the maximum-entropy reconstruction; the linear spatial resolution is better by a factor of three.

Figure 6 shows direct external validation of Pixon processing of 12μm scans taken by the *Infrared Astronomical Satellite.* The *top left* frame shows a collage of scans. For each point in each scan we plot the scan flux at the pixel corresponding to the center of the beam at the time the flux was measured. (When several scans overlap on the same pixel we take the average flux.) The sources are all point-like stars, yet they are spread significantly by the point-response function, particularly in the cross-scan direction. The *top right* panel shows a HIRES reconstruction of these scans 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 in the cross-scan direction is still visible. The *lower left* panel shows the Pixon reconstruction, which reveals many more sources than HIRES and little residual cross-scan blur. Finally, the *lower right* panel shows an image of the same area of the sky, taken twelve years later by the *Midcourse Space Experiment* [**AU: As meant?** Thank you for your correction] (MSX) satellite of the U.S. Air Force with a much improved imaging system, validating the Pixon reconstruction.

Finally, Figure 7 shows an example form nuclear medicine (Vija et al. 2005). 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 Mcts in the *left* panels through 0.8 Mcts in the *middle* panels, to 6.4 Mcts 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 the *lower left* frame and the raw counts in the *upper middle* frame shows that the Pixon images improve image quality in a way that can be achieved by the raw images only by increasing the count by an order of magnitude.

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.

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. It was not designed for the nonparametric case.

Third, the problem of image artifacts is exacerbated because a blurring point-response function has very small eigenvalues whose eigenfunctions are dominated by high spatial frequencies. Upon inversion, the high-frequency components of the data are therefore greatly magnified. When the high-frequency components 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, whereas 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 spatial adaptation results in image restriction that is too loose, so spatial adaptation 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 expected 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 being a random sampling of the parent statistical distribution, which we can and should test 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.

[**AU: Could you please redo the Lit Cited so that article titles are NOT included in the citations? Thanks.** Done]

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

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

Bertero M, Boccacci P. 1998. *Introduction to Inverse Problems in Imaging.* London: Inst. Phys. Publ. 351 pp.

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

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

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

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

Bhatnagar S, Cornwell TJ 2004, *Astron. Astrophys.* 426:747–754

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

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

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

Conchello JA, McNally JG. 1996. *Proc. SPIE* 2655:199–208

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

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

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. Dordrecht: Kluwer. 732 pp.

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. 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

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

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

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

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

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

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

Jones R. 1983. *Holographic and Speckle Interferometry: A Discussion of the Theory, Practice and Application of the Techniques.* Cambridge: Cambridge Univ. Press

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

Kalifa J, Mallat S, Rouge B. 2003. *IEEE Trans. Image Processing* 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

Lucy LB. 1974. *Astron. J.* 79:745–54

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

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

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

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. Press

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

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

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 Processing* 9:295–98

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

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

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

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

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

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

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

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.* In press[!**AU: Update?** To be published first half of 2005 (no date yet)!]

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.* In press [!**AU: Update?** To be published in early 2005 (no date yet)!]

Young EF, Puetter R, Yahil A. 2004. *Geophys. Res. Lett.* 31:L17S09

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

Figure 2: Variety of noniterative image reconstructions: (*a*) Wiener ($\beta = 1$) with (*b*) residuals, (*c*) wavelet ($\beta = 2$) with (*d*) residuals, (*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 ($\chi^2/n = 0.88$), (*e*) converged nonnegative least-squares fit with (*f*) residuals ($\chi^2/n = 0.76$)

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 60μm 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 image, (*c*) NASA high resolution reconstruction, (*d*) Richardson-Lucy reconstruction, (*e*) maximum entropy reconstruction, (*f*) Pixon reconstruction. Objects identified in optical images are also marked in (*a*): (Opt) stars, (Ha) $H\alpha$ emission knots. The black patches in (*e*) and (*f*) represent zero intensity.

Figure 6: Externally validated comparison of Pixon and NASA reconstructions of 12μm 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 twelve years later by the *Midcourse Space Experiment* (MSX) satellite of the U.S. Air Force and processed by the Space Dynamics Laboratory (Logan, UT).

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

Figure 1

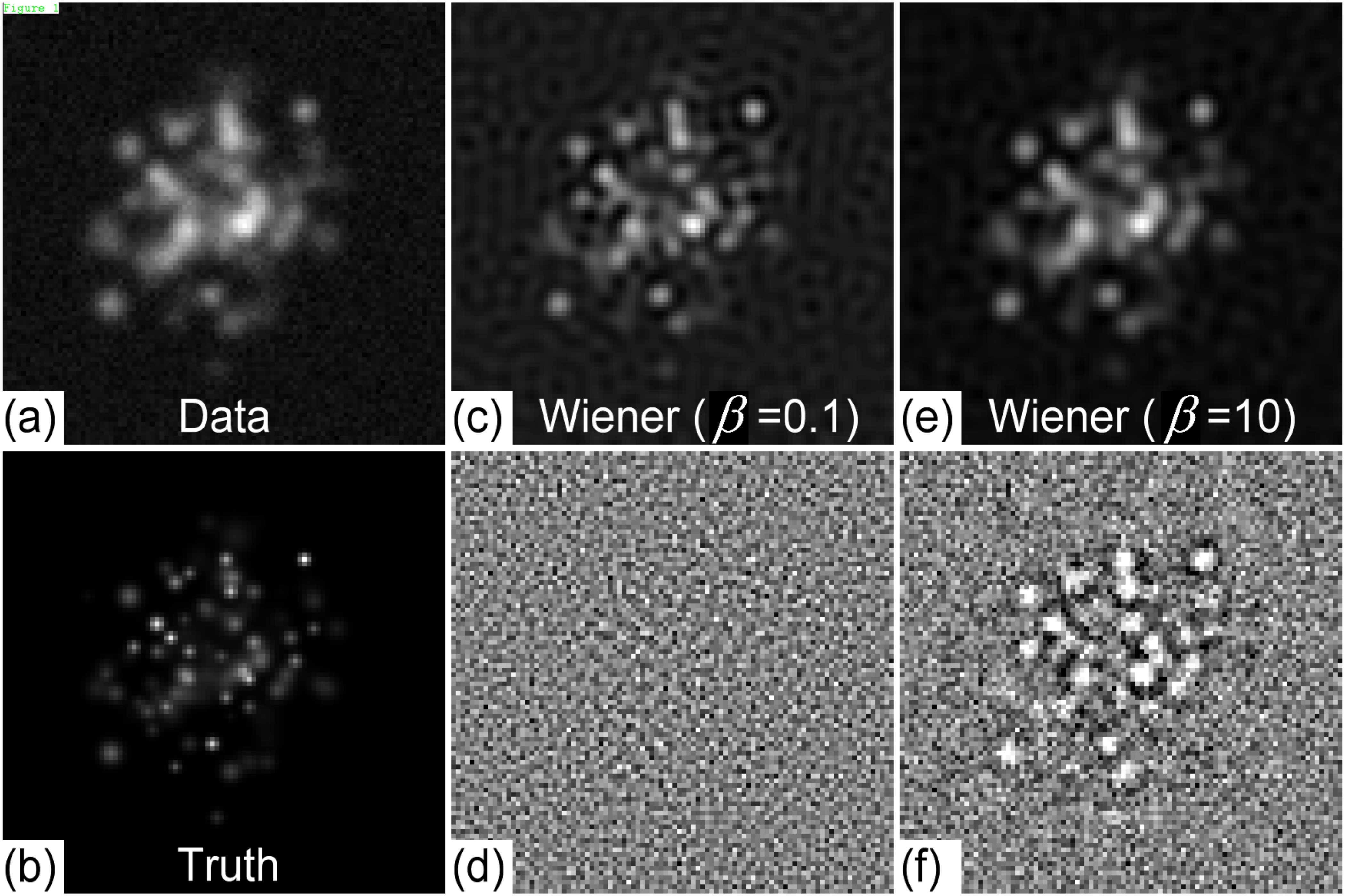


Figure 2

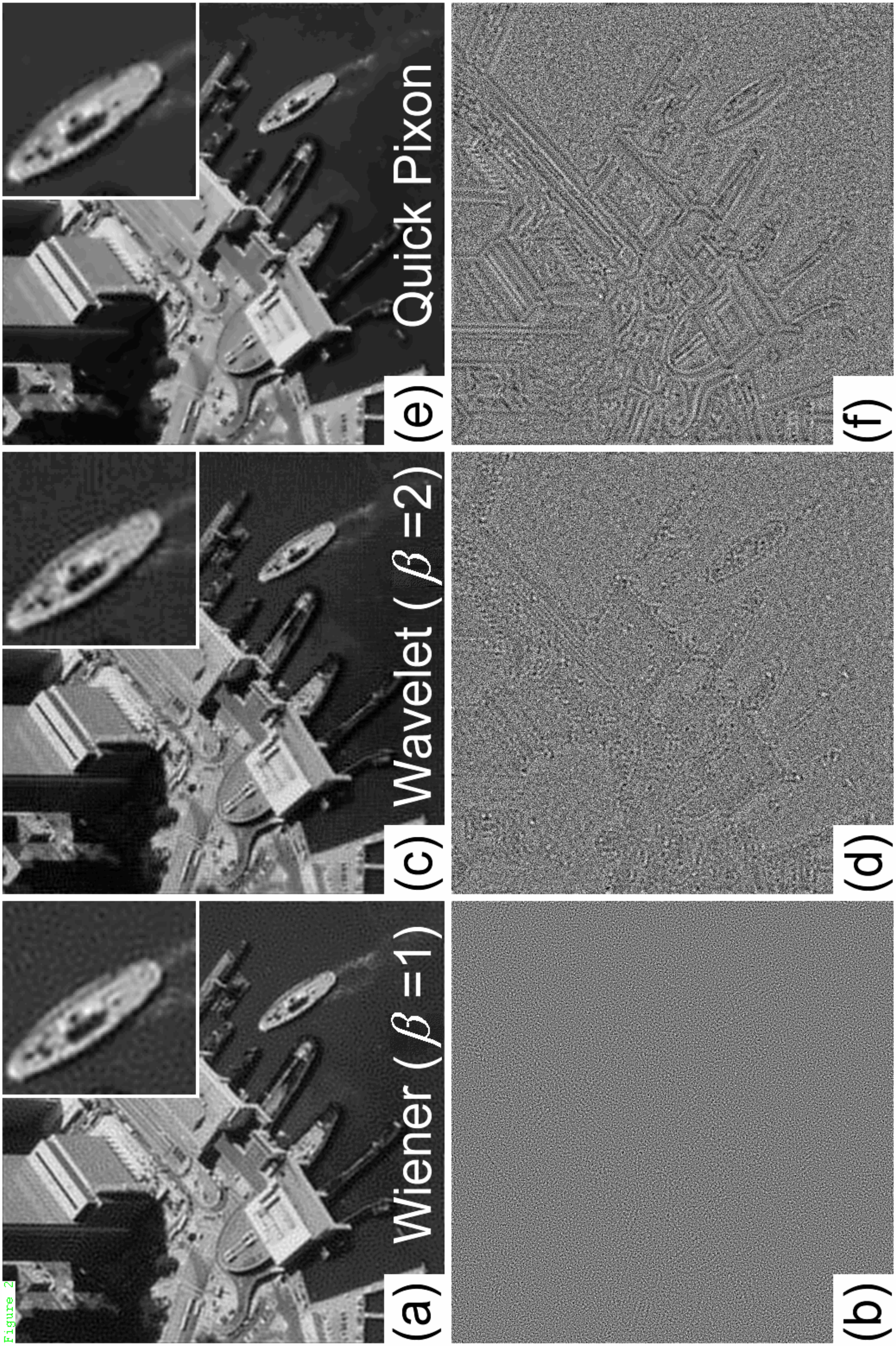


Figure 3

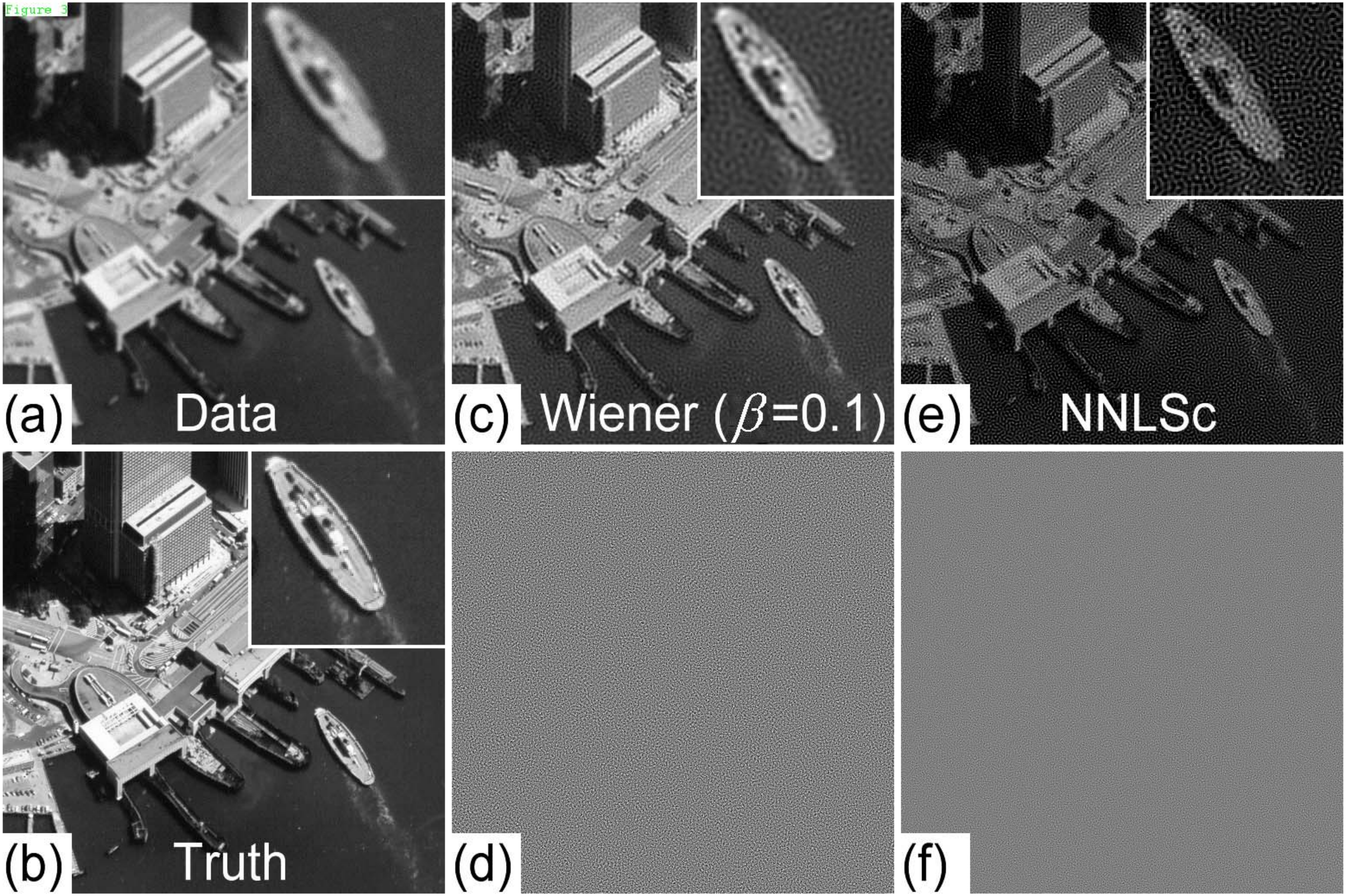


Figure 4

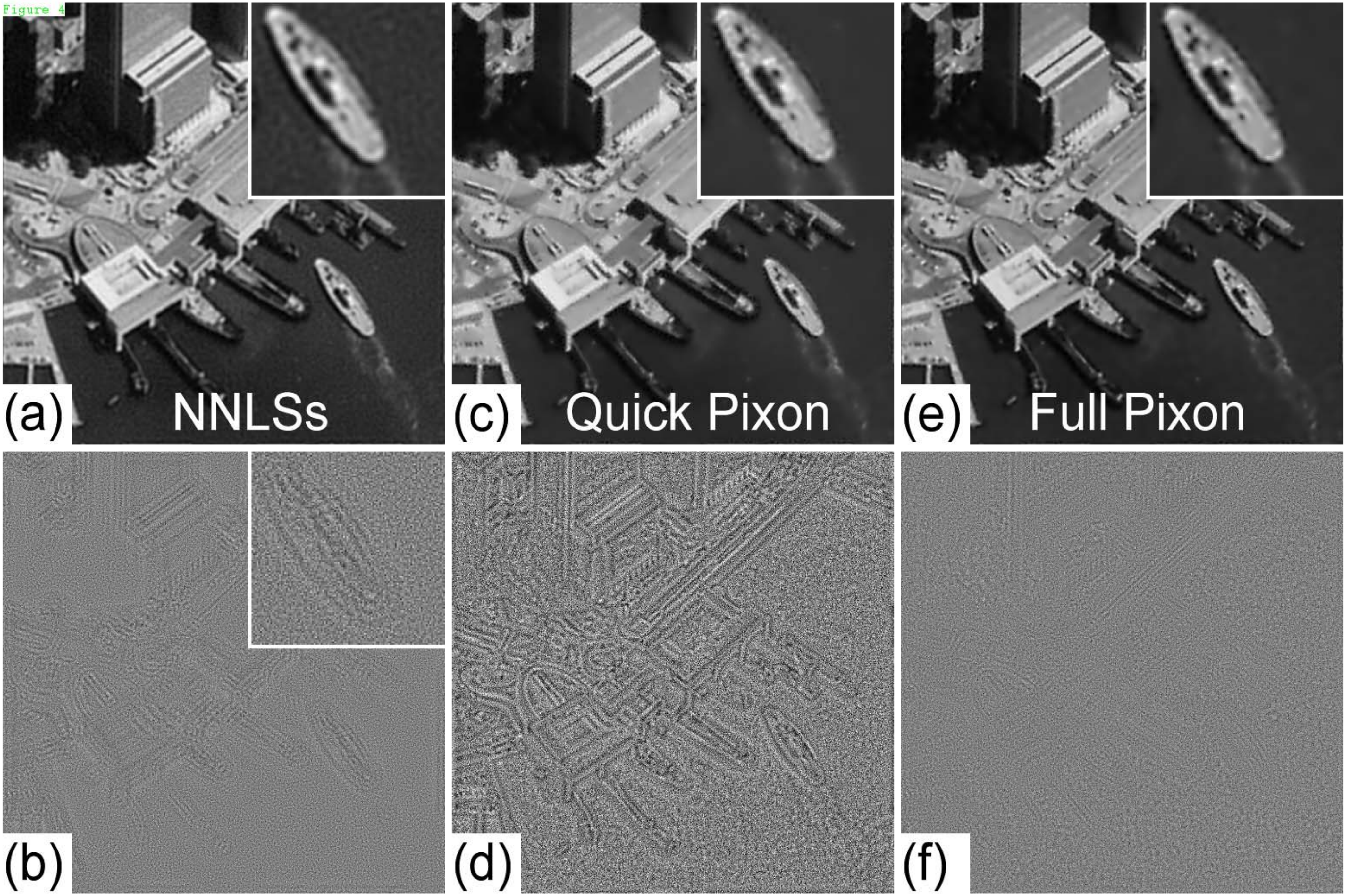
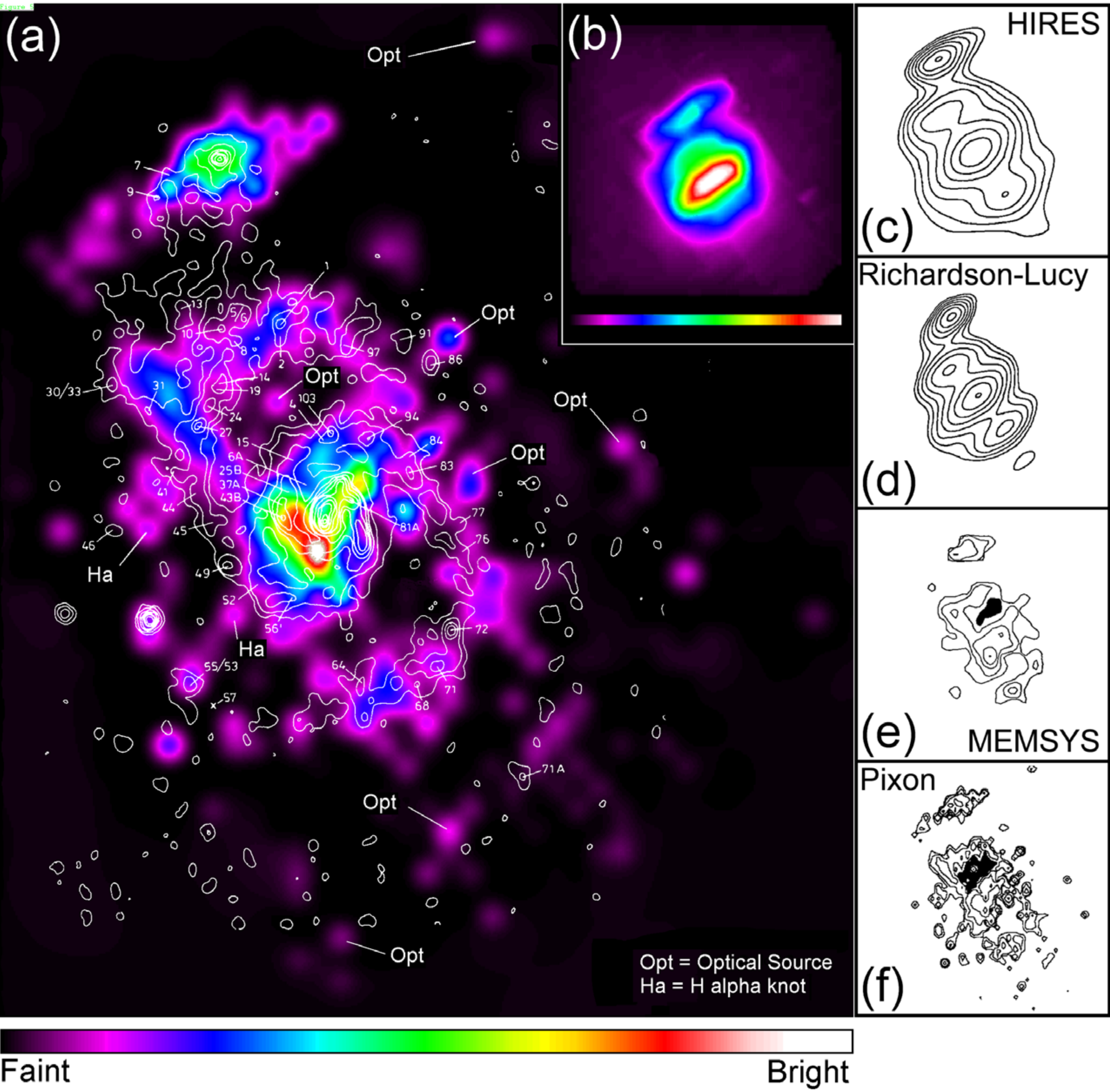


Figure 5


(a)
Opt
7
9
1
13
5/6
10
8
2
91
97
86
Opt
14
19
30/33
31
Opt
103
4
24
27
94
Opt
84
15
6A
25B
37A
43B
83
Opt
41
44
81A
77
45
76
46
Ha
49
52
56'
Ha
72
55/53
64
71
68
x 57
71A
Opt
Opt
Opt = Optical Source
Ha = H alpha knot
(b)
HIRES
(c)
Richardson-Lucy
(d)
(e)
MEMSYS
Pixon
(f)
Faint
Bright

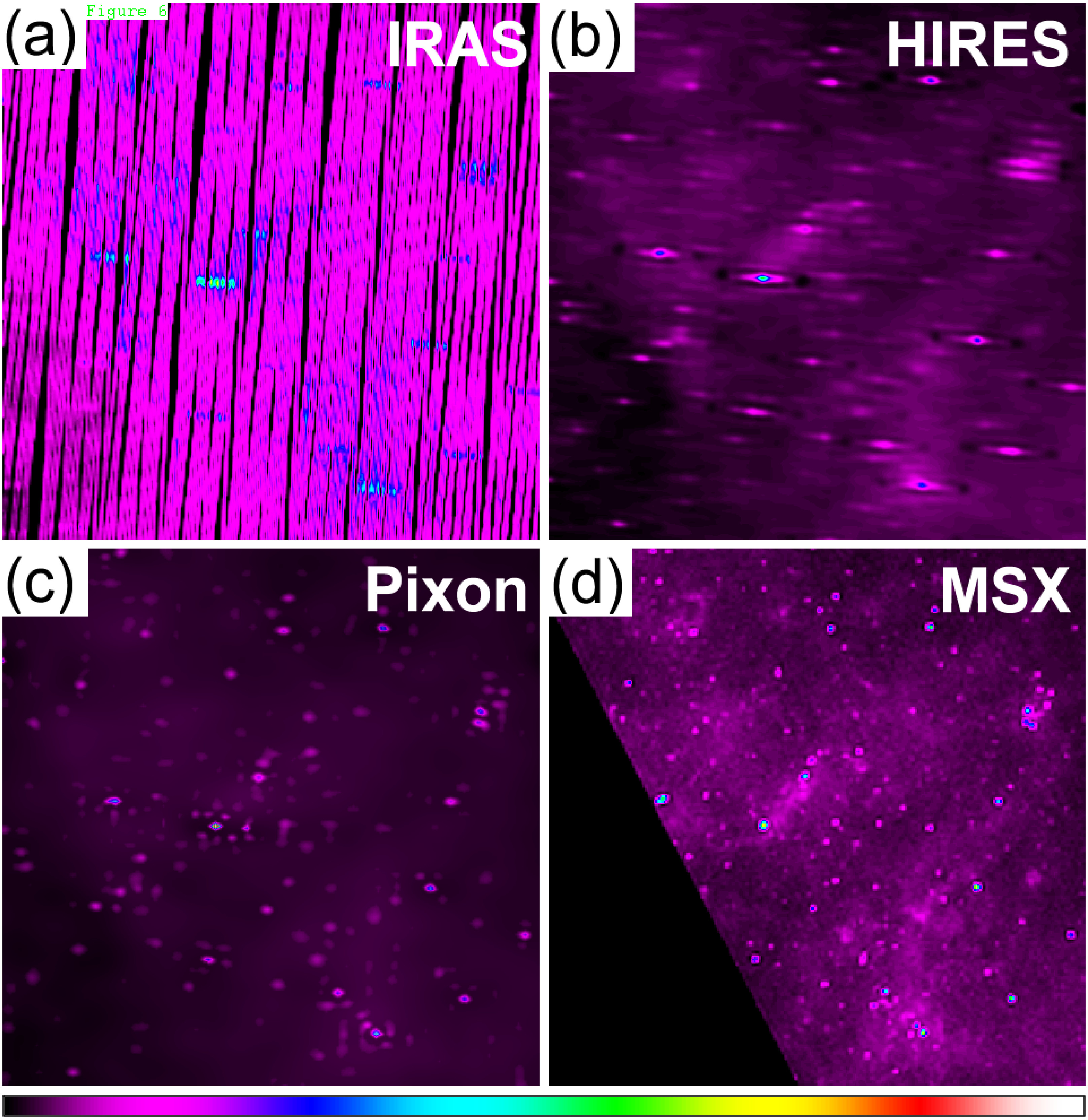
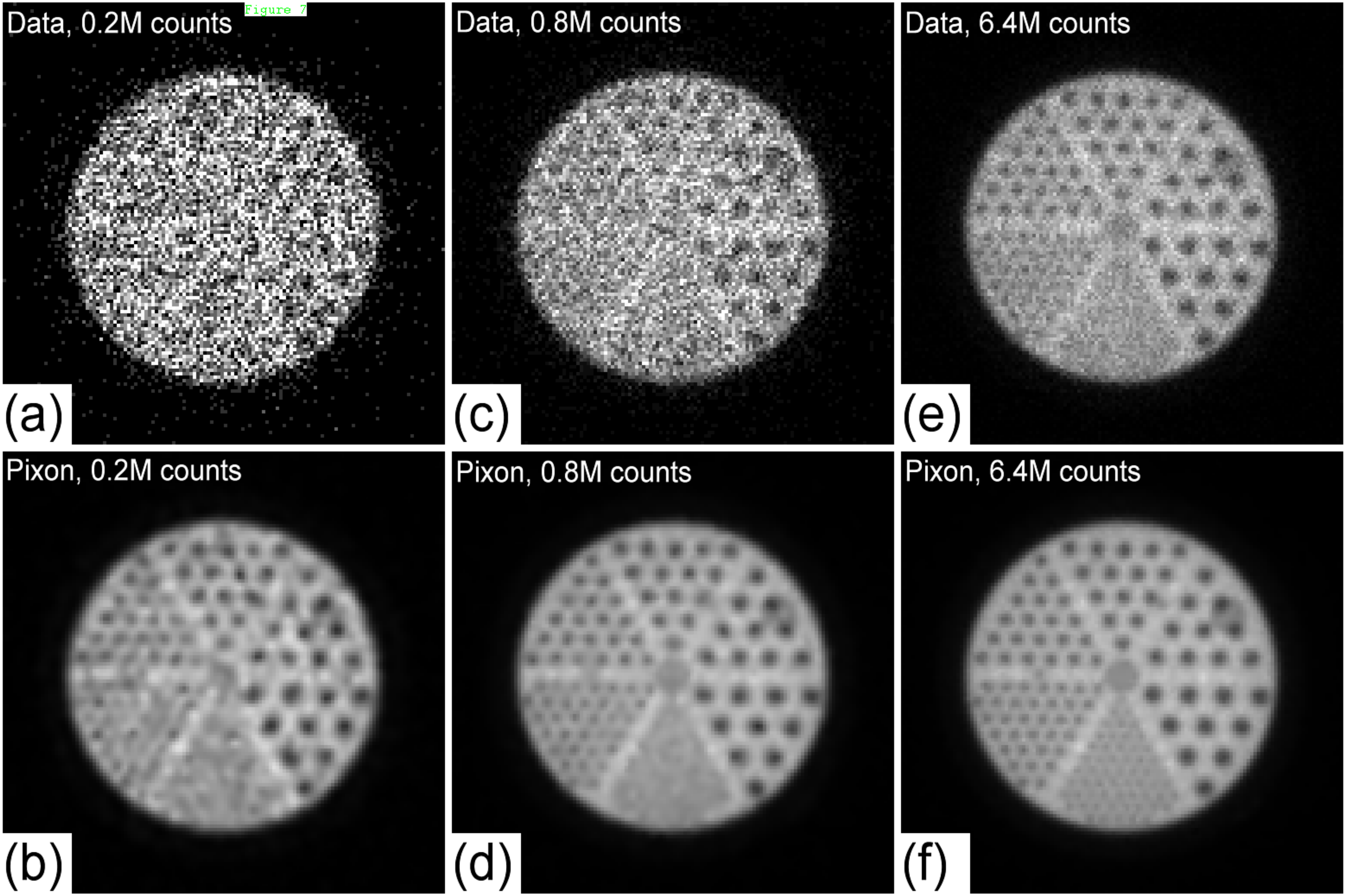


Figure 6
(a)
IRAS
(b)
HIRES
(c)
Pixon
(d)
MSX

Figure 7


Data, 0.2M counts
(a)
Data, 0.8M counts
(c)
Data, 6.4M counts
(e)
Pixon, 0.2M counts
(b)
Pixon, 0.8M counts
(d)
Pixon, 6.4M counts
(f)