Journal of the Royal Statistical Society
SERIES B Statistical Methodology

J. R. Statist. Soc. B (2011)
73, *Part* 4, *pp.* 579–599

Bayesian smoothing of photon-limited images with applications in astronomy

John Thomas White and Subhashis Ghosal

North Carolina State University, Raleigh, USA

[Received December 2009. Revised February 2011]

Summary. We consider a multiscale model for intensities in photon-limited images using a Bayesian framework. A typical Dirichlet prior on relative intensities is not efficient in picking up structures owing to the continuity of intensities. We propose a novel prior using the so-called 'Chinese restaurant process' to create structures in the form of equal intensities of some neighbouring pixels. Simulations are conducted using several photon-limited images, which are common in X-ray astronomy and other high energy photon-based images. Applications to astronomical images from the Chandra X-ray Observatory satellite are shown. The new methodology outperforms most existing methods in terms of image processing quality, speed and the ability to select smoothing parameters automatically.

Keywords: Bayesian smoothing; Chinese restaurant process; Denoising; Multiscale models

1. Introduction

One type of image reconstruction problem addresses photon-based images where relatively few photons are observed. Examples of these types of image can be found in high energy astronomical imaging (Starck and Murtagh, 2006). The methodology that is introduced in this paper addresses this specific type of image restoration and its uses in astronomical imaging.

An image consists of a grid of pixels and, more specifically, a count of photons is obtained in each pixel. These photons are collected by a detection device that causes error itself, but there is also error naturally occurring when photons are emitted from an object and travel to the detector. A common way to resolve this issue is to employ some type of image restoration or reconstruction technique that 'denoises' or smooths the image. This paper introduces one of these such techniques for photon-limited images that will estimate the underlying intensity by using a Poisson model applied to the photon counts that will provide the researcher with a better understanding of the underlying object.

Image reconstruction is not a new field and there are many existing methodologies that use a Gaussian model to smooth planar transmission images. However, in situations with very limited photons, a Gaussian model is not sensible whereas a Poisson model is much more appealing. Photon-limited images commonly occur in X-ray astronomy. For these images, photon counts of pixels may be assumed to be Poisson distributed with parameters given by the true intensity of the image at these pixels, i.e. Poisson noise corrupts the image. Sometimes, in addition to the noise, there are some other systematic aberrations such as the point spread function. In this paper, we consider the situation where the effect of these additional aberrations is negligible.

Address for correspondence: John Thomas White, 2105 Cedar Grove Drive, Durham, NC 27703, USA.
E-mail: John.Thomas.White@gmail.com

 1369–7412/11/73579

The other major type of images that are smoothed are medical images. These images will typically use smoothing methods based on Gaussian models. Since there are usually an abundance of photons in medical photon-based images, Gaussian models adequately describe photon counts. The methodologies that are discussed in this paper are specifically used in photon-limited images. Although there are some forms of medical imaging that may have limited photon counts, this scenario is also common in astronomical imaging (Starck and Murtagh, 2006).

Some methods of image reconstruction with Poisson noise already exist in the literature. Of these methods, a very popular class uses multiscale representations of the images, which are most popularly known from wavelet transforms but are used in the form of recursive partitions in this paper. These representations provide a way to show structure in very detailed or fine objects and other larger or coarse objects in the image simultaneously. In this paper, a new Bayesian multiscale model is proposed that uses mixtures of Dirichlet priors with mixing governed by the so-called 'Chinese restaurant process' (CRP) to denoise the image. Our method has many characteristics of the methods that were described in Kolaczyk (1999), Nowak and Kolaczyk (2000) and Esch *et al*. (2004). There are some key differences in the form of the prior that will be introduced, giving this method advantages over some of these earlier Bayesian models.

Other methods that will be used for comparison are a multiscale penalized likelihood method that was introduced in Kolaczyk and Nowak (2004), a Haar wavelet method using Poisson-corrected thresholds that was described in both Kolaczyk (1997) and Kolaczyk (1998), and platelets (Willett and Nowak, 2003). The platelets methodology is another multiscale image reconstruction technique that furthers the ideas of wedgelets (Donoho, 1999).

We compare the proposed methodology with existing methodologies in terms of the mean absolute deviation and Baddeley's delta metric for grey scale images found in Wilson *et al*. (1997). The latter is considered to be an intelligent metric which is appropriate for measuring similarity of images. It was derived from Baddeley's delta metric for binary images that was introduced in Baddeley (1992). This metric established itself as a useful measure of comparing image smoothing methodologies.

In Section 2, the new Bayesian multiscale statistical model is described in detail. Consistency results are provided in Section 3. Comparisons are made using test images where the maximum intensity and background emission can be decided and the true relative intensity is known. Images and results from these simulations are given in Section 4. Using Baddeley's delta metric for greyscale images demonstrates the superiority of the method that is introduced in this paper. Examples in astronomical imaging by using the Chandra X-ray Observatory are shown in Section 5. Finally, conclusions and remarks are given in Section 6. A basic introduction to the CRP, which is useful in understanding the method, is given in Appendix A.

2. Model and methodology

2.1. Bayesian multiscale statistical model

Multiscale representations of images decompose an image into different scales of data from coarse to fine with the two extremes in this case represented by the total photon flux and the pixel scale photon counts respectively. This representation will allow for different levels of smoothing to be used at the different scales of the image. Multiscale models have been used in many image reconstruction methods such as platelets, any wavelet-based methods, wedgelets and some Bayesian methods.

One common way to decompose an image in the multiscale representation is to start with the image as one block containing the total count of photons. This block is then sectioned into four smaller equal-sized blocks with the corresponding counts data. We continue to partition these

blocks into four equal cells until the pixel scale is obtained. For this decomposition, the image must have pixels on both sides of the form $n = 2^L$ where L is an integer that denotes the finest scale in the decomposition and $N = n^2 = 4^L$ is the total number of pixels in the image.

Let $X_{(j,k)} = X_{L,(j,k)}$ denote the photon count in pixel (j,k) and $X_{0,(1,1)}$ denote the total count of photons or photon flux of the image. At intermediate scales $l = 1, \ldots, L-1$, there are 4^l block pixels that are denoted by (j,k) where $j,k = 1, \ldots, 2^l$. To obtain the intermediate scale data, the following formula is used:

$$X_{l,(j,k)} = X_{l+1,(2j-1,2k-1)} + X_{l+1,(2j-1,2k)} + X_{l+1,(2j,2k-1)} + X_{l+1,(2j,2k)}, \quad j,k = 1, \ldots, 2^l, \quad l = 0, \ldots, L-1. \tag{1}$$

Collectively this is called a parent–child group, with the count of the photons of the parent on the left, and the counts of photons for each of the four children on the right. This decomposition can also be thought of as a quad tree where every node is split into four leaves as we move down the tree. Each level represents a different scale of data.

The observed image data comprise pixels in which there are counts of photons. Given the intensity of the image source at the pixel locations, these counts are modelled as independent with a Poisson distribution: $X_{(j,k)} \sim \text{Poisson}(\lambda_{(j,k)})$, where (j,k) denotes the pixel with j being the horizontal index and k being the vertical index. The underlying intensity λ represents the object that is being examined. These images may sometimes have background noise. This will be ignored since it is usually much fainter than the brightness of the object under study (Esch *et al.*, 2004). Images can have varying sizes. To fit in the multiscale setting, an image may have to be truncated or blank pixels may have to be added to make each side a length of 2^L.

The entire image can be thought of by $\mathbf{X} \sim \text{Poisson}(\mathbf{\Lambda})$. A multiscale statistical model, which was also used in Kolaczyk (1999), Nowak and Kolaczyk (2000) and Esch *et al.* (2004), is given by factoring this simple statistical model into

$$P(\mathbf{X}|\mathbf{\Lambda}) = \mathcal{P}(X_{0,(1,1)}|\lambda_{0,(1,1)}) \prod_{l=0}^{L-1} \prod_{j=1}^{2^l} \prod_{k=1}^{2^l} \mathcal{M}\left\{ \begin{pmatrix} X_{l+1,(2j-1,2k-1)} \\ X_{l+1,(2j-1,2k)} \\ X_{l+1,(2j,2k-1)} \\ X_{l+1,(2j,2k)} \end{pmatrix} | X_{l,(j,k)}, \begin{pmatrix} \rho^1_{l,(j,k)} \\ \rho^2_{l,(j,k)} \\ \rho^3_{l,(j,k)} \\ \rho^4_{l,(j,k)} \end{pmatrix} \right\}, \tag{2}$$

where $\mathcal{P}$ denotes the Poisson distribution and $\mathcal{M}$ the multinomial distribution, and ρ is a 4×1 vector of probabilities, distributing parent photons to its children. The ρs and λs have a one-to-one relationship; therefore, knowing the ρs will give λ by

$$\left.\begin{aligned} \rho^1_{l,(j,k)} \lambda_{l,(j,k)} &= \lambda_{l+1,(2j-1,2k-1)}, \\ \rho^2_{l,(j,k)} \lambda_{l,(j,k)} &= \lambda_{l+1,(2j-1,2k)}, \\ \rho^3_{l,(j,k)} \lambda_{l,(j,k)} &= \lambda_{l+1,(2j,2k-1)}, \\ \rho^4_{l,(j,k)} \lambda_{l,(j,k)} &= \lambda_{l+1,(2j,2k)}. \end{aligned}\right\} \tag{3}$$

To find an estimate of $\mathbf{\Lambda}$, the values of ρ must be estimated. Instead of directly obtaining estimates for the intensities, an estimate is obtained for all the split probabilities ρ, and their relationship is exploited to obtain the intensities. A pixel scale intensity estimator can be estimated as $\hat{\lambda}_{j,k} = E(\lambda_0 \rho_0 \rho_1 \ldots \rho_{L-1} | \mathbf{X})$, which is the posterior mean of the pixel scale intensity since $\lambda_0 \rho_0 \rho_1 \ldots \rho_{L-1} = \lambda_{(j,k)}$; here we leave off the horizontal and vertical subscripts for notational simplicity. Owing to the independence of each of these scales obtained from this multiscale

factorization, the expected value may be split into the expected value of each split probability, i.e.

$$\hat{\lambda}_{(j,k)} = E(\lambda_0|\mathbf{X}_0)\, E(\rho_0|\mathbf{X}_1) \ldots E(\rho_{L-1}|\mathbf{X}_L), \tag{4}$$

where

$$\mathbf{X}_l = (X_{l,(j,k)} : j,k = 1,\ldots,2^l; l = 0,\ldots,L). \tag{5}$$

The estimates of each of the split probabilities can be found separately as outlined in Kolaczyk and Nowak (2004).

2.2. Prior distributions

The prior distribution that is specified for the split probabilities is different in each of the earlier Bayesian multiscale models. Originally, when this idea was introduced in Kolaczyk (1999), the model was constructed for a one-dimensional signal instead of a two-dimensional image. The prior that was used in Kolaczyk (1999) in the one-dimensional setting with l as the scale and j as the position was given by

$$\rho_{l,(j)}|\gamma_{l,(j)} \sim \tfrac{1}{2}\gamma_{l,(j)} + (1-\gamma_{l,(j)})\,\text{beta}(1,1), \qquad \gamma_{l,(j)}|p_l \sim \text{Bernoulli}(p_l).$$

Later, in Nowak and Kolaczyk (2000), this prior was changed to

$$\rho_{l,(j)} \sim \text{beta}(\alpha_l, \alpha_l), \qquad \alpha_l = 2\alpha_{l-1}.$$

However, in Nowak and Kolaczyk (2000), they outlined the use of these Bayesian multiscale models for two-dimensional image problems where they used maximum *a posteriori* (MAP) estimates of these split probabilities to smooth images. They made this transfer to the two-dimensional image by splitting the parent once into vertical pieces and then next splitting each of these pieces, to end up with four splits of the parent. This idea was mentioned in Kolaczyk (1999) but was never formally examined for that type of prior. Finally, in Esch *et al.* (2004), the prior was modified again to incorporate a hyperprior on the parameters. This required the use of Markov chain Monte Carlo methods, which consequently increased the computation time.

In this paper, we take the idea from Kolaczyk (1999) and transfer it to two dimensions, but we split those into four groups at once, not first vertically and then horizontally, thus not spoiling the rotational invariance. In this way, it would not matter if the image was rotated 90° before beginning. Our motivation is to allow all types of configurations of ties between the children. Consider a group of the top two children and the bottom left child. This is only possible when splitting the four children at once.

The prior probabilities for this quad split are obtained from the CRP. The CRP, as discussed in Appendix A, allows for all configurations of the ρs to equal each other within a parent–child group with certain probability given by the parameter of the CRP. Since the CRP allows ties in the values of continuously distributed variables, the resulting prior encourages formations of structures in the image by identifying neighbouring pixels with equal intensity as part of a continuous object.

Let a configuration of ρs formed by grouping be denoted by $\mathcal{C}$, and let $\mathscr{C}$ be the space of all 15 possible configurations. These possibilities are

$\{(1234)\}$

$\{(123)4\}, \{(124)3\}, \{(134)2\}, \{(234)1\}$

$\{(12)(34)\}, \{(13)(24)\}, \{(14)(23)\}$

$\{(12)34\}, \{(13)24\}, \{(14)23\}, \{(23)14\}, \{(24)13\}, \{(34)12\}$

$\{1234\}$,

where $(\cdot)$ denote ties. Each parent–child group is represented as

1	3
2	4

The simplest configuration is the equal split into each of the four cells so that all ρs are equal in a parent–child group. This is given by $\rho^{(1)} = \rho^{(2)} = \rho^{(3)} = \rho^{(4)} = \frac{1}{4}$ or $\{(1234)\}$ in the notation that was defined above.

This allows for optimal smoothing in very fine scales when there is no significant difference between neighbouring pixels. Having the same ρs will ensure the same λs for children of the same parent–child group. The other extreme configuration is when all ρs are unique. This occurs when the children are very different from one another in a group. Out of the 15 possible configurations, three seem to be somewhat unnatural. They correspond to the cases where diagonally opposite children are tied—two for each diagonal $\{(14)23\}$ and $\{(23)14\}$, and one more for both diagonals $\{(14)(23)\}$. Since we tend to view ties as the formation of structures by adjacent pixels, it is natural to remove the three above-mentioned configurations and to redistribute their prior probabilities to configurations of the same type. In other words, groups must be made with adjacent blocks. We observed that these configurations are not typically supported by the data and hence seldom have high posterior probability. Therefore their inclusion or exclusion does not affect the final result to a large extent. However, their exclusion allows somewhat faster computation.

The probabilities are given as $P(\mathcal{C}|M) \sim \mathrm{CRP}(M)$, where M is the parameter of the CRP that is described in Appendix A. There are now 12 possible configurations of the split probabilities. Derived from information in Appendix A, an example of each type of configuration and their probability is

$$\left.\begin{aligned}
P(\mathcal{C} = \{(1234)\}|M) &= \frac{M}{M}\,\frac{1}{M+1}\,\frac{2}{M+2}\,\frac{3}{M+3},\\
P(\mathcal{C} = \{(123)4\}|M) &= \frac{M}{M}\,\frac{1}{M+1}\,\frac{2}{M+2}\,\frac{M}{M+3},\\
P(\mathcal{C} = \{(12)(34)\}|M) &= \frac{3}{2}\,\frac{M}{M}\,\frac{1}{M+1}\,\frac{M}{M+2}\,\frac{1}{M+3},\\
P(\mathcal{C} = \{(12)34\}|M) &= \frac{3}{2}\,\frac{M}{M}\,\frac{1}{M+1}\,\frac{M}{M+2}\,\frac{M}{M+3},\\
P(\mathcal{C} = \{1234\}|M) &= \frac{M}{M}\,\frac{M}{M+1}\,\frac{M}{M+2}\,\frac{M}{M+3}.
\end{aligned}\right\} \tag{6}$$

Note the constant of $\frac{3}{2}$ in expression (6). Since we disregard $\{(14)(23)\}$, $\{(14)23\}$ and $\{(23)14\}$, we redistribute their probabilities to the remaining configurations of the same type, such that some will have a factor of $\frac{3}{2}$ to make the sum of the configuration probabilities equal 1. We shall call this a modified CRP. A particular configuration can be written as $\mathcal{C} = \{C_1, \ldots, C_s\}$, where s is the number of groups in the configuration. All different configurations within each parent–child group form a mixture prior for the split probabilities. We can express $(\rho^{(1)}_{l,(j,k)}, \rho^{(2)}_{l,(j,k)}, \rho^{(3)}_{l,(j,k)}, \rho^{(4)}_{l,(j,k)})$ as a function of a vector $(q^{(1)}_{l,(j,k)}, \ldots, q^{(s)}_{l,(j,k)})$ on the unit simplex in $\mathbb{R}^s$:

$$(\rho^{(1)}_{l,(j,k)}, \rho^{(2)}_{l,(j,k)}, \rho^{(3)}_{l,(j,k)}, \rho^{(4)}_{l,(j,k)}) = H_{\mathcal{C}}(q^{(1)}_{l,(j,k)}, \ldots, q^{(s)}_{l,(j,k)}), \tag{7}$$

for some $s = 1, 2, 3, 4$.

To illustrate how this function $H_{\mathcal{C}}$ works, some examples are given.

(a) If $\mathcal{C} = \{1234\}$, then $\rho^{(1)}_{l,(j,k)} = q^{(1)}_{l,(j,k)}, \rho^{(2)}_{l,(j,k)} = q^{(2)}_{l,(j,k)}, \rho^{(3)}_{l,(j,k)} = q^{(3)}_{l,(j,k)}$ and $\rho^{(4)}_{l,(j,k)} = q^{(4)}_{l,(j,k)}$.
(b) If $\mathcal{C} = \{12(34)\}$, then $\rho^{(1)}_{l,(j,k)} = q^{(1)}_{l,(j,k)}, \rho^{(2)}_{l,(j,k)} = q^{(2)}_{l,(j,k)}$ and $\rho^{(3)}_{l,(j,k)} = \rho^{(4)}_{l,(j,k)} = q^{(3)}_{l,(j,k)}/2$.
(c) If $\mathcal{C} = \{(12)(34)\}$, then $\rho^{(1)}_{l,(j,k)} = \rho^{(2)}_{l,(j,k)} = q^{(1)}_{l,(j,k)}/2$ and $\rho^{(3)}_{l,(j,k)} = \rho^{(4)}_{l,(j,k)} = q^{(2)}_{l,(j,k)}/2$.
(d) If $\mathcal{C} = \{(123)4\}$, then $\rho^{(1)}_{l,(j,k)} = \rho^{(2)}_{l,(j,k)} = \rho^{(3)}_{l,(j,k)} = q^{(1)}_{l,(j,k)}/3$ and $\rho^{(4)}_{l,(j,k)} = q^{(2)}_{l,(j,k)}$.
(e) If $\mathcal{C} = \{(1234)\}$, then $\rho^{(1)}_{l,(j,k)} = \rho^{(2)}_{l,(j,k)} = \rho^{(3)}_{l,(j,k)} = \rho^{(4)}_{l,(j,k)} = \frac{1}{4}$.

Prior distributions are then specified for the q-parameters as follows:

$$P(q^{(1)}_{l,(j,k)}, \ldots, q^{(s)}_{l,(j,k)} | \mathcal{C} \in \mathscr{C}) \sim \text{Dirichlet}(s; \aleph_{C_1}, \ldots, \aleph_{C_s}), \tag{8}$$

where $\aleph$ represents the cardinality of a set and s is the number of distinct groups. The parameters in equation (5) are chosen to make the prior expectation of $\rho^{(1)}_{l,(j,k)}, \ldots, \rho^{(4)}_{l,(j,k)}$ equal to $\frac{1}{4}$, i.e. an *a priori* fair split. Further, since variation is already controlled by M through the probabilities of the ties, any additional parameter in distribution (8) is unnecessary. For instance, we did not need to use a Dirichlet$(s; \alpha\aleph_{C_1}, \ldots, \alpha\aleph_{C_s})$ prior with undetermined α. Again, some examples are given to illustrate this construction.

(a) If $\mathcal{C} = \{1234\}$, then $P(q^{(1)}_{l,(j,k)}, q^{(2)}_{l,(j,k)}, q^{(3)}_{l,(j,k)}, q^{(4)}_{l,(j,k)} | \mathcal{C}) \sim \text{Dirichlet}(4; 1, 1, 1, 1)$.
(b) If $\mathcal{C} = \{12(34)\}$, then $P(q^{(1)}_{l,(j,k)}, q^{(2)}_{l,(j,k)}, q^{(3)}_{l,(j,k)} | \mathcal{C}) \sim \text{Dirichlet}(3; 1, 1, 2)$.
(c) If $\mathcal{C} = \{(12)(34)\}$, then $P(q^{(1)}_{l,(j,k)}, q^{(2)}_{l,(j,k)} | \mathcal{C}) \sim \text{Dirichlet}(2; 2, 2)$.
(d) If $\mathcal{C} = \{(123)4\}$, then $P(q^{(1)}_{l,(j,k)}, q^{(2)}_{l,(j,k)} | \mathcal{C}) \sim \text{Dirichlet}(2; 3, 1)$.
(e) If $\mathcal{C} = \{(1234)\}$, then $\rho = \{0.25, 0.25, 0.25, 0.25\}$.

The parameter M drives the smoothing in this image reconstruction. Thus, determining M is of great importance. In general, decreasing the value of M will create more ties among the ρs whereas increasing its value will create more distinct groups.

An important property of an image reconstruction technique is to keep the photon flux of the original image. Therefore, instead of placing a prior on the overall intensity, it is set equal to the photon flux of the observed image. Thus, we set $\lambda_{0,(1,1)} = X_{0,(1,1)}$ rather than putting a prior on $\lambda_{0,(1,1)}$.

2.3. Posterior distributions

Assume, for now, that M is given. The goal is to find an estimate of $E(\lambda_{(j,k)} | \mathbf{X})$ by using equation (5). The posterior mean $E(\lambda_{(j,k)} | \mathbf{X})$ can be computed analytically exploiting posterior independence of the ρs and analytically computing their posterior means individually. To do this, it is necessary to find two general expressions: the discrete distribution $P(\mathcal{C} | M, X)$ and the continuous distribution $P(q_{l,(j,k)} | \mathcal{C} \in \mathscr{C}, X)$, both of which have the same general form for each parent–child group.

We start with the discrete distribution of configurations in a given parent–child group. Let (X_1, X_2, X_3, X_4) denote the observed photon counts of the children in that group and X denote the observed photon count of their parent (i.e. the sum of the photon counts of the children). Then we can express this distribution as

$$P(\mathcal{C} | M, X) \propto P(\mathcal{C} | M)\, P(X_1, X_2, X_3, X_4 | \mathcal{C}, X). \tag{9}$$

Since this is a discrete probability distribution, the values must sum to 1 giving the constant of proportionality. The first term $P(\mathcal{C} | M)$ is given by the modified CRP as in expression (6). The second term $P(X_1, X_2, X_3, X_4 | \mathcal{C}, X)$ is found by using the following integration:

$$P(X_1, X_2, X_3, X_4|\mathcal{C}, X) \propto \int_{\triangle_s} P(X_1, X_2, X_3, X_4|\mathcal{C}, X, \rho)\, P(q|\mathcal{C})\, \mathrm{d}q, \tag{10}$$

where $\triangle_s$ stands for the unit simplex in $\mathbb{R}^s$. This integration is performed for each of the 12 possible configurations to obtain a closed form solution. As an example, take $\mathcal{C} = \{(12)(34)\}$. Then

$$\begin{aligned} &P(X_1, X_2, X_3, X_4|X, \mathcal{C}, \rho^{(1)} = \rho^{(2)} = q^{(1)}/2, \rho^{(3)} = \rho^{(4)} = q^{(2)}/2)\, P(q^{(1)}, q^{(2)}|\mathcal{C}) \\ &\qquad = \frac{X!}{X_1!X_2!X_3!X_4!} \left(\frac{q^{(1)}}{2}\right)^{X_1+X_2} \left(\frac{q^{(2)}}{2}\right)^{X_3+X_4} \frac{\Gamma(4)}{\Gamma(2)\Gamma(2)} q^{(1)2-1} q^{(2)2-1}. \end{aligned}$$

Integrate out $(q^{(1)}, q^{(2)})$ to obtain

$$\begin{aligned} P(X_1, X_2, X_3, X_4|\mathcal{C}, X) &\propto 6 \frac{\Gamma(X_1 + X_2 + 2)}{\Gamma(X_1 + 1)\Gamma(X_2 + 1)} \frac{\Gamma(X_3 + X_4 + 2)}{\Gamma(X_3 + 1)\Gamma(X_4 + 1)} \frac{(\frac{1}{2})^X}{(X+3)(X+2)(X+1)} \\ &= 6 \frac{1}{\mathrm{beta}(X_1, X_2)} \frac{1}{\mathrm{beta}(X_3, X_4)} \frac{(\frac{1}{2})^X}{(X+3)(X+2)(X+1)} \\ &= 6 \frac{\exp\{-\mathrm{logbeta}(X_1, X_2) - \mathrm{logbeta}(X_3, X_4) - X \ln(2)\}}{(X+3)(X+2)(X+1)}. \end{aligned} \tag{11}$$

The standard library function logbeta (the logarithm of the beta function) is used to avoid numerical instability in computation since the values of X can become very large. By repeating this process, all 12 of these probabilities of configurations are obtained.

Given an M, we now have the discrete distribution $P(\mathcal{C}|M, X)$ for a parent–child group. The other posterior distribution, $P(q_{l,(j,k)}|\mathcal{C}, X)$, is simply a Dirichlet distribution by conjugacy since the likelihood of the data is a multinomial distribution. For instance, if $\mathcal{C} = \{(12)(34)\}$, then $(q^{(1)}, q^{(2)}|X_1, X_2, X_3, X_4, \mathcal{C}) \sim \mathrm{Dirichlet}(2; 2 + X_1 + X_2, 2 + X_3 + X_4)$. The posterior mean of the split probabilities can be obtained by using weights from each of the 12 configurations, i.e.

$$E(\rho|\mathbf{X}) = \sum_{\mathcal{C} \in \mathcal{C}} P(\mathcal{C}|\mathbf{X})\, E(\rho|\mathcal{C}, \mathbf{X}).$$

Each of the ρ-parameters is obtained in all scales, giving $E(\lambda_{(j,k)}|\mathbf{X})$ from equation (5). Cycle spinning (Coifman and Donoho, 1995) would have to be used to avoid staircase-like artefacts from the quad tree partitions. However, by obtaining a piecewise constant estimate, we can use a fast translation invariant approach to this image reconstruction as outlined in Willett and Nowak (2004).

Moreover, $E(\lambda^2_{(j,k)}|\mathbf{X}) = E(\lambda_0^2|\mathbf{X}_0)\, E(\rho_0^2|\mathbf{X}_1) \ldots E(\rho^2_{L-1}|\mathbf{X}_L)$ can be computed in the same fashion, giving posterior variances as a measure of precision of the posterior mean in estimating the intensity parameters. In fact, the entire posterior distribution can be described by drawing independent samples from the exact posterior distribution.

In comparison, for the original Bayesian multiscale models of Kolaczyk (1999), the posterior means were also obtained analytically. In Nowak and Kolaczyk (2000), when the Bayesian multiscale model was presented for two-dimensional images, MAP estimates were used. Both methods used cycle spinning. It is shown in simulations in Section 4 that our model outperforms the earlier Bayesian model that was specified for two-dimensional images.

2.4. Selection of smoothing parameters

The tuning parameter M decides how much smoothing will be done in our model. Two different methods of choosing M were used and compared. In both methods, M is different for the

multiple scales of data, which adjusts the level of smoothing. In the CRP, a lower value of M corresponds to more ties and hence a higher level of smoothing. It is intuitively obvious that more smoothing will be necessary as the scale becomes finer, i.e. as l increases and hence the value of M should become smaller. In the first approach, the value of M for different parent–child groups at a given scale is taken to be the same and an empirical estimate of M is obtained by maximizing the marginal likelihood. In the second approach, we let the smoothing parameter M corresponding to the different parent–child groups at the same level be independent and identically distributed following a gamma distribution with shape parameter α and scale parameter β. The values of α and β are possibly different in different scales and are chosen by maximizing the marginal likelihood at a pair (α, β) after integrating out the M-values.

The tuning parameters have a direct effect on the distribution of the configuration, which is given by

$$P(\mathcal{C}|X_1, X_2, X_3, X_4, M) \propto P(\mathcal{C}|M)\, P(X_1, X_2, X_3, X_4|\mathcal{C}, X) \tag{12}$$

in the first approach or

$$P(\mathcal{C}|X_1, X_2, X_3, X_4, \alpha, \beta) \propto P(X_1, X_2, X_3, X_4|\mathcal{C}, X) \int_0^\infty P(\mathcal{C}|M) P(M|\alpha, \beta)\, \mathrm{d}M \tag{13}$$

in the second approach.

In either case, $P(X_1, X_2, X_3, X_4|\mathcal{C}, X)$ has already been obtained in Section 2.3. The smoothing parameters are optimized to maximize the marginal distribution in both cases.

To find M directly, the marginal probability of obtaining the given sample is maximized by using a Newton–Raphson algorithm on

$$P(\mathbf{X}|M) = \sum_{\mathcal{C} \in \mathcal{C}} P(\mathcal{C}|M) P(X_1, X_2, X_3, X_4|\mathcal{C}, X). \tag{14}$$

When using the hyperprior, the marginal probability of obtaining the given sample is maximized for these hyperparameters:

$$P(\mathbf{X}|\alpha, \beta) = \sum_{\mathcal{C} \in \mathcal{C}} P(X_1, X_2, X_3, X_4|\mathcal{C}, X) \int_0^\infty P(\mathcal{C}|M) P(M|\alpha, \beta)\, \mathrm{d}M. \tag{15}$$

The integral on the right-hand part of the sum can be obtained for each parent–child group since M is allowed to vary freely in a scale coming from the same distribution. To calculate this integral, analytical computation is actually possible to some extent in the sense that we can reduce the integral to an expression involving only library functions in standard software like MATLAB. There are five distinct values of the CRP probabilities. They merely differ in the exponent of M and by a constant factor. Thus, we have the form

$$\begin{aligned}
\int_0^\infty P(\mathcal{C}|M)\, P(M|\alpha, \beta)\, \mathrm{d}M &\propto \int_0^\infty \frac{M^k}{(M+1)(M+2)(M+3)} M^{\alpha-1} \exp(-M\beta)\, \mathrm{d}M \\
&= \int_0^\infty \frac{M^{k+\alpha-1} \exp(-M\beta)}{(M+1)(M+2)(M+3)} \mathrm{d}M \\
&= \int_0^\infty \frac{M^{k+\alpha-1} \exp(-M\beta)}{2(M+3)} \mathrm{d}M - \int_0^\infty \frac{M^{k+\alpha-1} \exp(-M\beta)}{M+2} \mathrm{d}M \\
&\quad + \int_0^\infty \frac{M^{k+\alpha-1} \exp(-M\beta)}{2(M+1)} \mathrm{d}M.
\end{aligned} \tag{16}$$

It will suffice to solve the integral below with $k^* = k + \alpha - 1$ from above since each term in the sum is of the same form. If k^* is an integer,

$$\int_0^\infty \frac{M^{k^*}\exp(-M\beta)}{M+j}\,\mathrm{d}M = \int_j^\infty \frac{(t-j)^{k^*}\exp\{-\beta(t-j)\}}{t}\,\mathrm{d}t$$
$$= \exp(\beta j)\int_j^\infty \sum_{r=0}^{k^*}\binom{k^*}{r} t^{r-1} j^{k^*-r}(-1)^{k^*-r}\exp(-\beta t)\,\mathrm{d}t$$
$$= \exp(\beta j)\sum_{r=0}^{k^*}\binom{k^*}{r} j^{k^*-r}(-1)^{k^*-r}\int_j^\infty t^{r-1}\exp(-\beta t)\,\mathrm{d}t$$
$$= \exp(\beta j)(-j)^{k^*}\int_j^\infty \frac{\exp(-\beta t)}{t}\,\mathrm{d}t$$
$$+ \exp(\beta j)\sum_{r=1}^{k^*}\binom{k^*}{r}(-j)^{k^*-r}\int_j^\infty t^{r-1}\exp(-\beta t)\,\mathrm{d}t. \qquad (17)$$

The integral in the first term is the exponential integral, and the integral in the second term is the incomplete gamma function, both of which are available as standard library functions in most mathematical software, and hence can be evaluated much faster than by using numerical integration. Substituting expression (17) into equation (16), we optimize for β and restrict α to the integers since it is used in the sum. The probability of each configuration is proportional to expression (13), giving the discrete distribution.

Both of these methods above are shown for one parent–child group. However, we shall utilize all parent–child groups in a given scale to obtain smoothing parameters for that scale. This can be done since the marginal distributions have the same form in each parent–child group, but they need to be combined as shown below.

In the first scale of the data, there is only one parent–child group. For the second scale, there are four different parent–child groups. The following approach is applicable to either method of finding smoothing parameters, but we use the notation of the first method based on finding an M for notational simplicity. Denote $\mathcal{C}_{(1,1)}$ as the configuration in the first group, and let

$$P_{(1,1)} = P(\mathcal{C}_{(1,1)}|M)\,P(X_{2,(1,1)}, X_{2,(1,2)}, X_{2,(2,1)}, X_{2,(2,2)}|\mathcal{C}_{(1,1)}, X_{1,(1,1)}). \qquad (18)$$

Similarly define $P_{(1,2)}$, $P_{(2,1)}$ and $P_{(2,2)}$. To find an optimal M for the entire scale, the following sum would have to be maximized:

$$\sum_{\mathcal{C}_{(1,1)}\in\mathcal{C}}\sum_{\mathcal{C}_{(1,2)}\in\mathcal{C}}\sum_{\mathcal{C}_{(2,1)}\in\mathcal{C}}\sum_{\mathcal{C}_{(2,2)}\in\mathcal{C}} P_{(1,1)}P_{(1,2)}P_{(2,1)}P_{(2,2)}. \qquad (19)$$

This would take $12^{4^{2-1}}$ operations since there are 12 different configurations in each of the four parent–child groups, and 2 is the scale of data. However, expression (19) can be factorized as

$$\sum_{\mathcal{C}_{(1,1)}\in\mathcal{C}} P_{(1,1)} \sum_{\mathcal{C}_{(1,2)}\in\mathcal{C}} P_{(1,2)} \sum_{\mathcal{C}_{(2,1)}\in\mathcal{C}} P_{(2,1)} \sum_{\mathcal{C}_{(2,2)}\in\mathcal{C}} P_{(2,2)}, \qquad (20)$$

which takes only $4^{2-1}\times 12$ operations. More generally, if l is the scale of data, the number of operations reduces to $4^{l-1}\times 12$ instead of $12^{4^{l-1}}$. This is a substantial gain in computation efficiency when L can be as high as 10.

Instead of summing over all possible configurations for the entire scale of data, each parent–child group can be summed over configurations and then multiplied by all other parent–child groups. The logarithm of this product is used to optimize for a value of M or α and β across a scale of the data, i.e.

$$\sum_{z=1}^{4^{l-1}} \log\left\{ \sum_{\mathcal{C}_z \in \mathcal{C}_l} P(\mathcal{C}|M)\, P(X_1, X_2, X_3, X_4|\mathcal{C}, X) \right\},$$

where z represents the parent–child groups in scale l.

The two methods based on expression (12) and on expression (13) performed almost the same, but we prefer the second because of an increase in flexibility and stability of parameters. Using this method, we optimize for β whereas α can be 1 or 2 only. As indicated earlier, an integer value of α is essential whereas small values avoid occasional numerical instability due to rounding problems that are associated with large alternating terms in equation (17). To avoid overfitting the likelihood when optimizing for the smoothing parameters, we introduce a monotonicity constraint that restricts M or α/β to be decreasing as the scale becomes finer. Once the smoothing parameters have been chosen, regardless of which method is used to find them, a discrete distribution is obtained for the configuration given the data and smoothing parameters in each parent–child group. This will give the posterior mean of each ρ-parameter, and then intensity estimates for each pixel.

3. Asymptotic behaviour

To understand how well our method works, it is important to study theoretical convergence properties. Posterior consistency is a basic, but very important, theoretical property. If it were possible to expose the photon detector for an infinite amount of time, would the posterior distribution of the image concentrate near the true image? The longer that we expose the detector to the object, the larger in magnitude is the overall intensity. Therefore, the relative values of the intensities can be thought of as describing an image. Let $\bar{\boldsymbol{\lambda}}^0 = (\bar{\lambda}^0_{(j,k)}, j,k = 1, \ldots, 2^L)$ stand for the true value of the relative intensity parameters. We would require the relative intensities $\bar{\lambda}_{(j,k)} := \lambda_{(j,k)}/\lambda_{0,(1,1)}$ to converge to their true values, $\bar{\lambda}^0_{(j,k)}$. Note that, given the number of photons, $X_{0,(1,1)} = n$, the model is multinomial with parameters n and $\bar{\boldsymbol{\lambda}} = (\bar{\lambda}_{(j,k)}, j,k = 1, \ldots, 2^L)$ denoted by $\mathcal{M}(n; \bar{\boldsymbol{\lambda}})$. As $X_{0,(1,1)}$ is Poisson distributed with parameter $\lambda_{0,(1,1)}$, it follows that $X_{0,(1,1)\infty} \to \infty$ in probability as $\lambda_{0,(1,1)} \to \infty$. Thus it suffices to study asymptotics under the regime $n \to \infty$, disregarding the randomness in $X_{0,(1,1)}$.

We need some terminology to describe the models under consideration. In a parent–child group, if two block pixels share the same parent, we call them siblings, such as $\{1,2\}$, $\{1,3\}$, $\{1,4\}$, $\{2,3\}$, $\{2,4\}$ and $\{3,4\}$, referring to their relative positions. Among these, $\{1,4\}$ and $\{2,3\}$ are non-adjacent siblings and others are adjacent siblings. The common ancestor of two pixels (j,k) and (j',k') refers to the finest block pixel whose division leads ultimately to these two pixels and is denoted by $\mathrm{com}\{(j,k),(j'k')\}$. In general, any block pixel whose divisions lead to a pixel (j,k) is called an ancestor and (j,k) is called a descendant of it. The two children of this block pixel, whose divisions eventually lead to (j,k) and (j',k'), are called their separators and are denoted by $\mathrm{sep}\{(j,k)|(j',k')\}$ and $\mathrm{sep}\{(j',k')|(j,k)\}$. Clearly, these two block pixels are siblings.

In our setting, a structure can be defined by equality among neighbouring values of the split parameters at any level. The full model for $X_{(j,k)}$ is given by

$$\mathcal{M} = \{\mathcal{M}(n; \bar{\boldsymbol{\lambda}}) : \bar{\lambda}_{(j,k)} \text{ are not related to each other}\}.$$

Since the prior allows some ties between $\{\bar{\lambda}_{(j,k)}, j,k = 1, \ldots, 2^L\}$, various submodels, defined by the corresponding equalities, are also of interest. For $G_1, \ldots, G_r$ some groups of pixels, define

$$\mathcal{M}_{G_1,\ldots,G_r} = \{\mathcal{M}(n; \bar{\boldsymbol{\lambda}}) : \bar{\lambda}_{(j,k)} = \bar{\lambda}_{(j',k')} \text{ if } (j,k),(j',k') \in G_s, s = 1, \ldots, r\}.$$

Only certain collections of $G_1, \ldots, G_r$ need to be considered. If $(j,k), (j',k') \in G_s$ for some $s = 1, \ldots, r$, then any pixel (j'',k'') having either $\text{sep}\{(j,k)|(j',k')\}$ or $\text{sep}\{(j',k')|(j,k)\}$ as an ancestor will necessarily be in G_s. Further, if $\text{sep}\{(j,k)|(j',k')\}$ or $\text{sep}\{(j',k')|(j,k)\}$ are non-adjacent siblings, then G_s will also contain all pixels which have ancestor one of the two other siblings of $\text{sep}\{(j,k)|(j',k')\}$ and $\text{sep}\{(j',k')|(j,k)\}$ or will contain all descendants of $\text{com}\{(j,k),(j',k')\}$. For a model $\mathcal{M}_{G_1,\ldots,G_r}$, the number of equalities between relative intensity parameters $\Sigma_{s=1}^{r} \# G_s - r$ will be called the sharing index of the model. For two models $\mathcal{M}_{G_1,\ldots,G_r}$ and $\mathcal{M}_{G'_1,\ldots,G'_{r'}}$, we shall say that the latter is broader (or the former is narrower) if $r' \geqslant r$ and, for any $s'' = 1, \ldots, r'$, $G'_{s'} \subset G_s$ for some $s = 1, \ldots, r$. In this case, we shall write $\mathcal{M}_{G'_1,\ldots,G'_{r'}} \succeq \mathcal{M}_{G_1,\ldots,G_r}$. The narrowest model corresponds to the case when all $\bar{\lambda}_{(j,k)}$ are equal, whereas the broadest model is $\mathcal{M}$. The narrowest model which contains the true parameter $\bar{\boldsymbol{\lambda}}^0$ will be referred to as the true model and will be denoted by $\mathcal{M}_0$. Any model that is broader than the true model will be called a compatible model. Thus compatible models never contain an incorrect specification of ties between relative intensity parameters. However, some structure may be missed by such a model. The number of equalities that are missed by a compatible model $\mathcal{M}_{G_1,\ldots,G_r}$ is the difference between the sharing indices of $\mathcal{M}_0$ and $\mathcal{M}_{G_1,\ldots,G_r}$ and will be referred to as the redundancy of the model. In contrast, an artefact is formed by an incorrect forcing of ties by the model. Such models are incompatible.

We recall that the posterior distribution of $\bar{\boldsymbol{\lambda}}$ is consistent at $\bar{\boldsymbol{\lambda}}^0$ if, for any $\varepsilon > 0$, the posterior probability of $\mathcal{N}_\varepsilon(\bar{\boldsymbol{\lambda}}^0) = \{\bar{\boldsymbol{\lambda}} : \|\bar{\boldsymbol{\lambda}} - \bar{\boldsymbol{\lambda}}^0\| < \varepsilon\}$ converges to 1 almost surely as $n \to \infty$ under a distribution induced by $\bar{\boldsymbol{\lambda}}^0$.

Theorem 1. For the smoothing method based on the modified CRP, we have the following properties.

(a) The posterior distribution of $\bar{\boldsymbol{\lambda}}$ is consistent at $\bar{\boldsymbol{\lambda}}^0$.
(b) For any incompatible model $\mathcal{M}_{G_1,\ldots,G_r}$, $\Pi(\mathcal{M}_{G_1,\ldots,G_r}|X) \leqslant \exp(-cn)$ for some $c > 0$ almost surely for all sufficiently large n.
(c) For any compatible model $\mathcal{M}_{G_1,\ldots,G_r}$ that is different from the true model, $\Pi(\mathcal{M}_{G_1,\ldots,G_r}|X) = O_p(n^{-d/2})$, where d stands for the redundancy of $\mathcal{M}_{G_1,\ldots,G_r}$.

Proof. Recall that, conditionally on $X_{0,(1,1)} = n$, X follows $\mathcal{M}(n; \bar{\boldsymbol{\lambda}})$. To study consistency, we need to consider only the largest model $\mathcal{M}$ with the mixture prior governed by the modified CRP. According to Schwartz (1965), the posterior distribution is consistent if the prior assigns positive probability to every neighbourhood of $\bar{\boldsymbol{\lambda}}^0$ in the sense of the Kullback–Leibler divergence between the probability distributions $P(\mathbf{X}|\bar{\boldsymbol{\lambda}}^0)$ and $P(\mathbf{X}|\bar{\boldsymbol{\lambda}})$ corresponding to $\bar{\boldsymbol{\lambda}}^0$ and $\bar{\boldsymbol{\lambda}}$ respectively, and if the hypotheses $\bar{\boldsymbol{\lambda}} = \bar{\boldsymbol{\lambda}}^0$ against $\bar{\boldsymbol{\lambda}} \in \mathcal{N}_\varepsilon(\bar{\boldsymbol{\lambda}}^0)^c$ can be tested with exponentially small error probabilities. Now $\bar{\boldsymbol{\lambda}} = \bar{\boldsymbol{\lambda}}^0$ against $\bar{\boldsymbol{\lambda}} \in \mathcal{N}_\delta(\bar{\boldsymbol{\lambda}}^1)$, $\bar{\boldsymbol{\lambda}}^1 \neq \bar{\boldsymbol{\lambda}}^0$, can be tested with exponentially small error probabilities for sufficiently small $\delta > 0$ by using the fact that the Hellinger distance between $\mathcal{M}(n; \bar{\boldsymbol{\lambda}})$ and $\mathcal{M}(n; \bar{\boldsymbol{\lambda}}^0)$ is equivalent to the Euclidean distance $\|\bar{\boldsymbol{\lambda}} - \bar{\boldsymbol{\lambda}}^0\|$. Covering $\mathcal{N}_\varepsilon(\bar{\boldsymbol{\lambda}}^0)^c$ by balls of the type $\mathcal{N}_\delta(\bar{\boldsymbol{\lambda}}^1)$, $\bar{\boldsymbol{\lambda}}^1 \neq \bar{\boldsymbol{\lambda}}^0$, and using the compactness of the 4^L-dimensional unit simplex, it follows that the testing requirement is met by the family of multinomial distributions.

Now the Kullback–Leibler divergence number between $P(\mathbf{X}|\bar{\boldsymbol{\lambda}}^0)$ and $P(\mathbf{X}|\bar{\boldsymbol{\lambda}})$ is given by

$$E[\log\{P(X|\bar{\boldsymbol{\lambda}}_0)\} - \log\{P(X|\bar{\boldsymbol{\lambda}})\}] = \sum_{j=1}^{2^L} \sum_{k=1}^{2^L} \bar{\lambda}^0_{(j,k)} \log(\bar{\lambda}^0_{(j,k)} / \bar{\lambda}_{(j,k)}).$$

By a simple Taylor series expansion, this expression is bounded by a multiple of $\|\bar{\boldsymbol{\lambda}} - \bar{\boldsymbol{\lambda}}^0\|^2$. Thus the prior positivity condition holds provided that $\Pi(\|\bar{\boldsymbol{\lambda}} - \bar{\boldsymbol{\lambda}}^0\| < \delta) > 0$ for all $\delta > 0$. The mixture prior that is governed by the modified CRP can be written as a non-trivial convex combination

$a\Pi_1 + (1-a)\Pi_0$, where Π_1 is some prior distribution and Π_0 has positive density throughout. Thus $\Pi_0(\|\bar{\lambda} - \bar{\lambda}^0\| < \delta) > 0$ for all $\delta > 0$, which implies the required assertion. Consequently, by Schwartz (1965), $\Pi\{\mathcal{N}_\varepsilon(\bar{\lambda}^0)^c|\mathbf{X}\} \to 0$ almost surely under the true distribution, proving assertion (a). Indeed, it actually follows from Schwartz (1965) that $\Pi\{\mathcal{N}_\varepsilon(\bar{\lambda}^0)^c|\mathbf{X}\} \leqslant \exp(-nc)$ for some $c > 0$ almost surely for all sufficiently large n under the true distribution.

Now, to prove assertion (b), observe that in any incompatible model there is at least one incorrect equality between co-ordinates. Thus the distance between $\bar{\lambda}^0$ and the hyperplane, which defines the corresponding equality between co-ordinates, is positive. Since $\bar{\lambda}^0$ is fixed and there are only a finite number of such hyperplanes, it follows that, for some $\varepsilon > 0$, $\|\bar{\lambda} - \bar{\lambda}^0\| > \varepsilon$ for any $\bar{\lambda} \in \mathcal{M}_{G_1,\ldots,G_r}$. Thus $\mathcal{M}_{G_1,\ldots,G_r} \subset \mathcal{N}_\varepsilon(\bar{\lambda}^0)^c$, and hence assertion (b) follows the stronger version of assertion (a) that was established above.

Now it remains to prove assertion (c). Let $\mathcal{M}_{G_1,\ldots,G_r}$ be any compatible model. The number of free parameters in this model is given by

$$p = 4^L - 1 - \sum_{s=1}^{r} \#G_s + r.$$

Then, by Schwarz (1978), $\log\{\Pi(\mathcal{M}_{G_1,\ldots,G_r}|\mathbf{X})\}$ can be approximated by the Bayesian information criterion, namely

$$\log\{\Pi(\mathcal{M}_{G_1,\ldots,G_r}|\mathbf{X})\} = l(\hat{\bar{\lambda}}|\mathcal{M}_{G_1,\ldots,G_r}) - \frac{p}{2}\log(n) + o_p(1),$$

where $l(\hat{\bar{\lambda}}|\mathcal{M}_{G_1,\ldots,G_r})$ stands for the log-likelihood under the model $\mathcal{M}_{G_1,\ldots,G_r}$ at the corresponding maximum likelihood estimate. If p_0 stands for the total number of free parameters in the true model $\mathcal{M}_0$, then also

$$\log\{\Pi(\mathcal{M}_0|\mathbf{X})\} = l(\hat{\bar{\lambda}}|\mathcal{M}_0) - \frac{p_0}{2}\log(n) + o_p(1).$$

Consequently, with $d = p - p_0$ standing for the redundancy of $\mathcal{M}_{G_1,\ldots,G_r}$,

$$\log\left\{\frac{\Pi(\mathcal{M}_{G_1,\ldots,G_r}|\mathbf{X})}{\Pi(\mathcal{M}_0|\mathbf{X})}\right\} = \{l(\hat{\bar{\lambda}}|\mathcal{M}_0) - l(\hat{\bar{\lambda}}|\mathcal{M}_{G_1,\ldots,G_r})\} - \frac{d}{2}\log(n) + o_p(1). \tag{21}$$

The term in brackets on the right-hand side stands for the log-likelihood ratio test statistics for testing $\bar{\lambda} \in \mathcal{M}_0$ against $\bar{\lambda} \in \mathcal{M}_{G_1,\ldots,G_r}$. By Wilks's theorem,

$$2\{l(\hat{\bar{\lambda}}|\mathcal{M}_0) - l(\hat{\bar{\lambda}}|\mathcal{M}_{G_1,\ldots,G_r})\} \xrightarrow[\mathrm{d}]{} \chi^2_d.$$

Hence, in particular, the term in brackets on the right-hand side of equation (21) is $O_p(1)$. Therefore,

$$\frac{\Pi(\mathcal{M}_{G_1,\ldots,G_r}|\mathbf{X})}{\Pi(\mathcal{M}_0|\mathbf{X})} = O_p(n^{-d/2}) = o_p(1),$$

because $d > 0$. Since, by assertion (b), $\Pi(\mathcal{M}_{G'_1,\ldots,G'_{r'}}|\mathbf{X}) = o_p(1)$ for any incompatible model $\mathcal{M}_{G'_1,\ldots,G'_{r'}}$, it follows that $\{1 + o_p(1)\}\,\Pi(\mathcal{M}_0|\mathbf{X}) + o_p(1) = 1$, and hence $\Pi(\mathcal{M}_0|\mathbf{X}) = 1 + o_p(1)$. This implies that $\Pi(\mathcal{M}_{G_1,\ldots,G_r}|\mathbf{X}) = O_p(n^{-d/2})$ for any compatible model with redundancy d, completing the proof of assertion (c).

4. Empirical results and simulations

Our method of the proposed reconstruction of images with photon counts shows theoretical

promise, but empirical evidence is essential to verify actual performance. In this section, the Bayesian CRP method is compared with platelets (Willett and Nowak, 2003), the two-dimensional Bayesian multiscale model (Nowak and Kolaczyk, 2000) called MAP, a translation invariant Haar wavelet smoothing method, called TIPSH, using corrected thresholds for Poisson data (Kolaczyk, 1997) and the multiscale complexity regularization method MSCR that was introduced in Kolaczyk and Nowak (2004). These five methods will be compared by using simulated test images to compare each method's error with the true image. All methodologies use the fast translation invariant process that was proposed in Willett and Nowak (2004), except for platelets for which the technique is inapplicable.

An important distinction between images is the photon flux that is observed in an image. Poisson-based methods are much more applicable to the photon-limited images; thus simulations are conducted for low intensity test images. Another characteristic that is common to these astronomical images is background noise. There is typically an almost constant background photon source across the entire image. The simulated images will incorporate a constant amount of background noise to make the test images more realistic. Two distance metrics are used to compare the images—Baddeley's delta metric BDM for greyscale images (Wilson *et al.*, 1997) and the mean absolute difference MAD of the relative intensities. When viewing these images, it is common to view them as greyscale images, and so we employ Baddeley's delta metric that was specifically designed to compare these types of image. Since the estimates of each methodology

Table 1. Simulation results for the Saturn image for maximum intensity 0.3 and background intensity of 0.01†

Method	*MAD* ($\times 10^{-6}$)	*BDM* ($\times 10^{-4}$)	*Time (s)*
Bayesian CRP	2.40 (0.01)	4.66 (0.07)	9
Platelets (20 iterations)	2.23 (0.01)	5.77 (0.14)	38
Bayesian multiscale MAP	3.10 (0.01)	—	<1
TIPSH	2.61 (0.01)	—	<1
Multiscale complexity	2.83 (0.01)	—	<1

†The mean absolute difference MAD is given for all methods with 50 simulations. Baddeley's delta metric is given for platelets and the Bayesian CRP method with 50 simulations. Standard errors are given in parentheses.

Table 2. Simulation results for the Shepp–Logan phantom for maximum intensity 5 and background intensity 1†

Method	*MAD* ($\times 10^{-7}$)	*BDM* ($\times 10^{-4}$)	*Time (s)*
Bayesian CRP	1.19 (0.004)	2.88 (0.19)	40
Platelets (20 iterations)	1.27 (0.005)	2.91 (0.16)	1460
Bayesian multiscale MAP	2.63 (0.003)	—	2
TIPSH	2.94 (0.004)	—	1
Multiscale complexity	1.45 (0.005)	—	2

†The mean absolute difference MAD is given for all methods with 50 simulations using a 512×512 image. Baddeley's delta metric BDM is given for platelets and the Bayesian CRP method with 50 simulations using a 256×256 image owing to time constraints of the BDM. Standard errors are given in parentheses.

Fig. 1. Simulation images of Saturn with maximum intensity 0.3 and constant level of background 0.01 (the original image has both the regular intensities and the background intensity added to it; the image is a test image that is available on line and was also used in other image processing references such as Willett (2006)): (a) original image; (b) Poisson observations; (c) MSCR; (d) platelet; (e) translation invariant Haar; (f) Bayesian MAP; (g) Bayesian CRP

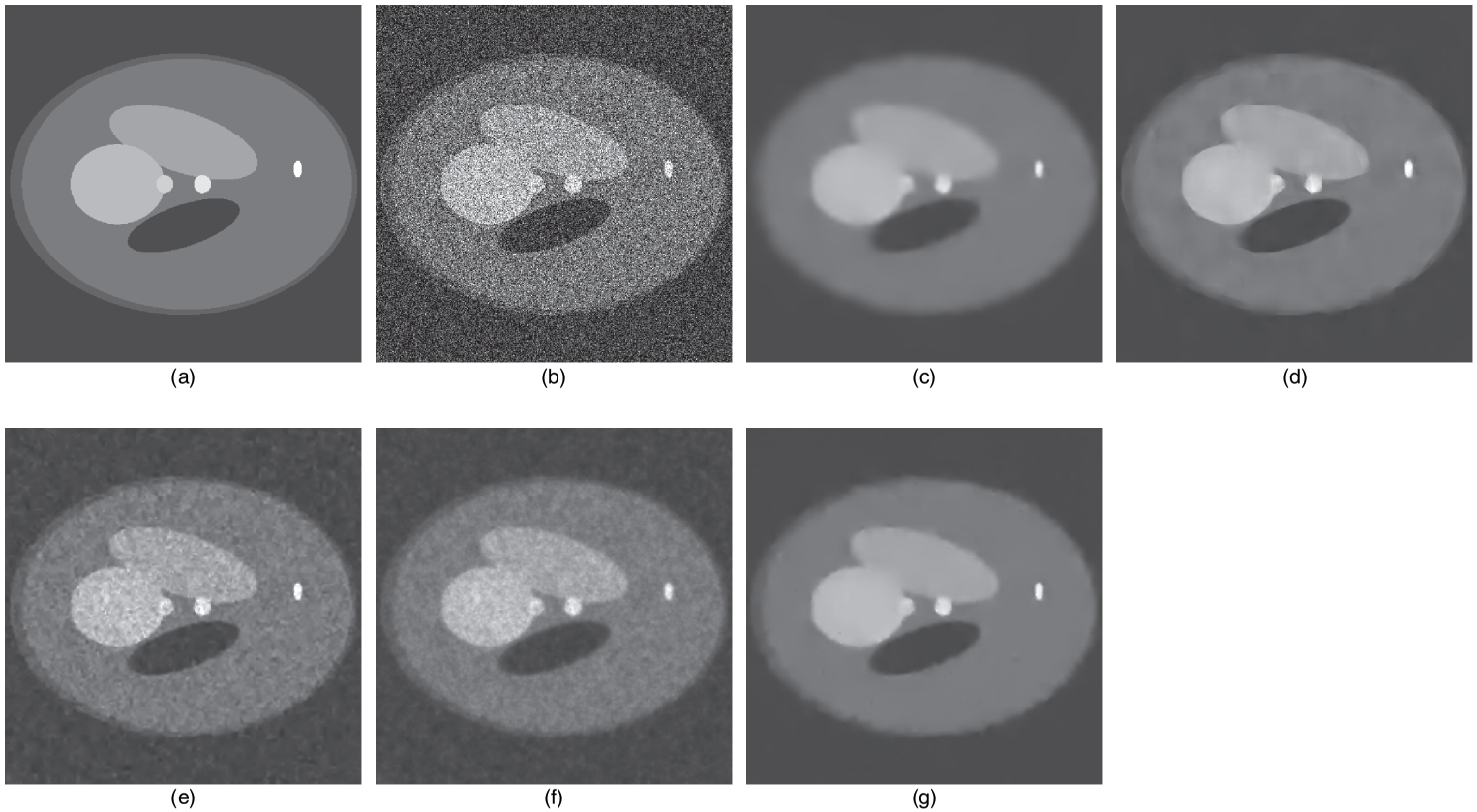


Fig. 2. Simulation images of the Shepp–Logan phantom with maximum intensity 5 and constant level of background 1 (the original image has both the regular intensities and the background intensity added to it; this image is a test image that is available in MATLAB and is commonly used in many image restoration techniques): (a) original; (b) Poisson observations; (c) MSCR; (d) platelet; (e) translation invariant Haar; (f) Bayesian MAP; (g) Bayesian CRP

allow for continuous values of intensities, to define the delta metric intensity values are grouped into 256 discrete bins standing for greyscales. Computing speed of one replication of each methodology is also compared in the tables.

An image of Saturn that is 256×256 pixels is the first test image. This image was also used in Willett (2006). We create an extremely photon-limited image with a maximum intensity level set at 0.3, and a constant background of 0.01 added uniformly across the entire image. Poisson samples based on the images emulate what would be seen in practice if looking at an astronomical X-ray image, or any image with Poisson noise. More specifically, the intensities are created from the original image by scaling back the maximum intensity pixel to 0.3 and then adding 0.01 to each pixel. This matrix of intensities is then used to create a Poisson sample. As a second comparison, the Shepp–Logan phantom image of size 512×512 is used with a maximum intensity of 5 and a constant background intensity of 1. This image is an imitation of a brain scan with the smaller ellipses representing features within the brain. It can be directly generated by the image processing software that is available in MATLAB and is a very popular test image.

Tables 1 and 2 compare methods by using the two metrics for the two images in Figs 1 and 2. All simulations were run using MATLAB on a personal computer with 6 Gbytes of random-access memory and an Intel Pentium Core i7 processor. 50 simulations were run using the test images computing the MAD and BDM for each replication. When comparing the images with BDM, only the top two methodologies were compared since the calculation of BDM for each replication and methodology required a large amount of time. Thus, only platelets and the Bayesian CRP are compared with this metric. The first test image of Saturn always uses a 256×256 image; however, when using the second test image, the sizes differ depending on the metric. More specifically, when using MAD, the second test image was 512×512; however, when using BDM, the second test image was generated to be 256×256 to save time in calculating BDM.

Table 1 and Fig. 1 represent the extremely photon-limited type of observation. The Bayesian CRP model performs quite well in this scenario. It is a clear improvement over the earlier Bayesian MAP method and in fact outperforms all other methods except for platelets when using MAD. When comparing platelets and the CRP method by using the more sophisticated BDM, the CRP method outperforms platelets also. The CRP method is slower than the MAP, TIPSH and MSCR methods but is considerably faster than platelets. For the CRP method, a substantial portion of time is spent estimating smoothing parameters. All methods other than platelets give only positive estimates for pixel scale intensities. However, platelets may sometimes give negative estimates, which could be bothersome.

As the intensity increases, all the methods obtain better results with respect to the quality of the image. Except for platelets, it can be shown that the computing time of every method does not increase with the total number of photons. However, the time does increase for all methods with a larger sized image as in Fig. 2. The results in Table 2 are slightly different from the results in Table 1. The CRP and platelet methodologies are the top performers when looking at MAD with the CRP method outperforming platelets. When using BDM, the CRP method appears to be slightly better than platelets.

When implementing all the methods other than the Bayesian CRP, there is a smoothing parameter that must be chosen. The Bayesian CRP method is completely data driven in that it chooses smoothing parameters on the basis of the noisy observed image only. All other smoothing parameters were chosen to be the default value given in downloaded software, as was the case for platelets and multiscale complexity methods, or based on the references describing them, as was the case for the Bayesian MAP and Haar methodology. It should also be noted

that, with more cycle spins, the platelet methodology may be able to improve the estimation error but will also take longer to compute.

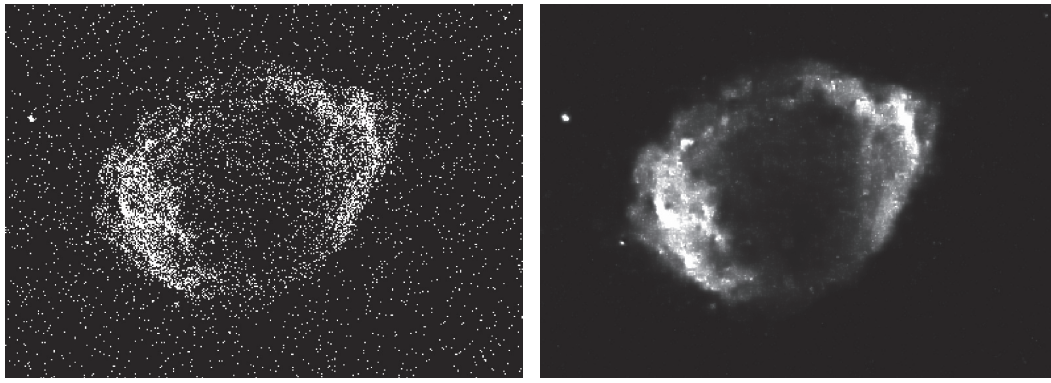
5. Astronomical images

X-ray astronomy is a very well-researched field (Trumper and Hasinger, 2008). The methodology that is described in this paper is particularly well suited in this type of imaging since the Poisson distribution is a natural model for noisy photon-limited X-ray images. Each photon can also have a different energy level but, if only the counts are obtained, then a two-dimensional Poisson model is appropriate. With imaging devices like the Chandra X-ray Observatory, X-ray images of high energy objects in space have been taken. Noise creeps in the images due to instrument error and photon scattering as well as having such a short exposure time.

One particular interest in X-ray astronomy is observing supernova remnants (Gull and Daniell, 1978). A supernova remnant is comprised of the material that is ejected after a star explodes. These remnants are characterized by their expanding shock. Obtaining a good representation of these remnants in terms of X-ray energy allows astronomers to determine their characteristics that provide details of their origin. Two real images of supernova remnants were reconstructed with this methodology. One is called G1.9 (Reynolds *et al*., 2008), which is the youngest supernova remnant discovered thus far, and the other is the famous Kepler supernova remnant.

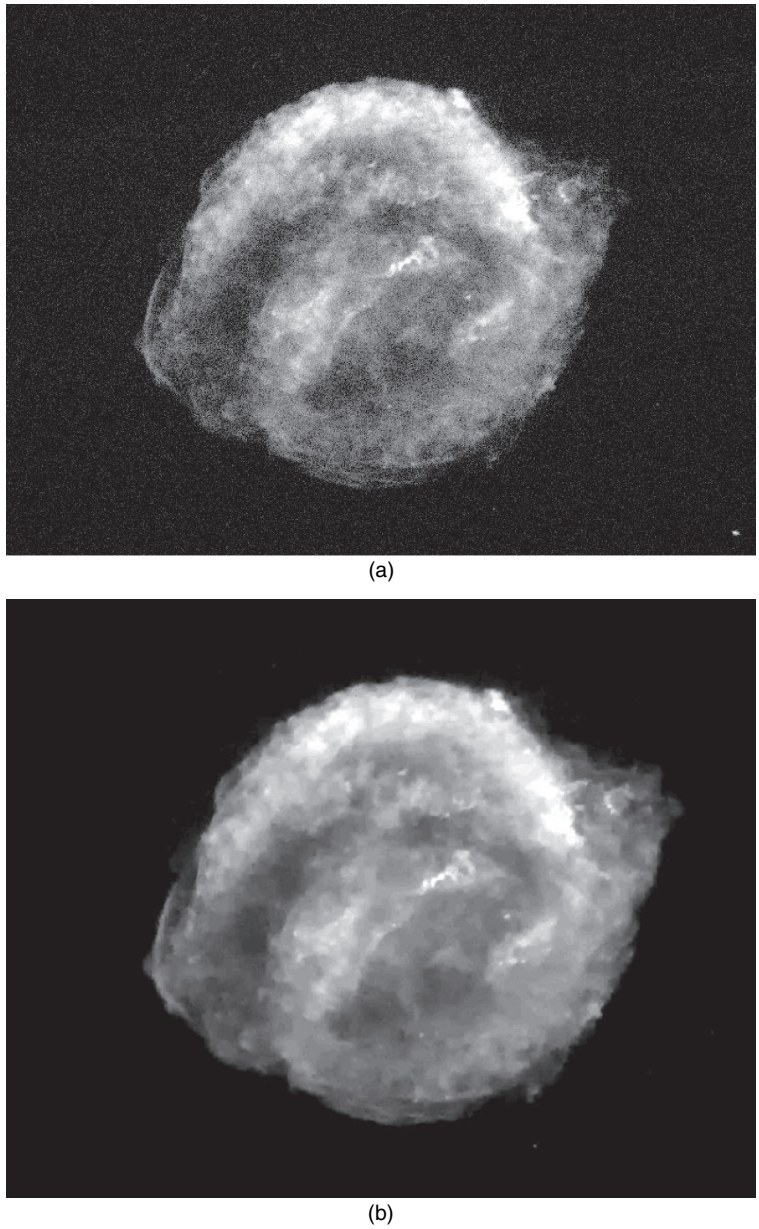
An image of G1.9 is shown in Fig. 3 and represents an extremely photon-limited image. The observed image in Fig. 3(a) has a maximum photon count of 8 in a pixel and an average of 0.0473 photons per pixel. The smoothed version in Fig. 3(b) is the result of applying the Bayesian CRP method to reconstruct the image. This image has a much more defined shape and also minimizes the background noise.

The Kepler supernova remnant is shown in Fig. 4. The observed image has an average photon count of 2.172 per pixel with a maximum count of 245 photons in a pixel. Even though the maximum value is large, most pixel counts are much smaller. The noise appears to be considerably reduced after processing. This is very noticeable towards the outside edges of the object.



(a) (b)

Fig. 3. X-ray image of G1.9 from the Chandra X-ray observatory (both images have the same contrast to give an idea of how they related to one another; these images can be obtained on line through the Chandra database; the mean photon count per pixel is 0.0473 with a maximum count of 8): (a) observed image; (b) smoothed image



(a)

(b)

Fig. 4. X-ray image of the Kepler supernova remnant from the Chandra X-ray Observatory (both images have the same contrast to give an idea of how they related to one another; these images can be obtained on line through the Chandra database; other images are available for this supernova remnant that have much more exposure; however, this example provides a case where the photon count is more limited; the mean photon count per pixel is 2.172 with a maximum count of 245): (a) observed image; (b) smoothed image

There are many other images in X-ray astronomy that could benefit from using the methodology proposed instead of familiar but less efficient older methodologies. Typically images are made of 1024×1024 pixels if the full detector is used, especially for the Chandra X-ray Observatory. These techniques may be employed for any image that records photon counts in pixels. However, it is more suitable for images with relatively few photons counts per pixel.

6. Conclusions

The Bayesian CRP method that was introduced in this paper is a promising method when smoothing an image with Poisson noise and is particularly well suited for images with small counts of photons in each pixel, such as those in astronomical imaging. The CRP allows any group to be made from the four possible children providing structure in the smoothed image as long as the children are adjacent. The posterior mean of these estimates is obtained analytically after optimizing for smoothing parameters. Since these are piecewise constant estimates, a fast translation invariant algorithm is used to obtain the smoothed image, avoiding the need for cycle spinning as well as any type of Markov chain Monte Carlo algorithm. Thus, the Bayesian CRP method is extremely fast.

Theoretical consistency properties are given in Section 3 and provide another distinct improvement over the Bayesian multiscale model in Nowak and Kolaczyk (2000), namely model selection consistency. In the context of images, this means that artefacts will tend to disappear whereas real structures will show up in posterior sampling as photon counts approach ∞.

In empirical simulations, the CRP method was compared with four of the leading methods for smoothing images with Poisson noise. Unlike other methods, the choice of the smoothing parameters in the CRP method is completely data driven. The CRP method outperforms MAP methodology, Haar wavelets with corrected Poisson thresholds and multiscale complexity regularization in terms of the mean absolute deviation error metric in a variety of simulations. The CRP method outperforms the leading platelet methodology in terms of Baddeley's delta metric for greyscales and also in terms of the mean absolute deviation metric for the Shepp–Logan phantom test image. The computing time for the platelet method can be prohibitively long, and thus the methodology that was introduced in this paper is a very useful competitor when smoothing photon-limited images.

Acknowledgements

The code for each of these methods, except the Bayesian CRP, was obtained from Dr Rebecca Willett or downloaded from her Web site. The supernova image was obtained from Dr Stephen Reynolds at North Carolina State University and can also be found on the Chandra Observatory Web site (`http://chandra.harvard.edu`). The suggestions of the referees and the Associate Editor significantly improved the presentation of the paper. In particular, the authors thank a referee for suggesting the use of Baddeley's delta metric for comparison. This research was partially supported by National Science Foundation grant DMS-034911.

Appendix A: Chinese restaurant process

The CRP is a sampling scheme that is very closely related to the Dirichlet process that is used in nonparametric Bayesian statistics. It is a method of randomly assigning objects to groups or a discrete time process that partitions integers $(1, \ldots, N)$ at time N into separate groups. A more thorough explanation of the CRP can be obtained in Pitman (1995) and Teh *et al.* (2006). This sampling scheme is described by the following analogy.

Suppose that there are an infinite number of tables with infinite capacities in a restaurant. Each customer that comes in will sit at a table. When customers arrive, they will either sit at a new table or sit at a table with other customers. The probability that a new customer will sit at an existing table is proportional to the number of customers already sitting at the table. However, there is always some probability that a new customer will sit at a new table. At time $N + 1$, the following probability distribution governs the choice of a table:

$$P(\text{joins group } k) = \frac{n_k}{M+N},$$
$$P(\text{forms a new group}) = \frac{M}{M+N},$$

where n_k is the number of integers in group k at time N. The parameter M decides how likely a new group is formed. Larger values of M will tend to make more groups. When considered in a model for creating ties in continuous variables, the CRP allows the formation of clusters.

The Dirichlet process prior is related to the CRP as follows. Suppose that $X_1, X_2, \ldots, X_n | P \sim P$ and $P \sim \mathrm{DP}(M, G)$, where $\mathrm{DP}(M, G)$ stands for the Dirichlet process prior with G as the centre measure and M as the total mass or precision parameter; see Ferguson (1973). The following properties of a Dirichlet process prior are well known:

$$X_1 \sim G,$$
$$X_2 | P, X_1 \sim P,$$
$$P | X_1 \sim \mathrm{DP}\left(M+1, \frac{M}{M+1}G + \frac{1}{M+1}\delta_{X_1}\right),$$
$$X_2 | X_1 \sim \frac{M}{M+1}G + \frac{1}{M+1}\delta_{X_1}.$$

Thus, with probability $1/(M+1)$, X_2 will replicate X_1. With probability $M/(M+1)$, X_2 will be a new draw from distribution G. Continuing in a similar manner

$$X_{N+1} | X_1, \ldots X_N \sim \frac{M}{M+N}G + \sum_{i=1}^{N} \frac{1}{M+N}\delta_{X_i}.$$

The presence of point masses leads to ties, or a table sharing in a CRP, which tracks the labels of observations in a Dirichlet process.

References

Baddeley, A. (1992) An error metric for binary images. In *Robust Computer Vision: Quality of Vision Algorithms*, pp. 59–78. Karlsruhe: Wichmann.

Coifman, R. R. and Donoho, D. L. (1995) Translation-invariant de-noising. In *Wavelets and Statistics* (eds A. Antoniadis and G. Oppenheim), pp. 125–150. New York: Springer.

Donoho, D. (1999) Wedgelets: nearly minimax estimates of edges. *Ann. Statist.*, **27**, 859–897.

Esch, D. N., Connors, A., Karovska, M. and van Dyk, D. A. (2004) An image restoration technique with error estimates. *Astrophys. J.*, **610**, 1213–1227.

Ferguson, T. S. (1973) A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, **1**, 209–230.

Gull, S. and Daniell, G. (1978) Image reconstruction from incomplete and noisy data. *Nature*, **272**, 686–690.

Kolaczyk, E. D. (1997) Nonparametric estmation of gamma-ray burst intensities using Haar wavelets. *Astrophys. J.*, **483**, 340–349.

Kolaczyk, E. D. (1998) Wavelet shrinkage estimation of certain Poisson intensity signals using corrected thresholds. *Statist. Sin.*, **9**, 119–135.

Kolaczyk, E. D. (1999) Bayesian multiscale models for Poisson processes. *J. Am. Statist. Ass.*, **94**, 920–933.

Kolaczyk, E. D. and Nowak, R. D. (2004) Multiscale likelihood analysis and complexity penalized estimation. *Ann. Statist.*, **32**, 500–527.

Nowak, R. D. and Kolaczyk, E. D. (2000) A statistical multiscale framework for Poisson inverse problems. *IEEE Trans. Inform. Theor.*, **46**, 1811–1825.

Pitman, J. (1995) Exchangeable and partially exchangeable random partitions. *Probab. Theor. Reltd Flds*, **102**, 145–158.

Reynolds, S., Borkowski, K., Green, D., Hwang, U., Harrus, I. and Petre, R. (2008) The youngest galactic supernova remnant: G1.9+0.3. *Astrophys. J.*, **680**, 41–44.

Schwartz, L. (1965) On Bayes procedures. *Z. Wahrsch. Ver. Geb.*, **4**, 10–26.

Schwarz, G. (1978) Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.

Starck, J. and Murtagh, F. (2006) *Astronomical Image and Data Analysis*, 2nd edn. New York: Springer.

Teh, Y. W., Jordan, M. I., Beal, M. J. and Blei, D. M. (2006) Hierarchical Dirichlet processes. *J. Am. Statist. Ass.*, **101**, 1566–1581.

Trumper, J. and Hasinger, G. (2008) *The Universe in X-rays*. New York: Springer.

Willett, R. (2006) Multiscale analysis of photon-limited astronomical images. *4th Conf. Statistical Challenges in Modern Astronomy, State College, June 12th–15th*.

Willett, R. M. and Nowak, R. D. (2003) Platelets: a multiscale approach for recovering edges and surfaces in photon-limited medical imaging. *IEEE Trans. Med. Imgng*, **22**, 332–350.

Willett, R. and Nowak, R. (2004) Fast multiresolution photon-limited image reconstruction. In *Proc. Int. Symp. Biomedical Imaging, Arlington, Apr. 15th–18th*. New York: Institute of Electrical and Electronics Engineers.

Wilson, D. L., Baddeley, A. J. and Owens, R. A. (1997) A new metric for grey-scale image comparison. *Int. J. Comput. Visn*, **24**, 5–17.