SIAM J. IMAGING SCIENCES
Vol. 3, No. 3, pp. 619–645

Multiscale Photon-Limited Spectral Image Reconstruction*

Kalyani Krishnamurthy†, Maxim Raginsky†, and Rebecca Willett†

Abstract. This paper studies photon-limited spectral intensity estimation and proposes a spatially and spectrally adaptive, nonparametric method for estimating spectral intensities from Poisson observations. Specifically, our method searches through estimates defined over a family of recursive dyadic partitions in both the spatial and spectral domains, and finds the one that maximizes a penalized log likelihood criterion. The key feature of this approach is that the partition cells are *anisotropic* across the spatial and spectral dimensions, so that the method adapts to varying degrees of spatial and spectral smoothness, even when the respective degrees of smoothness are not known a priori. The proposed approach is based on the key insight that spatial boundaries and singularities exist in the same locations in every spectral band, even though the contrast or perceptibility of these features may be very low in some bands. The incorporation of this model into the reconstruction results in significant performance gains. Furthermore, for spectral intensities that belong to the anisotropic Hölder–Besov function class, the proposed approach is shown to be near-minimax optimal. The upper bounds on the risk function, which is the expected squared Hellinger distance between the true intensity and the estimate obtained using the proposed approach, matches the best possible lower bound up to a log factor for certain degrees of spatial and spectral smoothness. Experiments conducted on realistic data sets show that the proposed method can reconstruct the spatial and the spectral inhomogeneities very well even when the observations are extremely photon-limited (i.e., less than 0.1 photon per voxel).

Key words. spectral imaging, Poisson intensity estimation, complexity regularization, wavelets, photon-limited imaging

AMS subject classifications. 68U10, 26B35, 30H25, 47A52

DOI. 10.1137/090756259

1. Spectral Poisson intensity estimation.

Spectral images consist of a spatial map of intensity variation across a large number of spectral bands or wavelengths; alternatively, they can be thought of as a measurement of the spectrum of light transmitted or reflected from each spatial location in a scene. Because spectral signatures are unique for every chemical element, observing these spectra at a high spatial and spectral resolution provides information about the material properties of the scene with much more accuracy than is possible with conventional three-color images. Spectral imaging is used in a variety of applications, including remote sensing, astronomical imaging, and fluorescence microscopy [35, 32, 43]. In the remote sensing of forest covers, for example, the spectral information helps to identify different types of vegetation and their chlorophyll content, which leads to better understanding of the forest

*Received by the editors April 16, 2009; accepted for publication (in revised form) June 3, 2010; published electronically September 29, 2010. This work was supported by NSF career award CCF-06-43947, DARPA grant HR0011-07-1-003, and AFRL grant FA8650-07-D-1221.
http://www.siam.org/journals/siims/3-3/75625.html
†Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708 (kk63@duke.edu, m.raginsky@duke.edu, willett@duke.edu).

ecosystem [35]. Similarly, in fluorescence microscopy, the spectra help to identify unwanted emissions due to background, autofluorescence, and other contaminants [43].

Although spectral imaging has the potential to be very useful, its use poses significant challenges. For instance, the number of photons hitting the detector might be very low (this is referred to as the *photon-limited regime*) because of either a weak source or a large distance between the source and the detector. Even if the source is not weak, the observed photon counts are binned by the imaging system according to their spatial locations and wavelengths. This means that increasing the spatial or spectral resolution of an imaging system decreases the number of photons detected in each spatiospectral bin. Thus, low photon counts always impact the quality of spectral images, either as photon noise or resolution limits. In addition, the geometry of imaging systems may introduce further distortions of the scene (either intentionally or unintentionally). For instance, lenses in optical systems introduce spatial blurring artifacts, and innovative new spectral imaging systems [19, 52, 53] measure pseudorandom projections of the scene to exploit recent theoretical developments in compressed sensing [5, 6, 13]. These indirect measurements of a scene result in challenging photon-limited inverse problems.

In this paper, we describe and analyze an approach to estimating a spectral intensity from either indirect or direct photon-limited measurements. In particular, we assume that the observations follow a spatially and spectrally inhomogeneous Poisson distribution, and we make the following two contributions: (a) develop a spatially and spectrally *adaptive* multiscale algorithm for recovering spectral image intensities from Poisson observations, and (b) provide a theoretical characterization of the proposed approach and prove that it is *near-minimax optimal*. The work presented here is an extension of our previous work [29], in which we proposed an algorithm for reconstructing spectral solar flare intensities from photon-limited observations. The previous work was focused on a specific model of spectra that was characterized by a wide dynamic range. This paper proposes and analyzes a much more broadly applicable extension and provides experimental results comparing the proposed approach with other methods described in the literature and demonstrating its effectiveness. We also present new theoretical upper and lower bounds on performance, which prove that the proposed algorithm is near-minimax optimal over a broad class of spectral images.

The remainder of the paper is organized as follows. In section 1.1, we provide a mathematical formulation of the problem of interest. This is followed by a review of the existing literature on Poisson intensity estimation and inverse problems in section 1.2. We describe the complexity-penalized likelihood estimation algorithm for denoising in section 2 and discuss the near-optimality of this approach and its computational complexity in sections 3 and 4, respectively. In section 5 we extend the denoising algorithm using an expectation-maximization (EM) framework to solve inverse problems. We demonstrate the effectiveness of the proposed approach in performing denoising and image reconstruction with numerical experiments in sections 6.1 and 6.2, respectively, followed by a discussion in section 7. Proofs of the main theorems are relegated to the appendices.

1.1. Problem formulation. Let f be the true spatially and spectrally varying spectral image on $[0,1]^3$, where the first two dimensions correspond to the spatial locations and the third dimension corresponds to the spectral bands. We let x_1, x_2, and λ denote the two spatial

and one spectral arguments of f, respectively. Let $\boldsymbol{f}$ denote a sampled version of f, where

$$\boldsymbol{f}_{i_1,i_2,i_3} = \int_{i_1/N_1}^{(i_1+1)/N_1} \int_{i_2/N_2}^{(i_2+1)/N_2} \int_{i_3/M}^{(i_3+1)/M} f(x_1, x_2, \lambda) d\lambda \, dx_2 \, dx_1 \tag{1.1}$$

for $i_1 = 0, 1, \ldots, N_1 - 1$, $i_2 = 0, 1, \ldots, N_2 - 1$, and $i_3 = 0, 1, \ldots, M - 1$. At the detector, we observe noisy and distorted photon counts which are given by

$$\boldsymbol{y} \sim \text{Poisson}(A\boldsymbol{f}), \tag{1.2}$$

where A corresponds to the distortion induced by the measurement system and "Poisson" denotes the independent observations of an inhomogeneous Poisson process with intensity $A\boldsymbol{f}$ on a grid of size $N_x \times N_y \times M$. The methods described in this paper can easily be generalized to the case $N_x \neq N_y$, but carrying this notation throughout our proof makes it less transparent. We thus assume that $N_x = N_y = N$, for simplicity of presentation. Let n denote the total number of observed events: $n \triangleq \sum_{i_1,i_2,i_3} \boldsymbol{y}_{i_1,i_2,i_3}$.

Our goal is to estimate $\boldsymbol{f}$ from $\boldsymbol{y}$ as accurately as possible by exploiting *anisotropic* correlations in the spectral and spatial dimensions. The inference methods and associated theoretical properties described below assume that f is *piecewise smooth*, meaning that spatially it is composed of smooth surfaces separated by smooth boundaries in the two spatial dimensions (and hence can be modeled as a member of a piecewise Hölder function class [28]), and that each spectrum in f varies smoothly as a function of wavelength except for a finite number of singularities and discontinuities (and hence can be modeled as a member of a Besov function class [9]). Such a function class can model several complex spectral intensities like the flare spectral intensities [32] and thus has wide applicability.

We thus assume that f belongs to an anisotropic Hölder–Besov function class (formally defined in detail in section 3) and that $0 < C_\ell \leq f \leq C_u < \infty$. As discussed in section 3, this anisotropy will allow us to model spectral images that exhibit different degrees of smoothness and irregularity in spatial and spectral domains, and precludes using isotropic basis functions (such as three-dimensional wavelets) without sacrificing performance.

In this paper, we will use the conventional $\mathcal{O}$ notation to specify computational complexity, and the symbols $\preceq$ and $\asymp$ to describe the error rates, where $f_n \preceq g_n$ means there exists some $C > 0$ such that the sequences f_n and g_n satisfy $f_n \leq C g_n$ for n sufficiently large and $f_n \asymp g_n$ means there exist some $C_1, C_2 > 0$ such that $C_1 g_n \leq f_n \leq C_2 g_n$ for n sufficiently large.

1.2. Background review. Many researchers have studied multiresolution methods to perform intensity estimation from Poisson data, because of the ability of those methods to capture the inhomogeneities in the data; see [38, 22, 40, 11, 10, 1, 4, 51]. Preserving discontinuities is critical because they potentially convey important information about the signal or image under observation. The key challenge in Poisson intensity estimation problems is that the mean and the variance of the observed counts are the same. As a result, the conventional wavelet-based approaches like hard thresholding [14] and soft thresholding [12], originally designed to denoise Gaussian data, will yield suboptimal results when applied to Poisson data with low intensities.

Variance-stabilizing transforms (VSTs), such as the Anscombe transform [16] and the Haar–Fisz [17, 18] transform, are widely used to address this issue and to approximate the

Poisson observations by Gaussian random variables [10, 21]. Jansen proposes a wavelet-based Poisson estimation method based on data-adaptive VSTs and Bayesian priors on the stabilized coefficients [20]. However, as pointed out in [26, 46], such approximations are inaccurate when the observed number of photons per pixel or voxel is very low, and they tend to oversmooth the resulting estimate. In a more recent work, Zhang, Fadili, and Starck [55] propose a multiscale variance-stabilizing transform (MSVST), which applies a VST to the empirical wavelet, ridgelet, or curvelet transform coefficients. However, theoretical analysis of this approach is not available, and it is not clear how to extend the MSVST to Poisson inverse problems.

Several authors have investigated signal and image estimation methods specifically designed for Poisson noise without the need for VSTs. For example, Kolaczyk proposes scale-dependent corrected Haar wavelet thresholds for Poisson data [23, 26]. Bayesian approaches offer an elegant way of incorporating prior knowledge into the estimation process to improve the performance. Kolaczyk [22] and Timmermann and Nowak [44] propose a multiscale approach using the unnormalized Haar wavelet transform in conjunction with Bayesian methods to perform denoising of Poisson data. In [38, 37], Nowak and Kolaczyk extend the multiscale and Bayesian approaches mentioned above to Poisson inverse problems, where they present an MAP (maximum a posteriori) estimation algorithm in an expectation-maximization (EM) framework to reconstruct two-dimensional intensities. While Bayesian methods are optimal when the prior distribution accurately reflects the true distribution underlying the phenomenon being observed (i.e., f), it is not clear how the performance of these approaches changes with inaccurate approximations of the true prior.

In their seminal paper [25], Kolaczyk and Nowak present a multiscale framework for likelihoods similar to the multiresolution analysis on wavelets and propose a denoising algorithm based on the complexity-penalized likelihood estimation (CPLE). Compared to the Bayesian methods discussed above, the CPLE algorithm has fewer tuning parameters (or, alternately, avoids complex Markov chain Monte Carlo (MCMC) methods required when replacing tuning parameters with Bayesian hyperparameters) and is also minimax optimal over a wide range of isotropic likelihood models. There are variants of the CPLE method that depend upon the nature of the image or signal being denoised [51, 24, 50].

Though there is a rich literature on one- and two-dimensional Poisson intensity estimation problems, there are not many algorithms developed for photon-limited Poisson spectral intensity estimation, which is the focus of this paper. *Extending the multiresolution methods discussed above in the context of one- and two-dimensional signals and images to spectral intensities will, in many applications, yield suboptimal results because the spatial and the spectral content of a spectral intensity might exhibit different degrees of smoothness.* Conventional wavelet-based approaches (e.g., three-dimensional wavelets), cannot optimally adapt in this setting.

Atkinson, Kamalabadi, and Jones [2] propose a wavelet-based method for estimating spectral intensities from noisy Gaussian observations. The basic idea behind their approach is to decorrelate the spectral data along the spectral dimension using the discrete Fourier transform, denoise each spatial map of Fourier coefficients independently using two-dimensional wavelet-based thresholding algorithms, and reform the images using the inverse Fourier transform. Manjón, Robles, and Thacker [34] propose an extension of the nonlocal means [3] filter

to spectral measurements for denoising spectral magnetic resonance (MR) images from Gaussian data. Scheunders suggests denoising spectral images from Gaussian observations using the interband correlations in the data [41]. However, these spectral intensity estimation algorithms are all designed for Gaussian noise statistics, and the impact of Poisson noise on their performance is not well understood. Some researchers have also considered the related problem of compressing spectral images [15], which, like our work, often relies on a low-complexity representation of the spectral image.

Recent work on marked Poisson processes suggests that the use of marks (e.g., spectral dimensions of a spectral image) can improve the spectral estimation accuracy when the observations are in the photon-limited regime. In this paper, we extend the approach of [48, 47] to a broader setting and show that the proposed algorithm is near-minimax optimal in a certain anisotropic Hölder–Besov function class. We also demonstrate its effectiveness through experiments conducted on realistic data sets.

2. Multiscale spatiospectral intensity estimation. In this section, we assume that the measurements collected at the detector are noisy but not distorted, so that A in (1.2) is the identity matrix, and describe our approach and analysis in this special case. The extension of these ideas to inverse problems will be discussed later in the paper. Thus, we let $\boldsymbol{y} \sim \text{Poisson}(\boldsymbol{f})$.

Our approach consists of finding the spectral image within a class of candidate estimates which optimizes a penalized log likelihood function. The class of candidate estimates and the penalty term are chosen to yield an estimator which is both near-minimax optimal for a broad and realistic class of spectral images and easy to compute rapidly. Specifically, we search over a family of *recursive dyadic partitions* (RDPs) of $[0,1]^3$, and for each partition consider the estimate formed by computing maximum likelihood model fits on each partition cell. We then choose the partition which gives the best fit to the data (in a log likelihood sense) and which has low complexity. This can formally be expressed as

$$\widehat{\boldsymbol{f}} \equiv \underset{\widetilde{\boldsymbol{f}} \in \Gamma_{M,N}}{\arg\min} \left\{ -\log p(\boldsymbol{y}|\widetilde{\boldsymbol{f}}) + \text{pen}(\widetilde{\boldsymbol{f}}) \right\}, \tag{2.1}$$

where $\Gamma_{M,N}$ is the class of candidate estimates and $\text{pen}(\cdot)$ is a complexity penalization term which satisfies the Kraft inequality [8] given by $\sum_{\widetilde{\boldsymbol{f}} \in \Gamma_{M,N}} e^{-\text{pen}(\widetilde{\boldsymbol{f}})} \leq 1$. The Kraft inequality plays an important role in our proof of an upper bound on the estimation error. Intuitively, the penalties can be thought of as negative log prior probabilities assigned to each candidate estimator, and the Kraft inequality ensures that the prior probabilities sum to something less than one.

The class of candidate estimators, $\Gamma_{M,N}$, corresponds to functions which are piecewise constant spatially and piecewise polynomial spectrally, where the breakpoints between the constant and polynomial pieces are constrained by an RDP. The role of the RDP framework is to allow more localized model fits in regions where the intensity is very inhomogeneous (such as near a boundary) and hence preserve that inhomogeneity, and yet use much less local model fits in regions where there is strong homogeneity. This encourages significant smoothing in homogeneous regions and removes noisy artifacts without eliminating key features.

The penalty of the estimate $\widehat{\boldsymbol{f}}$, which will be specified shortly, is proportional to the sum

of the polynomial order of the model fits across all the cells in the partition corresponding to $\widehat{\boldsymbol{f}}$. It is thus a measure of the estimator complexity and helps to balance the bias and the variance of the estimator; a higher penalty value favors smoother estimates with lower variance and higher bias, and a lower penalty value encourages complex estimates that have high fidelity to the observed data and a high variance.

The proposed approach takes advantage of correlations in the spectral image both between different spectral bands and between neighboring pixels by (a) dividing the spatial domain into spatially homogeneous regions, and (b) computing an estimate of the spectrum in each spatial region by dividing it into spectrally homogeneous or smooth bands. A sample partition of this kind is displayed in Figure 1. This approach leverages the key fact that spatial features such as boundaries between different spatial structures are manifested in the same spatial locations at all spectral bands, even though such features may have low contrast or be difficult to detect in certain spectral bands. Intuitively, we use high-contrast spectral bands to infer important information about the locations of the boundaries of these structures even in low-contrast spectral bands. Estimation procedures that do not leverage this fact can be much more vulnerable to noise or oversmoothing.

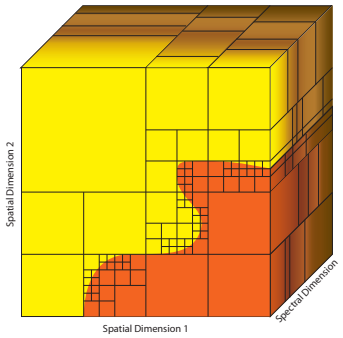

Figure 1. *Sample partition of a spatiospectral data cube. The spatial partition is the same at each spectral band, ensuring that the spatial boundaries are at the same locations in every spectral band.*

2.1. RDP estimators. We define a dyadic *spatial* partition $\mathcal{P}$ on $[0,1]^3$ to be a disjoint union of dyadic cuboids that covers $[0,1]^3$, where each cuboid is of the form $[k_1 2^{-j}, (k_1+1)2^{-j}] \times [k_2 2^{-j}, (k_2+1)2^{-j}] \times [0,1]$ for a scale parameter $j \in \{0,1,\dots,\log_2(N)\}$ and indices $k_1, k_2 \in \{0,1,\dots,2^j-1\}$. Similarly, each spectrum can be represented by an RDP defined using disjoint dyadic intervals on $[0,1]$. A spatiospectral partition $\mathcal{Q}$ on $[0,1]^3$ is thus a disjoint union of dyadic cuboids, where each cuboid has the form $[k_1 2^{-j}, (k_1+1)2^{-j}] \times [k_2 2^{-j}, (k_2+1)2^{-j}] \times [k_3 2^{-j'}, (k_3+1)2^{-j'}]$ for scale parameters $j \in \{0,1,\dots,\log_2(N)\}$, $j' \in \{0,1,\dots,\log_2(M)\}$ and indices $k_1, k_2 \in \{0,1,\dots,2^j-1\}$ and $k_3 \in \{0,1,\dots,2^{j'}-1\}$. The spatial and spectral RDPs can be represented in terms of quadtrees and binary trees, respectively, in which tree branches correspond to partition cells being subdivided into smaller partition cells with half the sidelength. We refer the reader to [25, 51] for more details on this.

We define the class of possible estimates $\Gamma_{M,N}$ as follows. Let r be a positive integer which is an upper bound on the smoothness of the spectra. Consider the set of all one-dimensional (1d) functions (i.e., spectra) $g' : [0,1] \to [C_\ell, C_u]$ that are piecewise polynomial

where each polynomial piece is of order r. The polynomial pieces are defined on dyadic intervals corresponding to a 1d RDP, and the polynomial coefficients are quantized to one of $\sqrt{n}$ levels. Each of these functions can then be clipped according to

$$g'_{\text{clipped}}(\lambda) = g'(\lambda) \cdot I_{\{g'(\lambda) > C_\ell\}}$$

so that the resulting function is positive. Let $\Gamma'_{M,N}$ denote the collection of all g'_{clipped} satisfying the above construction. (We focus on the above piecewise polynomial model for our analysis, but note that piecewise exponential models, such as those described in [29], are a simple extension of the above, where $\log(g')$ is a piecewise polynomial.)

The class of candidate spectral image estimates, $\Gamma_{M,N}$, can now be defined in terms of the class of candidate spectral estimates, $\Gamma'_{M,N}$. Consider all functions $g : [0,1]^3 \to [C_\ell, C_u]$ with a corresponding RDP $\mathcal{P}$ such that

$$g(x_1, x_2, \lambda) = \sum_{c \in \mathcal{P}} g'_c(\lambda) I_{\{(x_1,x_2) \in c\}},$$

where $g'_c \in \Gamma'_{M,N}$ for all $c \in \mathcal{P}$; i.e., g is spatially partitioned according to an RDP $\mathcal{P}$, and every spatial location in a given cell $c \in \mathcal{P}$ has a corresponding spectrum g'_c. Each element of the class $\Gamma_{M,N}$ corresponds to a sampled version of g, where the sampling is similar to sampling of f described in (1.1).

The penalty, as outlined earlier, is a measure of the estimator complexity and is proportional to the number of cells in the spatiospectral partition $\mathcal{Q}$ of the estimate $\widehat{\boldsymbol{f}}$ and the order of the polynomial fits to the partition cells of $\mathcal{Q}$. Specifically, we set the penalty of $\widehat{\boldsymbol{f}}$ to

$$\text{pen}(\widehat{\boldsymbol{f}}) = |\mathcal{Q}| \left(\frac{10}{3} + \frac{r}{2} \log_2 n \right) \log_e 2, \tag{2.2}$$

where n is the total number of observed photon counts. This penalty is simply proportional to the complexity of the estimate, i.e., the number of RDP cells times the number of polynomial coefficients in each cell. The other terms in the penalty expression are scaling factors which ensure than our penalty satisfies the Kraft inequality since our proof hinges on the application of the Li–Barron theorem (explained in detail in Appendix A), which presupposes that the Kraft inequality is satisfied. The origin of these terms is detailed in Appendix A. This particular choice of penalty leads to near-minimax estimators, as discussed in section 3, and yields an optimal balance of the bias and the variance terms of the estimates.

2.2. Estimation algorithm. The optimization problem in (2.1) can be solved accurately and efficiently using the approach described in this section. An initial, *complete* RDP of $[0,1]^2$ is obtained by recursively partitioning $[0,1]^2$ into cells with dyadic (power of two) sidelengths until the finest resolution (pixel-level) is reached. The optimal spatiospectral partition of the data is found by ascending through every level of the quadtree (starting at one level above the leaves), finding the best spectral estimate to the data in each RDP cell at that level, and pruning quadtree branches based on the penalized likelihood criterion. We explain this in detail below.

Given a spatial partition $\mathcal{P}$, the estimate $\boldsymbol{f}(\mathcal{P})$ can be calculated by finding the "best" model fit to the observed spectrum over each cuboid in $\mathcal{P}$. The spectral estimate for a given

cuboid can be computed using 1d penalized-likelihood Poisson intensity estimation methods, such as those described in [51]. In particular, for each cuboid c and for each spectral band we sum the observations. This yields Poisson observations, denoted $\widetilde{\boldsymbol{y}}^{(c)}$, of the aggregate spectrum $\widetilde{\boldsymbol{f}}^{(c)}_{i_3} = \sum_{(i_1,i_2)\in c} \boldsymbol{f}_{i_1,i_2,i_3}$. We then estimate the intensity of each spectrum of this form. This can be accomplished by pruning an RDP representation of the spectrum; the spectral RDP can be represented using a binary tree, and the models fit to each terminal interval in the spectral RDP can be constants, polynomials, or exponentials.

In each partition cell c, we compute the following penalized log likelihood:

$$\widehat{\boldsymbol{f}}^{(c)} = \underset{\widetilde{\boldsymbol{f}}^{(c)}\in\Gamma_{M,N}}{\arg\min} \left\{ L^{(c)} \right\}, \tag{2.3}$$

where $L^{(c)} = -\log p(\widetilde{\boldsymbol{y}}^{(c)}|\widehat{\boldsymbol{f}}^{(c)}) + \mathrm{pen}(\widehat{\boldsymbol{f}}^{(c)})$ and $p(\widetilde{\boldsymbol{y}}^{(c)}|\widetilde{\boldsymbol{f}}^{c})$ corresponds to the Poisson likelihood given by

$$p(\widetilde{\boldsymbol{y}}^{(c)}|\widetilde{\boldsymbol{f}}^{(c)}) = \prod_{i_3=0}^{M-1} \frac{e^{-\widetilde{\boldsymbol{f}}^{(c)}_{i_3}} \left(\widetilde{\boldsymbol{f}}^{(c)}_{i_3}\right)^{\widetilde{\boldsymbol{y}}_{i_3}}}{\widetilde{\boldsymbol{y}}_{i_3}!}. \tag{2.4}$$

We define $\mathrm{pen}(\widetilde{\boldsymbol{f}}^{(c)})$ to be the penalty proportional to the number of terminal intervals in the pruned binary RDP; the penalties are discussed in detail in [25, 51].

As the algorithm iterates over every level of the quadtree, it prunes the branches of the quadtree to find the partition $\mathcal{P}$ with the minimal sum of the $\left\{L^{(c)}\right\}$. The final spatiospectral estimate is then calculated by finding the partition $\widehat{\mathcal{P}}$ that minimizes the total penalized likelihood function:

$$\widehat{\mathcal{P}} \equiv \underset{\mathcal{P}}{\arg\min} \left\{ \sum_{c\in\mathcal{P}} L^{(c)} \right\} \quad \text{and} \quad \widehat{\boldsymbol{f}}_{i_1,i_2,i_3} \equiv \sum_{c\in\widehat{\mathcal{P}}} \widehat{\boldsymbol{f}}^{(c)}_{i_3} I_{\{(i_1,i_2)\in c\}}. \tag{2.5}$$

In this paper, we refer to this approach as the *full spatiospectral denoising algorithm* since at every level of quadtree pruning we perform both spatial and spectral smoothing. Each of the terminal intervals in the pruned RDP could correspond to a homogeneous or smoothly varying region of intensity. This endows our estimators with spatially and spectrally varying resolution to automatically increase the smoothing in very regular regions of the intensity and preserve detailed structures in less homogeneous regions.

This approach is similar to the image estimation method described in [25, 49], with the key distinction that the proposed method *forces the spatial RDP to be the same at every spectral band*. This constraint ensures that the method preserves the spatial boundaries at the same locations in every spectral band, irrespective of the contrast differences among different spectral bands. In other words, when a tree branch is pruned, it means partition cells are merged in every spectral band simultaneously at the corresponding spatial location. This approach is effective because an outlier observation in one spatiospectral voxel may not be recognized as such when spectral bands are considered independently, but may be correctly pruned when the corresponding spectrum is very similar to a spatially neighboring spectrum.

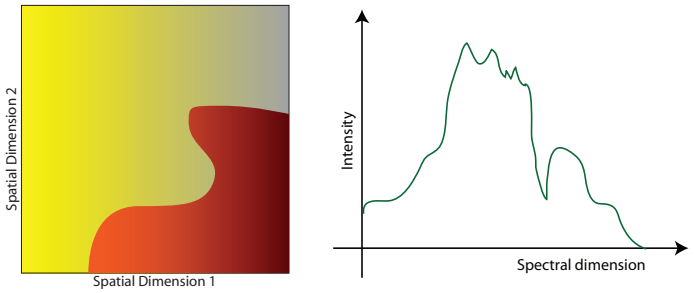

Figure 2. *Illustration of the anisotropic Hölder–Besov function class.* (a) *Spatial variations in* $f(x_1, x_2, \lambda')$ *for a fixed* λ'. *The intensity is piecewise smooth along the spatial coordinates with Hölder smooth surfaces separated by a Hölder smooth boundary.* (b) *Besov smooth spectrum corresponding to fixed spatial location* (x_1', x_2') *that is piecewise smooth with finitely many discontinuities.*

3. Error analysis. In this section, we present a minimax upper bound on the risk of the *full spatiospectral denoising algorithm* proposed in section 2.2, and also derive minimax lower bounds on the risk of any spectral intensity estimation procedure to demonstrate the near-optimality of our approach over an anisotropic Hölder–Besov function class to be defined precisely below.

To facilitate the error analysis, let us reformulate the intensity estimation problem using a multinomial framework. The conditional distribution of a random variable $\boldsymbol{y} \sim \text{Poisson}(\boldsymbol{f})$ with unit total intensity (i.e., the components of $\boldsymbol{f}$ sum to one), given that the observed number of photon counts n is multinomial. Estimating the Poisson intensity can be broken into two components: (1) estimating the total intensity $I_{\boldsymbol{f}}$ of the spectral image and (2) estimating the normalized spectral image $\boldsymbol{f}/I_{\boldsymbol{f}}$. The multinomial framework allows us to bound the error between $\boldsymbol{f}/I_{\boldsymbol{f}}$ and $\widehat{\boldsymbol{f}}/I_{\widehat{\boldsymbol{f}}}$. Unfortunately, the error in estimating $I_{\boldsymbol{f}}$ can make it very difficult to bound the error between $\boldsymbol{f}$ and $\widehat{\boldsymbol{f}}$ using the Poisson framework, even though the normalized error is primarily important to most end-users. Assuming that the true continuous-domain intensity f integrates to unity, $I_f \equiv \int f = 1$, the components of $\boldsymbol{f}$ will sum to unity as well. Consequently, we restrict each $\widetilde{\boldsymbol{f}} \in \Gamma_{M,N}$ to be positive and to sum to one. The observations $\boldsymbol{y}$ are now assumed to follow a multinomial distribution with parameter vector $\boldsymbol{f}$; that is, $\boldsymbol{y} \sim \text{Multinomial}(\boldsymbol{f}; n)$. We enumerate the voxels according to the lexicographic ordering and let $\boldsymbol{f}_j$ be the component of f corresponding to the jth voxel. The likelihood of observing $\boldsymbol{y}$ given $\boldsymbol{f}$ under this model is

$$p(\boldsymbol{y}|\boldsymbol{f}) = \frac{n!}{\boldsymbol{y}_1!\boldsymbol{y}_2!\dots\boldsymbol{y}_{N^2M}!}\boldsymbol{f}_1^{\boldsymbol{y}_1}\boldsymbol{f}_2^{\boldsymbol{y}_2}\dots\boldsymbol{f}_{N^2M}^{\boldsymbol{y}_{N^2M}}. \tag{3.1}$$

3.1. Specification of a Hölder–Besov function class. We assume that the underlying density f lies in an anisotropic Hölder–Besov function class $\mathcal{F}$, which we define as the space of functions on the unit cube $[0,1]^3$ that exhibit piecewise Hölder smoothness in the first two (spatial) dimensions and Besov smoothness in the third (spectral) dimension, as illustrated in Figure 2. To describe such a Hölder–Besov function class we will draw upon the machinery of *anisotropic smoothness classes* [36]. We start by developing an appropriate notion of a

modulus of smoothness, following [9]. Consider a continuous function f on $[0,1]^d$ for some $d \geq 1$. Given $r = 1, 2, \ldots$ and $h = (h_1, h_2, \ldots, h_d) \in \mathbb{R}^d$, let $\Delta_h^r f : \mathbb{R}^d \to \mathbb{R}$ be the *rth difference of f with step h*:

$$\Delta_h^r f(x_1, x_2, \ldots, x_d) \triangleq \sum_{k=0}^{r} \binom{r}{k} (-1)^{r-k} f(x_1 + kh_1, x_2 + kh_2, \ldots, x_d + kh_d).$$

We will mostly deal with the case when d is equal to 1, 2, or 3. Note that the mapping $f \mapsto \Delta_h^r f$ is linear, and the function $\Delta_h^r f$ is supported on the set $A_{r,h}^{(d)} \triangleq \prod_{i=1}^{d} [0, \min\{0, 1 - rh_i\}]$. The corresponding moduli of smoothness are given by

$$\omega_r(f,t)_p \triangleq \sup_{h \in \mathbb{R}^d, \|h\| \leq t} \|\Delta_h^r f\|_{L_p},$$

where $r = 1, 2, \ldots$, $1 \leq p \leq \infty$, and $\omega(f,t) \triangleq \omega_1(f,t)_\infty$.

We are interested in functions on $[0,1]^3$ that exhibit different kinds of smoothness in different coordinate directions. In particular, we consider functions that are Hölder in the first two directions and Besov in the third direction. We will say that a function $f : [0,1]^3 \to \mathbb{R}$ is *spatially piecewise Hölder-(α, γ) smooth* if, for any fixed $\lambda \in [0,1]$ and all $(x_1, x_2) \in [0,1]^2$, we can write

$$f(x_1, x_2, \lambda) = f_1(x_1, x_2, \lambda) I_{\{H(x_1) \geq x_2\}} + f_2(x_1, x_2, \lambda) I_{\{H(x_1) < x_2\}}, \tag{3.2}$$

where, for each $\lambda \in [0,1]$, the surfaces $(x_1, x_2) \mapsto f_j(x_1, x_2, \lambda)$ for $j = 1, 2$ are Hölder-α for $\alpha \in (0,1]$, so that

$$|f_j(x_1, x_2, \lambda) - f_j(x_1', x_2', \lambda)| \leq C_\alpha ((x_1 - x_1')^2 + (x_2 - x_2')^2)^{\alpha/2} \tag{3.3}$$

for any $(x_1, x_2), (x_1', x_2') \in [0,1]^2$ [28]. Also, $H(x)$ is Hölder-γ for $\gamma \in (0,1]$ so that

$$|H(x) - H(x')| \leq C_\gamma |x - x'|^\gamma. \tag{3.4}$$

In other words, for each λ, $f(x_1, x_2, \lambda)$ consists of two Hölder surfaces separated by a Hölder boundary. Further assume that for any fixed $(x_1, x_2) \in [0,1]^2$ the 1d function $\lambda \mapsto f(x_1, x_2, \lambda)$ is in the Besov space $\mathcal{B}_p^\beta(L_p([0,1]))$, where the smoothness parameter $\beta > 0$ and $1/p = \beta + 1/\tau$, and where $L_\tau([0,1])$ is the approximation space [9]. In this paper, we fix $\tau = 2$.

In terms of the moduli of smoothness, the surfaces f_1 and f_2 that enter into the definition in (3.2) belong to a family of functions $f : [0,1]^3 \to \mathbb{R}$ such that for a given $1 \leq p, q < \infty$, $0 < \alpha \leq 1$, and $\beta > 0$,

$$|f|_{H_\alpha}^{(1,2)} \triangleq \sup_{0 \leq \lambda \leq 1} \sup_{t > 0} \left(t^{-\alpha} \omega(f(\cdot, \lambda), t) \right) < +\infty \tag{3.5}$$

and

$$|f|_{B_q^\beta(L_p)}^{(3)} \triangleq \sup_{(x,y) \in [0,1]^2} \left(\int_0^1 \left(t^{-\beta} \omega_r(f(x, y, \cdot), t)_p \right)^q \frac{dt}{t} \right)^{1/q} < +\infty, \tag{3.6}$$

where $r = \lfloor \beta \rfloor + 1$ [9]. The superscript $(1,2)$ on $|f|^{(1,2)}_{H^\alpha}$ in (3.5) refers to the fact that we measure the modulus of smoothness only along the first two spatial dimensions; the same goes for the superscript (3) in (3.6). With these definitions, we can formalize the notion of different types of smoothness in different coordinate directions. In particular, the condition in (3.5) defines the functions f on $[0,1]^3$ that are uniformly Hölder in the first two coordinate directions. Similarly, the condition (3.6) defines the functions f on $[0,1]^3$ that are uniformly Besov in the third coordinate direction. Thus, f_1 and f_2 in (3.2) belong to the function class

$$\mathcal{F}_{\alpha,(\beta,p,q)}(L) \triangleq \left\{ f : [0,1]^3 \to \mathbb{R} \middle| \|f\|_{L^2} + |f|_{H^{(1,2)}_\alpha} + |f|_{B^{\beta,(3)}_q(L_p)} < L \right\}$$

for some $L > 0$. We will see that, although our estimate must be a member of the class $\Gamma_{M,N}$, we can accurately estimate any function in the Hölder–Besov space.

3.2. Upper bounds on the risk function. The derivation of the upper bounds on the risk function follows that in [51], but with some significant differences because of the fact that we consider *anisotropic* 3d Hölder–Besov densities here, whereas [51] deals with 1d and 2d densities in the Besov and Hölder function spaces, respectively, and explicitly considers discrete-domain intensities.

We define the risk function between the true intensity $\boldsymbol{f}$ and its penalized log likelihood estimate $\widehat{\boldsymbol{f}}$ (defined in (2.1)) as follows: $R(\boldsymbol{f}, \widehat{\boldsymbol{f}}) \equiv \mathbb{E}\left[\mathcal{H}^2(\boldsymbol{f}, \widehat{\boldsymbol{f}})\right]$, where

$$\mathcal{H}^2(\boldsymbol{f}, \widehat{\boldsymbol{f}}) = \sum_{j=1}^{N^2 M} \left(\sqrt{\boldsymbol{f}_j} - \sqrt{\widehat{\boldsymbol{f}}_j} \right)^2 \tag{3.7}$$

is the squared Hellinger distance between $\boldsymbol{f}$ and $\widehat{\boldsymbol{f}}$. Here, the expectation is taken with respect to the observations.

Theorem 3.1. *Let $\mathcal{F}$ denote the class of functions of the form (3.2). Suppose that the observation $\boldsymbol{y} \sim \text{Multinomial}(\boldsymbol{f}; n)$, where $\boldsymbol{f}$ is obtained from an unknown intensity $f \in \mathcal{F}$ with $I_f = 1$ via integration sampling. Let $\Gamma_{M,N}$ be a class of candidate estimators, as detailed in section 2, and let $\widehat{\boldsymbol{f}}$ be the estimate obtained by the full spatiospectral denoising algorithm according to (2.5). Assume that every $\widetilde{\boldsymbol{f}} \in \Gamma_{M,N}$ satisfies the condition $\widetilde{\boldsymbol{f}}_j \geq \frac{C}{N^2 M}$, for $j = 1, \ldots, N^2 M$ and some $C \in (0,1)$, and that the penalty is given by (2.2). Assume that N and M are sufficiently large, so that*

$$M \log_2 M > C' \left(\frac{n}{\log_e n} \right)^{\frac{(2\beta+1)\nu}{2\beta(\nu+2-\gamma)+\nu}},$$

$$N \log_2 M > C' \left(\frac{n}{\log_e n} \right)^{\frac{2\beta}{2\beta(\nu+2-\gamma)+\nu}},$$

where C' is a constant independent of n, N, and M and dependent on the smoothness parameters α, β, and γ. Assume that the polynomial fits in (2.5) are of order $r \geq \lceil \beta \rceil$. Then

$$R(\boldsymbol{f}, \widehat{\boldsymbol{f}}) \preceq \left(\frac{n}{\log_2 M \log_e n} \right)^{\frac{-2\beta\nu}{2\beta(\nu+2-\gamma)+\nu}}, \tag{3.8}$$

where $\nu = \min(\alpha, \gamma)$.

The proof of Theorem 3.1 is given in Appendix A. As detailed in the proof, the $\log_2 M$ term arises from a combination of known approximation error bounds for free knot polynomials and our restriction of polynomial endpoints to boundaries in a recursive dyadic partition.

For a fixed n, the lower bounds on N and M place a limit on the spatial and spectral resolution of the observed data to ensure that the photon noise is dominant over the errors due to sampling. In particular, if M and N are held fixed (and small), but the number of photons n increases, at some point photon noise will be negligible relative to errors from sampling the spectral image to fit on the $N \times N \times M$ grid. In that case, the estimation error will not continue to decay with n at the rate our bounds suggest. Since approximation errors due to sampling are not the focus of this paper, we consider the case where M and N are relatively large.

3.3. Lower bounds on the risk function. In this section, we present lower bounds on the Hellinger risk when the unknown target intensity is a member of the Hölder–Besov class. To keep the technical details to a minimum, we consider the problem of estimating the actual continuous-domain intensity $f : [0,1]^3 \to [0,1]$, without discretization. Effectively, this corresponds to the case when the number of voxels $N^2 M \to \infty$ while the number of observed photon counts n is held fixed. Since $\int_{[0,1]^3} f = 1$, this asymptotic scenario is equivalent to observing n independent samples $z_1, \ldots, z_n$ from a *probability density* f with support in the unit cube $[0,1]^3$, and the goal is to estimate f. To measure the quality of a candidate estimator $\widehat{f}$, we adopt the Hellinger risk $R(f, \widehat{f}) \triangleq \mathbb{E}\big[\mathcal{H}^2(f, \widehat{f})\big]$, where

$$\mathcal{H}^2(f, \widehat{f}) = \int_{[0,1]^3} \left[\sqrt{f(x_1, x_2, \lambda)} - \sqrt{\widehat{f}(x_1, x_2, \lambda)} \right]^2 dx_1 dx_2 d\lambda$$

and the expectation is taken w.r.t. $z_1, \ldots, z_n$. The squared Hellinger distance $\mathcal{H}^2(\boldsymbol{f}, \widehat{\boldsymbol{f}})$ defined in (3.7) can be viewed as a discrete approximation of the above integral, and the limit $N, M \to \infty$ corresponds to taking increasingly fine approximations. Moreover, the set of all possible estimators of f includes those that first bin the observations into voxels. Thus, the lower bounds on the risk of estimating the actual continuous-domain intensity f also provide lower bounds for the discrete estimators of $\boldsymbol{f}$. We are interested in lower bounds on the minimax risk $R_n(\mathcal{F}) \triangleq \inf_{\widehat{f}} \sup_{f \in \mathcal{F}} R(f, \widehat{f})$, where the infimum is over all estimators $\widehat{f}$ based on n i.i.d. (independently and identically distributed) samples $z_1, \ldots, z_n$ from f.

Our derivation of the minimax lower bound relies on a powerful information-theoretic method of Yang and Barron [54]. The key idea behind the method of [54] is that the minimax rates of convergence for a wide variety of function classes can be determined from the global metric properties of a carefully chosen subset of the particular function class. These metric properties are encoded in the covering and the packing numbers of the class. Before stating the main result of [54], let us define the following. Given $\epsilon > 0$ and a function class $\mathcal{G} \subset L_2([0,1]^d)$, a finite set $S \subset \mathcal{G}$ is an *ϵ-packing set* for $\mathcal{G}$ if

$$\min \left\{ ||g - g'||_{L_2([0,1]^d)} : g, g' \in S; g \neq g' \right\} \geq \epsilon.$$

Let $M(\epsilon, \mathcal{G})$ denote the cardinality of the maximal ϵ-packing set for $\mathcal{G}$ with respect to $||.||_{L_2([0,1]^d)}$; the Kolmogorov ϵ-capacity of $\mathcal{G}$ is defined by $K_\epsilon(\mathcal{G}) \triangleq \log M(\epsilon, \mathcal{G})$ [27].

One of the key results of [54] is that, if $\mathcal{F} \supseteq \mathcal{G}$ is a class of density functions bounded above and below such that $0 < C_\ell \leq f \leq C_u < \infty$ for all $f \in \mathcal{F}$, then we have the minimax lower bound $R_n(\mathcal{F}) \succeq \epsilon_n^2$, where ϵ_n is the *critical separation distance*, determined from the Kolmogorov packing entropy by solving the equation

$$K_{\epsilon_n}(\mathcal{G}) = n\epsilon_n^2. \tag{3.9}$$

In this paper, as indicated in section 3, $\mathcal{F}$ represents the anisotropic Hölder–Besov function class consisting of piecewise Holder surfaces separated by a Hölder boundary in the spatial dimensions and piecewise Besov spectra in the spectral dimension. Using the result from [54] referred to above, we arrive at the following lower bound.

Theorem 3.2. *Let us assume that we observe n i.i.d. realizations drawn from a density $f \in \mathcal{F}$ and that $0 < C_\ell \leq f \leq C_u < \infty$. Let $\widehat{f}$ be any estimate of f based on these n realizations. Then the minimax lower bound is given by*

$$\mathbb{E}\left[\mathcal{H}^2(f, \widehat{f})\right] \succeq \max\left(n^{-2\gamma/(2\gamma+1)}, n^{-2\alpha\beta/(2\alpha\beta+2\beta+\alpha)}\right) \tag{3.10}$$

for $\alpha \in (0, 1]$, $\gamma \in (0, 1]$, and $\beta > 0$.

In particular, for $\gamma = 1$ we have $\nu = \min(\alpha, \gamma) = \alpha$, and

$$\mathbb{E}\left[\mathcal{H}^2(f, \widehat{f})\right] \succeq \max\left(n^{-2/3}, n^{-2\alpha\beta/(2\alpha\beta+2\beta+\alpha)}\right) \equiv n^{-2\alpha\beta/(2\alpha\beta+2\beta+\alpha)}.$$

Thus the lower bound matches the upper bound given in (3.8) up to a log factor when $\gamma = 1$. The proof of Theorem 3.2 is given in Appendix B.

4. Computational complexity. Implementing the full spatiospectral algorithm involves performing both spatial and spectral smoothing at every level of the tree. For a datacube of size $N \times N \times M$ the computational complexity of finding a spatiospectral estimate with piecewise constant fits in both the spatial and spectral dimensions is $\mathcal{O}(N^2 M)$. To see this, first note that the computational complexity of performing binary tree pruning of a spectrum of length M and fitting constants to each optimal partition interval for a single spectrum is $\mathcal{O}(M)$, because constant fits (and their likelihoods) at one node of a tree can be computed from the average of the constant fits of the children. In the full spatiospectral algorithm, as described in section 2, the spectral smoothing operation is performed on every unique spectrum at each level of the quadtree. At scale j of the quadtree, there are $N^2/2^{2j}$ spectra to estimate, where j ranges from 0 to $\log_2 N$. Since $\sum_{j=0}^{\log_2 N} \frac{N^2}{2^{2j}} = \mathcal{O}(N^2)$, the computational complexity of finding the optimal spatiospectral partition and fitting piecewise constants to the partition cells is $\mathcal{O}(N^2 M)$.

The computational complexity of finding a spatiospectral estimate with piecewise constant fits in the spatial dimension and piecewise polynomial fits in the spectral dimension depends on the complexity of the optimization routine used to find the polynomial coefficients. To prune a spectrum of length M and to fit polynomials to each partition interval, we will need $\mathcal{O}(M)$ likelihood function calls (which can be computed using a total of $\mathcal{O}(M \log_2 M)$ operations) and

$\mathcal{O}(M)$ optimization routine calls. Thus, the full spatiospectral denoising algorithm that makes piecewise constant fits in the spatial dimension and piecewise polynomial fits in the spectral dimension requires $\mathcal{O}(N^2 M \log_2 M)$ likelihood function calls and $\mathcal{O}(N^2 M)$ polynomial fitting routine calls.

The accuracy of the proposed estimator can be augmented by a process called cycle-spinning, or averaging over shifts, resulting in translation-invariant (TI) estimates [30, 7]. Cycle-spinning was derived in the context of undecimated wavelet coefficient thresholding in the presence of Gaussian noise, but it can be difficult to implement efficiently in our case when spectral smoothing is performed at every leaf of the quadtree. If spectral smoothing is not required (which might be the case when the spectral intensity is uncorrelated along the spectral dimension), then TI estimates with piecewise constant spatial fits can be obtained in $\mathcal{O}(NM \log_2 N)$ time using some novel computational methods discussed in [49].

5. Spectral image reconstruction. The proposed multiscale method for spatiospectral denoising can now be used in an EM framework to reconstruct a blurred noisy spectral image. As explained in section 1.1, the observations $\boldsymbol{y} \sim \text{Poisson}(A\boldsymbol{f})$ collected at the detector are noisy and distorted (A corresponds to the distortion operator), and our goal is to estimate $\boldsymbol{f}$ from $\boldsymbol{y}$ as accurately as possible.

To solve this challenging inverse problem, we perform the following optimization problem:

$$\widehat{\boldsymbol{f}} = \underset{\widetilde{\boldsymbol{f}} \in \Gamma_{M,N}}{\arg\min} \left\{ -\log p(\boldsymbol{y} | A\widetilde{\boldsymbol{f}}) + \text{pen}(\widetilde{\boldsymbol{f}}) \right\}, \tag{5.1}$$

where $\Gamma_{M,N}$ is the collection of estimators corresponding to a pruned spatiospectral tree, as described in section 2, and the penalty $\text{pen}(\widetilde{\boldsymbol{f}})$ is proportional to the number of cells in the pruned RDP; hence the penalty term encourages solutions with small numbers of leaves. We compute the solution to this problem using an EM algorithm, which in this case is a regularized version of the Richardson–Lucy algorithm [39, 33, 38]. The method consists of two alternating steps:

E-step: $\boldsymbol{x}^{(t)} = \widehat{\boldsymbol{f}}^{(t)} . \times A^T(\boldsymbol{y} ./ A\widehat{\boldsymbol{f}}^{(t)})$, where $.\times$ and $./$ denote elementwise multiplication and division, respectively.

M-step: Compute $\widehat{\boldsymbol{f}}^{(t+1)}$ by denoising $\boldsymbol{x}^{(t)}$ as described in section 2.

6. Experimental results. In this section we demonstrate the effectiveness of the proposed spatiospectral algorithm on the NASA AVIRIS (airborne visible/infrared imaging spectrometer) Moffett field reflectance data set.

6.1. Denoising results. In these experiments, we focus on a region of the data cube which is $256 \times 256 \times 128$. Furthermore, we scale the data so that observations are truly photon-limited and the mean intensity per voxel is 0.0387. Figures 3(a) and 3(b) show the true intensity and the noisy observations, respectively, at spectral band 6. Here we compare our denoising algorithm to three other approaches: (a) Kolaczyk's corrected Haar thresholds extended to spectral data [26, 23], (b) Kolaczyk's and Nowak's multiscale CPLE method applied to every spectral image independently [25], and (c) the Fourier- and wavelet-based spectral image estimation algorithm (or Atkinson method) [2]. To provide a quantitative comparison among the results, we use the following error measure: $\varepsilon = \frac{\|\boldsymbol{f} - \widehat{\boldsymbol{f}}\|_1}{\|\boldsymbol{f}\|_1}$. In the experiments discussed

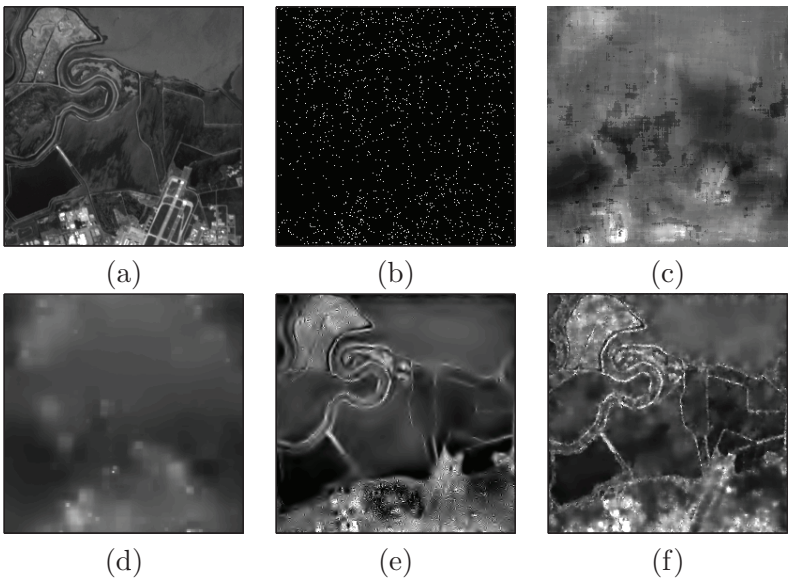

Figure 3. *Spatiospectral denoising results.* (a) *True intensity at spectral band* 6. (b) *Noisy observations at spectral band* 6. (c) *Result obtained by applying the corrected* 3*d Haar wavelet thresholds* [26, 23]*, seen at spectral band* 6*;* $\varepsilon = 0.3497$. (d) *Result obtained by performing the multiscale CPLE method* [25] *on every image separately, seen at spectral band* 6*;* $\varepsilon = 0.2887$. (e) *Result obtained by performing the Atkinson method* [2]*, seen at spectral band* 6*;* $\varepsilon = 0.2298$. (f) *Result obtained by performing the proposed full spatiospectral denoising algorithm, averaged over* 2000 *different spatial and spectral shifts, seen at spectral band* 6*;* $\varepsilon = 0.1917$.

below, the error numbers are obtained by averaging the error obtained in running independent experiments over 100 different noise realizations.

Kolaczyk's corrected Haar wavelet threshold algorithm extends the wavelet-based thresholding approaches to Poisson data. To account for the signal-dependence of the Poisson statistics, Kolaczyk suggested the use of *corrected*, scale-dependent thresholds at every level of decomposition based on the tail probabilities of the wavelet coefficients followed by hard or soft thresholding. Here we extend this approach to 3d Haar wavelet transform to denoise spectral data and use the following thresholds at every level $j = 0, 1, \ldots, J$ for $J = \log_2 N, \log_2 M$:

$$t_j = t' 2^{-3(J-j)/2} \left[\log(n_j) + \sqrt{\log^2(n_j) + 2\lambda_j \log(n_j)} \right],$$

where $n_j = 2^{3j}$, $\lambda_j = 2^{3(J-j)}\lambda'$, and t' and λ' are user-defined parameters that are chosen to minimize the error. In this experiment, we use hard thresholding. This approach assumes the same degrees of smoothness in the spatial and the spectral dimensions. However, if the underlying intensity has anisotropic smoothness in the spatial and the spectral dimensions, such an approach becomes less effective, as demonstrated by the results shown in Figures 3(c) and 4(a), respectively. The parameters t' and λ' are chosen to minimize the error, which is $\varepsilon = 0.3497$.

The multiscale CPLE method finds a spatial partition that maximizes the penalized log likelihood from the space of all possible recursive dyadic partitions in $[0, 1]^2$ and fits piecewise constants to every partition interval [25]. In this experiment, we weighted the spatial penalty

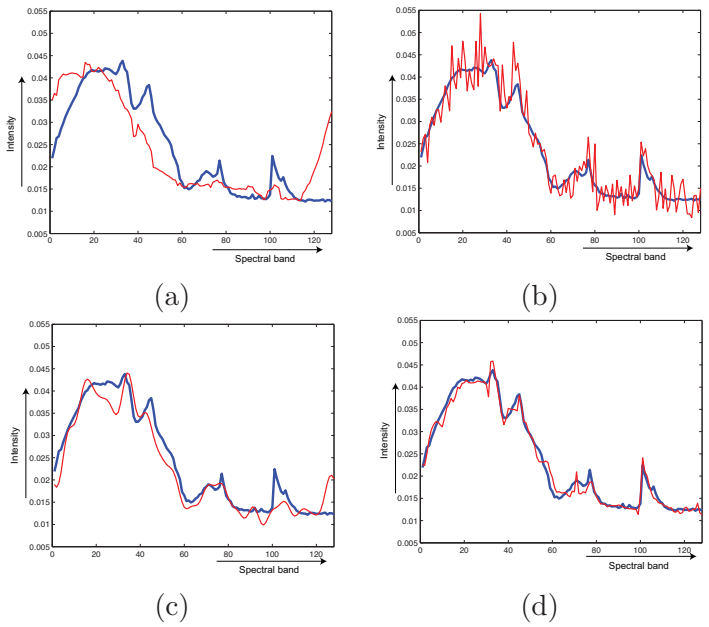

Figure 4. *Spatiospectral denoising results. True spectrum at location* $(31, 160)$ *is shown by the thick blue line, and estimates are shown by the thin red line.* (a) *The estimate obtained by using Kolaczyk's corrected Haar wavelet thresholding approach.* (b) *The estimate obtained by performing the multiscale CPLE method on every image independently.* (c) *The estimate obtained by using the Atkinson method.* (d) *The estimate obtained by using the proposed full spatiospectral denoising algorithm.*

to minimize the error. Since the observations are extremely photon-limited, performing the multiscale CPLE method on every spectral image independently yields oversmoothed results, as shown in Figures 3(d) and 4(b). The average error is $\varepsilon = 0.2887$.

The Atkinson method proposed in [2] is designed to perform spectral intensity estimation from Gaussian observations. A possible approach for handling Poisson data is to apply the Anscombe transform to the Poisson data and approximate them to Gaussian. However, such a transformation is inaccurate when the observations are extremely photon-limited; for example, when the mean intensity per voxel is 0.0387, then the number of photons measured at each voxel is either 0 or 1. Under such circumstances, the Anscombe transform breaks down and hence yields suboptimal results. In our experiments, we found that the error was much higher ($\varepsilon = 0.3471$) when we performed the wavelet-based spectral intensity estimation algorithm after applying the Anscombe transform to the Poisson data. The results obtained by this algorithm without variance stabilization are shown in Figures 3(e) and 4(c), and the error is $\varepsilon = 0.2298$. We chose a wavelet threshold that minimized the error. From the results, we can see that this algorithm performs better than the multiscale CPLE method because it accounts for the spectral correlation that exists between different spectral bands. However, the algorithm fails to preserve some fine edges when the number of observed photons is extremely low, as shown in Figure 3(e).

Figures 3(f) and 4(d) show the results obtained by applying the proposed full spatiospectral denoising algorithm with $r = 1$, averaged over 2000 random spatial and spectral shifts

to overcome the blocky artifacts introduced by the full spatiospectral denoising algorithm ($\varepsilon = 0.1917$). The penalty in (2.2) was multiplied by a small constant factor to yield a low error and also visually appealing results. From the results, we can see that the proposed approach outperforms the other two algorithms. The algorithm is very effective in estimating fine (high-frequency) details even when the observations are extremely photon-limited, as seen in Figure 3(f). Similar results were achieved using piecewise polynomial fits to the *log* of the spectrum using generalized linear models [29].

6.2. Reconstruction results. Reconstruction problems are generally much more challenging and computationally intensive than denoising problems because of the iterative algorithms often used to solve such problems. In the experiments discussed below, we restrict our attention to a smaller subset of the AVIRIS data of size $128 \times 128 \times 64$ to reduce the computational complexity. We also increase the mean intensity per voxel to be 20.7614 because of the difficulty associated with the reconstruction in the inverse problem setting. In this experiment, we apply the full spatiospectral denoising algorithm as the M-step of the EM algorithm and fit piecewise constants both spatially and spectrally. To overcome the blocky artifacts resulting from the full spatiospectral denoising algorithm, and to approximate cycle-spinning, we perform the following at every iteration:

E-step: $\boldsymbol{x}^{(t)} = \widehat{\boldsymbol{f}}_i^{(t)} . \times A^T(\boldsymbol{y}./A\widehat{\boldsymbol{f}}_i^{(t)})$.

M-step: Compute $\widehat{\boldsymbol{f}}_i^{(t+1)} = \mathcal{S}_i^{-1}(\text{Denoise}[\mathcal{S}_i(\boldsymbol{x}^{(t)})])$.

Estimate at iteration t: $\widehat{\boldsymbol{f}}^{(t+1)} = \frac{1}{k}\sum_{i=1}^{k} \widehat{\boldsymbol{f}}_i^{(t+1)}$.

The shift operator $\mathcal{S}_i(.)$ shifts the argument $(.)$ by i voxels in either of the three coordinate directions, and the inverse operator $\mathcal{S}_i^{-1}(.)$ undoes the shifting. This procedure is a very good approximation to cycle-spinning, as the number of shifts k approaches the detector resolution N^2M. We stop iterating when $\|\widehat{\boldsymbol{f}}^{(t+1)} - \widehat{\boldsymbol{f}}^{(t)}\|_2^2 / \|\widehat{\boldsymbol{f}}^{(t)}\|_2^2 < 1 \cdot 10^{-4}$.

Figure 5(a) shows the true intensity at spectral band 61. Figure 5(b) shows the blurred and noisy observations measured at the detector. Figure 5(c) shows the result of the Richardson–Lucy (RL) algorithm after convergence. The RL reconstruction is very noisy, as is evident from Figure 5; the error is $\varepsilon = 0.2700$. In order to provide a more meaningful comparison, we computed an ad hoc approximation to the EM algorithm, in which we used the Atkinson method of [2] in the M-step; this yielded the result presented in Figure 5(d). The method is considered an ad hoc, approximate EM algorithm because it does not correspond to the penalized likelihood criterion in (5.1) for any known penalization method or prior probability model on the wavelet and Fourier coefficients. In performing this experiment, we used the undecimated wavelet transform and found a threshold that minimizes the error and also yields visually sharp results. The error associated with this approach is $\varepsilon = 0.1467$. Figure 5(e) shows the result obtained by using the proposed reconstruction algorithm. As with the previous algorithm, the spatial and the spectral penalties were chosen to yield an estimate with minimum error and visually sharp results; the error is $\varepsilon = 0.1415$. Visually, the reconstruction results shown in Figures 5(d) and 5(e) are comparable. This is largely because of the fact that we use an undecimated wavelet transform in the wavelet-based spectral reconstruction algorithm and just average over nine different spatial shifts in the case of the proposed approach. To make a fair comparison, we conducted an experiment with the Atkinson method using the decimated wavelet transform and averaged over nine spatial shifts, similar to our

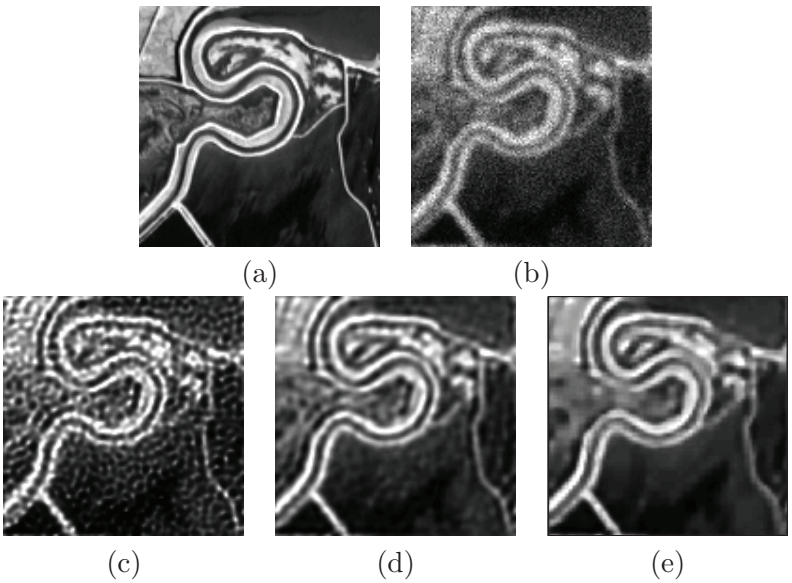

Figure 5. *Spatiospectral reconstruction results.* (a) *True intensity at spectral band* 61. (b) *Blurred and noisy observations at spectral band* 61. (c) *Result obtained by performing the RL algorithm, seen at spectral band* 61*;* $\varepsilon = 0.2700$. (d) *Result obtained by performing the Atkinson method* [2] *as the M-step of the EM algorithm, seen at spectral band* 61*;* $\varepsilon = 0.1467$. (e) *Result obtained by performing the proposed full spatiospectral denoising algorithm as the M-step of the EM algorithm, averaged over nine spatial shifts and seen at spectral band* 61*;* $\varepsilon = 0.1415$.

proposed approach. The reconstruction with the Atkinson method did not converge, and the results were unacceptable.

In all the experiments discussed above, the penalties were chosen to minimize the error and yield visually appealing results since the ground truth was available. When the true spectral data are unknown but multiple realizations of the data are available, then the data can be split into test and validation data, and methods such as cross validation can be used to find the tuning parameters.

In related work, our approach was shown to be similarly effective in the context of compressed sensing of spectral images. Many modern spectral imagers face a limiting trade-off between spatial and spectral resolution, with the total number of voxels measured constrained by the size of the detector array. To mitigate this trade-off, many researchers have developed spectral imaging systems and associated reconstruction methods that are designed to exploit the theory of "compressive sensing" [19, 45, 42, 53]. One example physical system collects coded projections of each spectrum in the spectral image [19]. Using the novel multiscale representation of the spectral image based upon adaptive partitions as described in this paper, we are able to accurately reconstruct spectral images with an order of magnitude more reconstructed voxels than measurements.

7. Discussion.

In this paper, we presented an efficient multiscale algorithm for estimating spectral Poisson intensities. The key feature of our algorithm is that it adapts to varying degrees of smoothness in the spatial and the spectral directions, unlike 3d wavelet transform–

based approaches that assume the same degree of smoothness in all coordinate directions. Furthermore, our approach is backed by strong theoretical results which show that the proposed spatiospectral algorithm is near-minimax optimal on a certain anisotropic Hölder–Besov function class. The experimental results suggest that the algorithm performs very well even when there are significantly fewer photons than voxels. In many practical applications, the observed data are binned because of the inability of the reconstruction software to reconstruct the true intensity from such low photon counts. In contrast, our algorithm can offer improved estimation accuracy even when the observations are photon-limited. The near-optimality of our algorithm on a wide range of spectral intensities and the superior performance in photon-limited scenarios suggest that it can be an important component of several applications.

Appendix A. Proof of Theorem 3.1. For the expected squared Hellinger loss function, it can be shown [25, 51] that, if we consider all density estimates in a finite or countable family $\Gamma_{M,N}$ and if the penalty corresponding to every estimate $\widetilde{\boldsymbol{f}} \in \Gamma_{M,N}$ satisfies the Kraft inequality, then, by applying the Li–Barron theorem [31], it can be shown that

$$\mathbb{E}\left[\mathcal{H}^2(\boldsymbol{f}, \widehat{\boldsymbol{f}})\right] \leq \min_{\widetilde{\boldsymbol{f}} \in \Gamma_{M,N}} \left\{ \frac{1}{n} K\left(p_{\boldsymbol{f}} || p_{\widetilde{\boldsymbol{f}}}\right) + \frac{2}{n} \mathrm{pen}(\widetilde{\boldsymbol{f}}) \right\}, \tag{A.1}$$

where $K\big(p_{\boldsymbol{f}} || p_{\widetilde{\boldsymbol{f}}}\big) \equiv \sum_{\boldsymbol{y}} p(\boldsymbol{y}|\boldsymbol{f}) \log_e\big(\frac{p(\boldsymbol{y}|\boldsymbol{f})}{p(\boldsymbol{y}|\widetilde{\boldsymbol{f}})}\big)$ is the Kullback–Leibler (KL) divergence between densities $p_{\boldsymbol{f}}$ and $p_{\widetilde{\boldsymbol{f}}}$ following the proof techniques in [25, 51]. Here, we use the subscript notation $p_{\boldsymbol{f}}(\boldsymbol{y}) \equiv p(\boldsymbol{y}|\boldsymbol{f})$. The Li–Barron theorem shows that the performance of our estimator is bounded by the optimal trade-off between approximation error $\big(K\big(p_{\boldsymbol{f}} || p_{\widetilde{\boldsymbol{f}}}\big)\big)$ and the estimation error $\big(\mathrm{pen}(\widetilde{\boldsymbol{f}})\big)$. The penalty term can be thought of as a bound on estimation errors since it is proportional to the number of degrees of freedom in an estimate.

When the observations follow a multinomial distribution, the risk function $R(\boldsymbol{f}, \widehat{\boldsymbol{f}})$ can be bounded by [25]

$$R(\boldsymbol{f}, \widehat{\boldsymbol{f}}) \leq \min_{\widetilde{\boldsymbol{f}} \in \Gamma_{M,N}} \left\{ \frac{K\left(p_{\boldsymbol{f}} || p_{\widetilde{\boldsymbol{f}}}\right) + 2\mathrm{pen}(\widetilde{\boldsymbol{f}})}{n} \right\} \leq \min_{\widetilde{\boldsymbol{f}} \in \Gamma_{M,N}} \left\{ \frac{N^2 M}{C} \|\boldsymbol{f} - \widetilde{\boldsymbol{f}}\|_2^2 + \frac{2}{n} \mathrm{pen}(\widetilde{\boldsymbol{f}}) \right\}. \tag{A.2}$$

The squared ℓ_2 approximation error between the discrete densities $\boldsymbol{f}$ and $\widetilde{\boldsymbol{f}}$ can be upper bounded by the L_2 approximation error between the continuous densities f and $\widetilde{f}$, respectively:

$$\|\boldsymbol{f} - \widetilde{\boldsymbol{f}}\|_{\ell_2}^2 \leq \frac{1}{N^2 M} \|f - \widetilde{f}\|_{L_2}^2. \tag{A.3}$$

Expressing $\|\boldsymbol{f} - \widetilde{\boldsymbol{f}}\|_2^2$ in terms of $\|f - \widetilde{f}\|_2^2$ allows us to exploit the theoretical properties of the Hölder and Besov function class to bound the approximation error between f and $\widetilde{f}$. The approximation error between f and $\widetilde{f}$ can be bounded using the triangle inequality as

$$\|f - \widetilde{f}\|_{L_2}^2 \leq 2\|f - g\|_{L_2}^2 + 2\|g - \widetilde{f}\|_{L_2}^2, \tag{A.4}$$

where g is obtained by forming a spatial partition of cuboids with sidelengths at least $1/m$, where $m \leq N$ is a dyadic number. In particular, g has $\mathcal{O}(m^{2-\gamma})$ cuboids with sidelength $1/m$

along the Hölder boundary in f and $\mathcal{O}(m)$ larger cuboids apart from the boundary. Using techniques similar to those presented in [50, 29], it can be shown that the approximation error between f and g decays as

$$\|f - g\|_{L_2}^2 \preceq m^{-\alpha} + m^{-\gamma} \preceq m^{-\nu}, \tag{A.5}$$

where $\nu = \min(\alpha, \gamma)$ [29]. Given g, $\widetilde{f}$ is obtained by finding the best p-piece spectral RDP of each spectrum in g and fitting polynomials to every partition cell. The second term on the right-hand side of (A.4) can be determined using techniques described in [51, 29] based on free-knot polynomial approximation errors [9] to yield

$$\|g(x_1, x_2, \cdot) - \widetilde{f}(x_1, x_2, \cdot)\|_{L_2} \preceq p^{-\beta} + \left(\frac{p}{M}\right)^{1/2} + n^{-1/2}. \tag{A.6}$$

The approximation error in (A.6) corresponds to a single spectrum. Integrating (A.6) over the spatial domain of observations $[0, 1]^2$, we have

$$\|g - \widetilde{f}\|_{L_2}^2 \preceq p^{-2\beta} + \frac{p}{M} + \frac{1}{n}. \tag{A.7}$$

Substituting (A.5) and (A.7) into (A.4),

$$\|f - \widetilde{f}\|_{L_2}^2 \preceq m^{-\nu} + p^{-2\beta} + \frac{p}{M} + \frac{1}{n}. \tag{A.8}$$

Now we show that the penalty corresponding to the density estimates in class $\Gamma_{M,N}$ as defined in (2.2) satisfies the Kraft inequality. Consider an estimate $\widetilde{\boldsymbol{f}} \in \Gamma_{M,N}$, which, according to our definition, is piecewise constant in the two spatial dimensions and piecewise polynomial in the spectral dimension. In a tree representation, it can be represented as a quadtree having binary splits in each of its leaves. We arrive at the penalty for the estimate $\widetilde{\boldsymbol{f}}$ by computing the number of bits needed to uniquely represent $\widetilde{\boldsymbol{f}}$. The structure of a quadtree with k leaves can be uniquely encoded using at most $4k/3$ bits [51]. Each of the k leaves of the quadtree is associated with a binary tree which can also be encoded using a prefix code, as described in [51]. In particular, a binary tree with p leaves consists of $2p - 1$ nodes, and hence can be encoded uniquely by at most $2p - 1$ bits, which can be verified by induction [51]. Let $\mathcal{P}$ be the spatial partition, and for each $c \in \mathcal{P}$ let p_c be the number of cells in the spectral RDP. Then the number of bits required to encode the structure of the spatiospectral RDP $\mathcal{Q}$ is

$$\frac{4}{3}k + \sum_{c \in \mathcal{P}} (2p_c - 1) \leq \frac{4}{3}k + 2|\mathcal{Q}| \leq \frac{10}{3}|\mathcal{Q}|.$$

Assuming that each polynomial coefficient is quantized to one of $\sqrt{n}$ levels, $\log_2 \sqrt{n}$ bits are needed to encode each coefficient in the partition $\mathcal{Q}$, and the total number of bits required to prefix encode the estimate is $|\mathcal{Q}| \left(\frac{10}{3} + \frac{r}{2} \log_2 n\right)$. Thus the penalty in (2.2) corresponds to a prefix codelength and is guaranteed to satisfy

$$\sum_{\widetilde{\boldsymbol{f}} \in \Gamma_{M,N}} 2^{-|\mathcal{Q}(\widetilde{\boldsymbol{f}})| \left(\frac{10}{3} + \frac{r}{2} \log_2 n\right)} \leq 1$$

or

$$\sum_{\widetilde{\boldsymbol{f}} \in \Gamma_{M,N}} e^{-|\mathcal{Q}(\widetilde{\boldsymbol{f}})|\left(\frac{10}{3}+\frac{r}{2}\log_2 n\right)\log_e 2} \leq 1,$$

as desired.

The approximation error in (A.8) assumed $k = \mathcal{O}\left(m^{2-\gamma}\right)$ cells in the spatial partition and a maximum of p cells in each spectral partition, for a total of at most kp cells. It further assumed that the spectral approximation was a free-knot piecewise polynomial. However, in the RDP construction, the knots are restricted to lie on dyadic interval endpoints. A p-piece piecewise polynomial can be subdivided into a $(p \log_2 M)$-piece piecewise polynomial with the same approximation error but with knots on dyadic interval endpoints. Thus the RDP corresponding to the approximation error in (A.8) has at most $kp \log_2 M$ cells, so its penalty is bounded by

$$\begin{aligned} \operatorname{pen}(\widetilde{\boldsymbol{f}}) &\leq kp \log_2 M \left(\frac{10}{3} + \frac{r}{2}\log_2 n\right) \log_e 2 \\ &< kp \log_2 M \left(\frac{10}{3} + \frac{r}{2}\right) \log_2 n \log_e 2 = kp \log_2 M \left(\frac{10}{3} + \frac{r}{2}\right) \log_e n. \end{aligned} \tag{A.9}$$

Applying (A.3), (A.8), and (A.9) to (A.2), the risk function can be rewritten as follows:

$$R(\boldsymbol{f}, \widehat{\boldsymbol{f}}) \preceq \min_{m,p} \left\{ \left(m^{-\nu} + p^{-2\beta} + \frac{p}{M} + \frac{1}{n}\right) + \frac{\left(\frac{10}{3}+\frac{r}{2}\right)}{n} m^{2-\gamma} p \log_2 M \log_e n \right\}.$$

The values of m and p that minimize the risk function given above can be shown to be

$$m = C' \left(\frac{n}{\log_2 M \log_e n}\right)^{\frac{2\beta}{2\beta(\nu+2-\gamma)+\nu}} \quad \text{and} \quad p = C' \left(\frac{n}{\log_2 M \log_e n}\right)^{\frac{\nu}{2\beta(\nu+2-\gamma)+\nu}},$$

where C' is a constant independent of n, N, and M and dependent on the smoothness parameters α, β, and γ. Thus, if we collect the measurements on a grid of size $N \times N \times M$, where $N \geq m$ and $M \geq p$, and estimate the true intensity $\boldsymbol{f}$ by the estimation procedure given in section 2, then the risk function between the true and the estimated density is upper bounded by

$$R(\boldsymbol{f}, \widehat{\boldsymbol{f}}) \preceq \left(\frac{n}{\log_2 M \log_e n}\right)^{\frac{-2\beta\nu}{2\beta(\nu+2-\gamma)+\nu}},$$

which is within a log factor of the lower bound when $\gamma = 1$.

Appendix B. Proof of Theorem 3.2. The proof of Theorem 3.2 involves finding a suitable packing set for the density class $\mathcal{G} \subseteq \mathcal{F}$ and bounding the packing entropy. Once a bound on the packing entropy is obtained, the lower bound on the risk function can be obtained using the relation (3.9).

B.1. Bound on the Kolmogorov ϵ-capacity. We will show that the L_2 packing numbers of $\mathcal{F}_{\alpha,(\beta,p,q)}(L)$ satisfy

$$K_\epsilon(\mathcal{F}_{\alpha,(\beta,p,q)}(L)) \succeq \epsilon^{-\frac{2\beta+\alpha}{\alpha\beta}} \tag{B.1}$$

for L large enough. The first step in arriving at (B.1) is to construct an appropriate packing set for $\mathcal{F}_{\alpha,(\beta,p,q)}(L)$. To this end, let us choose functions $\phi : [0,1]^2 \to \mathbb{R}$ and $\psi : [0,1] \to \mathbb{R}$ such that the following hold:

1. ϕ and ψ are bounded, respectively, by some C_ϕ and C_ψ;
2. $\|\phi\|_{L_2} = 1$ and $\|\psi\|_{L_2} = 1$;
3. ϕ is α-Hölder on $[0,1]^2$ with constant C_α: $\sup_{t>0}(t^{-\alpha}\omega(\phi,t)) = C_\alpha < \infty$;
4. ψ belongs to the Besov space $B_q^\beta(L_p[0,1])$:

$$|\psi|_{B_q^\beta(L_p)} = \left(\int_0^1 \left[t^{-\beta}\omega_r(\psi,t)_p\right]^d \frac{dt}{t}\right)^{1/q} = C_\beta < \infty.$$

Now let j and m denote fixed positive integers (to be chosen later), and for each $k = (k_1,k_2,k_3)$, $k_i = 0,1,2,\ldots$, define a function

$$\begin{aligned}\xi_k(x_1,x_2,\lambda) &\triangleq 2^{j+m/2}\phi_{j,k_1,k_2}(x_1,x_2)\psi_{m,k_3}(\lambda)\\ &\equiv 2^{j+m/2}\phi(2^j x_1 - k_1, 2^j x_2 - k_2)\psi(2^m\lambda - k_3).\end{aligned}$$

Let S_k denote the support of ξ_k. It is easy to see from definitions that $\|\xi_k\|_{L_2} = 1$ and $S_k \cap S_{k'} = \varnothing$ for $k \neq k'$. Let

$$\mathcal{S}_{j,m} \triangleq \left\{k = (k_1,k_2,k_3) : 0 \le k_1,k_2 \le 2^j - 1, 0 \le k_3 \le 2^m - 1\right\},$$

and for each $\Lambda \subseteq \mathcal{S}_{j,m}$ define the function $f_\Lambda \triangleq \frac{\epsilon}{\sqrt{J}}\sum_{k\in\Lambda}\xi_k$, where $J = 2^{2j+m}$. Note that $\bigcup_{k\in\mathcal{S}_{j,m}} S_k = [0,1]^3$. We first show that there exists some $L > 0$ (independent of ϵ) such that $\{f_\Lambda\} \subseteq \mathcal{F}_{\alpha,(\beta,p,q)}(L)$. Since the ξ_k have disjoint support, we have

$$\|f_\Lambda\|_{L_2}^2 = \frac{\epsilon^2}{J}\left\|\sum_{k\in\Lambda}\xi_k\right\|_{L_2}^2 = \frac{\epsilon^2}{J}\sum_{k\in\Lambda}\|\xi_k\|_{L_2}^2 \le \epsilon^2, \tag{B.2}$$

where we also used the fact that $\|\xi_k\|_{L^2} = 1$ for all k. Next, again using the fact that the ξ_k have disjoint support, we get

$$\begin{aligned}\omega(f_\Lambda(\cdot,\lambda),t) &= \sup_{h\in\mathbb{R}^2;\|h\|\le t}\|\Delta_h^1 f_\Lambda(\cdot,\lambda)\|_{L_\infty}\\
&= \sup_{h\in\mathbb{R}^2;\|h\|\le t}\ \sup_{(x_1,x_2)\in[0,1]^2}|f_\Lambda(x_1+h_1,x_2+h_2,\lambda) - f_\Lambda(x,y,\lambda)|\\
&= \frac{\epsilon}{\sqrt{J}}\sup_{h\in\mathbb{R}^2;\|h\|\le t}\ \sup_{(x_1,x_2)\in[0,1]^2}\left|\sum_{k\in\Lambda}\left(\xi_k(x_1+h_1,x_2+h_2,\lambda) - \xi_k(x_1,x_2,\lambda)\right)\right|\\
&\le \frac{\epsilon}{\sqrt{J}}\max_{k,k'\in\Lambda}\ \sup_{\substack{(x_1,x_2,\lambda),(x_1',x_2',\lambda)\in S_k\cup S_{k'}\\ \|(x_1,x_2)-(x_1',x_2')\|\le t}}\left\{\left|\xi_k(x_1,x_2,\lambda) - \xi_k(x_1',x_2',\lambda)\right| + \left|\xi_{k'}(x_1,x_2,\lambda) - \xi_{k'}(x_1',x_2',\lambda)\right|\right\}\\
&\le \frac{\epsilon}{\sqrt{J}}\max_{k,k'\in\Lambda}\left(\omega(\xi_k(\cdot,\lambda),t) + \omega(\xi_{k'}(\cdot,\lambda),t)\right) \le \frac{\epsilon}{\sqrt{J}}\cdot C_\psi\cdot 2\sqrt{J}\omega(\phi,2^j t) = 2C_\psi\epsilon\omega(\phi,2^{-j}t).\end{aligned}$$

Therefore,

$$|f_\Lambda|^{(1,2)}_{H_\alpha} = \sup_{0\le\lambda\le 1}\sup_{t>0}\left(t^{-\alpha}\omega\left(f_\Lambda(\cdot,\lambda),t\right)\right) \le 2C_\psi\epsilon\sup_{t>0}\left(t^{-\alpha}\omega(\phi,2^j t)\right)$$
$$= 2^{j\alpha+1}C_\psi\epsilon\sup_{t>0}\left(t^{-\alpha}\omega(\phi,t)\right) = 2^{j\alpha+1}C_\psi C_\alpha\epsilon.$$

Finally, we deal with the Besov norm along the third coordinate direction. We split the integral in (3.6) as follows:

$$|f_\Lambda|^{(3)}_{B^\beta_q(L_p)} \le 2^{1/q}\sup_{(x,y)\in[0,1]^2}\left(\int_0^{2^{-m}}\left[t^{-\beta}\omega_r(f_\Lambda(x,y,\cdot),t)_p\right]^q\frac{dt}{t}\right)^{1/q}$$
$$+\,2^{1/q}\sup_{(x,y)\in[0,1]^2}\left(\int_{2^{-m}}^1\left[t^{-\beta}\omega_r(f_\Lambda(x,y,\cdot),t)_p\right]^q\frac{dt}{t}\right)^{1/q} \equiv T_1+T_2.$$

First let us deal with T_1. Let us fix some $(x_1,x_2)\in[0,1]^2$, and for each $\Lambda\subseteq\mathcal{S}_{j,m}$ define the set $\Lambda_{x_1,x_2}\triangleq\left\{k=(k_1,k_2,k_3)\in\Lambda\,\middle|\,\exists\lambda \text{ s.t. } (x_1,x_2,\lambda)\in S_k\right\}$. Then there exist some $0\le k_1(x_1),k_2(x_2)\le 2^j-1$ such that $\Lambda_{x_1,x_2}=\left\{(k_1(x_1),k_2(x_2),k_3)\,\middle|\,0\le k_3\le 2^m-1\right\}$. Now, given some $t\le 2^{-m}$ and $0<h\le t$, we have

$$\|\Delta_h^r f_\Lambda(x_1,x_2,\cdot)\|^p_{L_p} = \frac{\epsilon^p}{J^{p/2}}\int\left|\sum_{k\in\Lambda}J^{1/2}\phi_{j,k_1,k_2}(x_1,x_2)\Delta_h^r\psi_{m,k_3}(\lambda)\right|^p d\lambda$$
$$\le \epsilon^p\sum_{k_3=0}^{2^m-1}|\phi_{j,k_1(x_1),k_2(x_2)}(x_1,x_2)|^p\int_{\mathrm{supp}(\Delta_h^r\psi_{m,k_3})}|\Delta_h^r\psi_{m,k_3}(\lambda)|^p\,d\lambda$$
$$\le \epsilon^p C_\phi^p\sum_{k_3=0}^{2^m-1}\int|\Delta_h^r\psi_{m,k_3}(\lambda)|^p d\lambda = (C_\phi\epsilon)^p\int|\Delta^r_{2^m h}\psi(2^m\lambda)|^p d\lambda$$
$$= (C_\phi\epsilon)^p 2^{-m}\int|\Delta^r_{2^m h}\psi(\lambda)|^p d\lambda \le (C_\phi\epsilon)^p\|\Delta^r_{2^m h}\psi\|^p_{L_p}.$$

Hence, we have for the modulus of continuity

$$\omega_r(f_\Lambda(x_1,x_2,\cdot),t) = \sup_{0<h\le t}\|\Delta_h^r f_\Lambda(x_1,x_2,\cdot)\|_{L_p} \le C_\phi\epsilon\sup_{0<h\le t}\|\Delta^r_{2^m h}\psi\|_{L_p} = C_\phi\epsilon\omega_r(\psi,2^m t)_p.$$

Therefore,

$$T_1 = 2^{1/q}\sup_{(x_1,x_2)\in[0,1]^2}\left(\int_0^{2^{-m}}\left[t^{-\beta}\omega_r(f_\Lambda(x_1,x_2,\cdot),t)_p\right]^q\frac{dt}{t}\right)^{1/q}$$
$$\le 2^{1/q}\epsilon C_\phi\left(\int_0^{2^{-m}}\left[t^{-\beta}\omega_r(\psi,2^m t)_p\right]^q\frac{dt}{t}\right)^{1/q} = 2^{1/q}\epsilon C_\phi 2^{m\beta}\left(\int_0^1\left[t^{-\beta}\omega_r(\psi,t)_p\right]^q\frac{dt}{t}\right)^{1/q}$$
$$= 2^{1/q}2^{m\beta}C_\phi C_\beta\epsilon.$$

Moreover, since ξ_k have disjoint support we have

$$|\Delta^r_{(0,0,h)} f_\Lambda(x_1,x_2,\lambda)| \le \frac{\epsilon}{\sqrt{J}} \sum_{k\in\Lambda} 1_{\{(x_1,x_2,\lambda)\in S_k\}} \left|\Delta^r_{(0,0,h)} \xi_k(x_1,x_2,\lambda)\right.$$
$$= \epsilon \sum_{k\in\Lambda} 1_{\{(x_1,x_2,\lambda)\in S_k\}} |\phi(2^j x_1 - k_1, 2^j x_2 - k_2)| \cdot |\Delta^r_{2^m h}\psi(2^m\lambda - k_3)| \le \epsilon 2^r C_\phi C_\psi.$$

Hence,

$$T_2 = 2^{1/q}\left(\int_{2^{-m}}^1 \left[t^{-\beta}\omega_r^{(3)}(f_\Lambda,t)_p\right]^q \frac{dt}{t}\right)^{1/q} \le 2^{1/q}\epsilon 2^r C_\phi C_\psi \left(\int_{2^{-m}}^1 t^{-(\beta q+1)}dt\right)^{1/q}$$
$$= 2^{1/q}\epsilon 2^r C_\phi C_\psi \left(\frac{2^{m\beta q}-1}{\beta q}\right)^{1/q} \le (\beta q)^{-1} 2^{1/q}\epsilon 2^r C_\phi C_\psi 2^{m\beta}.$$

Now let us choose j and m so that $2^j \asymp \epsilon^{-1/\alpha}$ and $2^m \asymp \epsilon^{-1/\beta}$. Then we will have $\max_{\Lambda\subseteq\mathcal{S}_{j,m}} |f_\Lambda|_{H^{(1,2)}_\alpha} \preceq 1$ and $\max_{\Lambda\subseteq\mathcal{S}_{j,m}} |f_\Lambda|_{B^{\beta,(3)}_q(L_p)} \preceq 1$, where the constants hidden in the asymptotic order notation depend on the choice of ϕ and ψ, as well as on α, β, p, and q.

We now extract an ϵ-packing set from the collection $\{f_\Lambda\}$. To this end, we will use a result from information theory known as the Gilbert–Varshamov bound (see, e.g., Lemma 2.7.4 in [28]): we can choose at least $2^{J/2}$ subsets Λ_i of $\mathcal{S}_{j,m}$ such that $|\Lambda_i \triangle \Lambda_j| \ge J/4$. For each of these Λ_i's, let us denote the corresponding f_{Λ_i} by f_i. Then

$$\|f_i - f_j\|^2_{L_2} = \frac{\epsilon^2}{J} \sum_{k\in\Lambda_i\triangle\Lambda_j} \|\xi_k\|^2_{L_2} = \frac{\epsilon^2}{J}|\Lambda_i\triangle\Lambda_j| \ge \frac{\epsilon^2}{4}.$$

Hence, the functions f_i are $\epsilon/2$-separated in the L_2 sense, and so $\{f_i\}$ is an $\epsilon/2$-packing set for the corresponding function class $\mathcal{F}_{\alpha,(\beta,p,q)}(L)$. Thus, $M(\epsilon/2, \mathcal{F}_{\alpha,(\beta,p,q)}(L)) \ge 2^{J/2}$ or

$$K_{\epsilon/2}(\mathcal{F}_{\alpha,(\beta,p,q)}(L)) = \log M(\epsilon/2, \mathcal{F}_{\alpha,(\beta,p,q)}(L)) \ge J/2 \succeq \epsilon^{-(2/\alpha+1/\beta)} = \epsilon^{-(2\beta+\alpha)/(\alpha\beta)}.$$

B.2. Minimax lower bound. To arrive at the lower bound we consider a subset $\mathcal{G} \subseteq \mathcal{F}$, which consists of all functions $g : [0,1]^3 \to [0,\infty)$ satisfying the following conditions:

1. There exist constants $0 < C_\ell, C_u < +\infty$ such that for every $g \in \mathcal{G}$ we have
$$C_\ell \le g(x_1,x_2,\lambda) \le C_u \qquad \forall (x_1,x_2,\lambda) \in [0,1]^3.$$
2. There exist constants $\alpha,\gamma \in (0,1]$, $c_\gamma > 0$, $\beta \in (1,2]$ such that each $g \in \mathcal{G}$ has the form
$$g(x_1,x_2,\lambda) = g_1(x_1,x_2,\lambda) I_{\{H(x_1)\ge x_2\}} + g_2(x_1,x_2,\lambda) I_{\{H(x_1)<x_2\}},$$
where H is Hölder-γ with constant c_γ, and g_1 and g_2 are both in $\mathcal{F}_{\alpha,(\beta,p,p)}(L)$ with $1/p = 1/\beta + 1/2$ and a large but finite L (which implicitly depends on C_ℓ and C_u).
3. There exists a constant $0 < h < 1/2$ such that $1/2 - h \le |H(x)| \le 1/2 + h$ for all $x \in [0,1]$.

The last constraint described above restricts the Hölder-γ boundary that separates the two Hölder-α surfaces to lie within a small strip of height $2h < 1$ centered on the line $x_2 = 1/2$.

We now establish the following result:

$$R_n(\mathcal{G}) \succeq \max\left\{n^{-2\gamma/(2\gamma+1)}, n^{-2\alpha\beta/(2\alpha\beta+2\beta+\alpha)}\right\}.$$

The reasoning is similar to the proof of Theorem 3.3.7 in [28]. Let us first consider the subset $\mathcal{G}_0$ of $\mathcal{G}$ consisting of functions that are equal to C_ℓ below the boundary and to C_u above the boundary. Then $R_n(\mathcal{G}_0) \asymp R_n(H_\gamma(c_\gamma)) \succeq n^{-2\gamma/(2\gamma+1)}$. In addition, since the boundary lies within the strip $[0,1]\times[1/2-h, 1/2+h]$ with $0 < h < 1/2$, we have $R_n(\mathcal{G}) \succeq R_n(\mathcal{F}_{\alpha,(\beta,p,p)}(L))$, where the constant hidden in the asymptotic order notation depends on h. We can determine the minimax lower bound on $R_n(\mathcal{F}_{\alpha,(\beta,p,p)}(L))$ using the Yang–Barron method [54]: $R_n(\mathcal{F}_{\alpha,(\beta,p,p)}(L)) \succeq \epsilon_n^2$, where ϵ_n^2 is the solution of the equation $K_{\epsilon_n}(\mathcal{F}_{\alpha,(\beta,p,p)}(L)) = n\epsilon_n^2$. This yields $R_n(\mathcal{F}_{\alpha,(\beta,p,p)}(L)) \succeq n^{-2\alpha\beta/(2\alpha\beta+2\beta+\alpha)}$. Overall,

$$R_n(\mathcal{G}) \succeq \max\left\{n^{-2\gamma/(2\gamma+1)}, n^{-2\alpha\beta/(2\alpha\beta+2\beta+\alpha)}\right\}.$$

Now, let us consider the case $\gamma = 1$. Then the minimax lower bound becomes

$$R_n(\mathcal{G}) \succeq \max\left\{n^{-2/3}, n^{-2\nu\beta/(2\nu\beta+2\beta+\nu)}\right\},$$

since $\nu = \min\{\alpha, \gamma\} = \alpha$. Thus, the upper bound given in (3.8) matches the lower bound up to a log factor.

REFERENCES

[1] A. Antoniadis and T. Sapatinas, *Wavelet shrinkage for natural exponential families with quadratic variance functions*, Biometrika, 88 (2001), pp. 805–820.

[2] I. Atkinson, F. Kamalabadi, and D. L. Jones, *Wavelet-based hyperspectral image estimation*, in Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, 2003, IEEE Press, Piscataway, NJ, Vol. 2, pp. 743–745.

[3] A. Buades, B. Coll, and J.-M. Morel, *A non-local algorithm for image denoising*, in Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), Washington, DC, 2005, IEEE Computer Society Press, Piscataway, NJ, Vol. 2, pp. 60–65.

[4] E. Candès and D. Donoho, *Curves and surface fitting*, in Curvelets: A Surprisingly Effective Nonadaptive Representation for Objects with Edges, A. Cohen, C. Rabut, and L. L. Schumaker, eds., Vanderbilt University Press, Nashville, TN, 2000, pp. 105–120.

[5] E. Candès, J. Romberg, and T. Tao, *Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information*, IEEE Trans. Inform. Theory, 52 (2006), pp. 489–509.

[6] E. Candés and T. Tao, *Near optimal signal recovery from random projections: Universal encoding strategies*, IEEE Trans. Inform. Theory, 52 (2006), pp. 5406–5425.

[7] R. Coifman and D. Donoho, *Translation invariant de-noising*, in Lecture Notes in Statistics: Wavelets and Statistics, Springer-Verlag, New York, 1995, pp. 125–150.

[8] T. Cover and J. Thomas, *Elements of Information Theory*, 2nd ed., Wiley, New York, 2006.

[9] R. A. DeVore, *Nonlinear approximation*, Acta Numer., 7 (1998), pp. 51–150.

[10] D. Donoho, *Nonlinear wavelet methods for recovery of signals, densities, and spectra from indirect and noisy data*, in Proceedings of Symposia in Applied Mathematics, American Mathematical Society, Providence, RI, 1993, pp. 173–205.

[11] D. Donoho and I. Johnstone, *Minimax estimation via wavelet shrinkage*, Ann. Statist., 26 (1998), pp. 879–921.
[12] D. Donoho, *De-noising by soft-thresholding*, IEEE Trans. Inform. Theory, 41 (1995), pp. 613–627.
[13] D. Donoho, *Compressed sensing*, IEEE Trans. Inform. Theory, 52 (2006), pp. 1289–1306.
[14] D. Donoho and I. Johnstone, *Ideal spatial adaptation by wavelet shrinkage*, Biometrika, 81 (1994), pp. 425–455.
[15] L. Dragotti, G. Poggi, and A. Ragozini, *Compression of multispectral images by three-dimensional SPIHT algorithm*, IEEE Trans. Geosci. Remote Sensing, 38 (2000), pp. 416–428.
[16] F. J. Anscombe, *The transformation of Poisson, binomial, and negative-binomial data*, Biometrika, 15 (1948), pp. 246–254.
[17] P. Fryzlewicz, *Data-driven wavelet-Fisz methodology for nonparametric function estimation*, Electron. J. Statist., 2 (2008), pp. 863–896.
[18] P. Fryzlewicz and G. Nason, *A Haar-Fisz algorithm for Poisson intensity estimation*, J. Comput. Graph. Statist., 13 (2004), pp. 621–638.
[19] M. E. Gehm, R. John, D. J. Brady, R. M. Willett, and T. J. Schulz, *Single-shot compressive spectral imaging with a dual-disperser architecture*, Opt. Express, 15 (2007), pp. 14013–14027.
[20] M. Jansen, *Multiscale Poisson data smoothing*, J. R. Stat. Soc. Ser. B Stat. Methodol., 68 (2006), pp. 27–48.
[21] C. Kervrann and A. Trubuil, *An adaptive window approach for Poisson noise reduction and structure preserving in confocal microscopy*, in Proceedings of the IEEE International Symposium on Biomedical Imaging: From Nano to Macro, Arlington, VA, 2004, IEEE Press, Piscataway, NJ, pp. 788–791.
[22] E. Kolaczyk, *Bayesian multi-scale models for Poisson processes*, J. Amer. Statist. Assoc., 94 (1999), pp. 920–933.
[23] E. Kolaczyk and D. Dixon, *Nonparametric estimation of intensity maps using Haar wavelets and Poisson noise characteristics*, Astrophys. J., 534 (2000), pp. 490–505.
[24] E. Kolaczyk and R. Nowak, *Multiscale generalised linear models for nonparametric function estimation*, Biometrika, 92 (2005), pp. 119–133.
[25] E. Kolaczyk and R. Nowak, *Multiscale likelihood analysis and complexity penalized estimation*, Ann. Statist., 32 (2004), pp. 500–527.
[26] E. D. Kolaczyk, *Wavelet shrinkage estimation of certain Poisson intensity signals using corrected thresholds*, Statist. Sinica, 9 (1999), pp. 119–135.
[27] A. Kolmogorov and V. Tihomirov, *ϵ-entropy and ϵ-capacity of sets in function spaces*, Uspekhi Mat. Nauk, 14 (1959), pp. 3–86.
[28] A. P. Korostelev and A. B. Tsybakov, *Minimax Theory of Image Reconstruction*, Springer-Verlag, New York, 1993.
[29] K. Krishnamurthy and R. Willett, *Multiscale reconstruction of photon-limited hyperspectral data*, in Proceedings of the IEEE Statistical Signal Processing Workshop, Madison, WI, 2007, IEEE Press, Piscataway, NJ, 2007, pp. 596–600.
[30] M. Lang, H. Guo, J. E. Odegard, C. S. Burrus, and R. O. Wells, *Noise reduction using an undecimated discrete wavelet transform*, IEEE Signal Process. Lett., 3 (1996), pp. 10–12.
[31] J. Li and A. Barron, *, Mixture density estimation*, in Advances in Neural Information Processing Systems 12, Denver, CO, 1999, MIT Press, Cambridge, MA, 2000, pp. 279–285.
[32] R. P. Lin, B. R. Dennis, and A. O. Benz, eds., *The Reuven Ramaty High-Energy Solar Spectroscopic Imager (RHESSI)—Mission Description and Early Results*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2003.
[33] L. B. Lucy, *An iterative technique for the rectification of observed distributions*, Astron. J., 79 (1974), pp. 745–754.
[34] J. V. Manjón, M. Robles, and N. Thacker, *Multispectral MRI de-noising using non-local means*, in Proceedings of the Conference on Medical Image Understanding and Analysis (MIUA), University of Wales Aberystwyth, BMVA, Worcs, UK, 2007, pp. 41–45.
[35] J. Miller, C. Elvidge, B. Rock, and J. Freemantle, *An airborne perspective on vegetation phenology from the analysis of avris data sets over the jasper ridge biological preserve*, in Proceedings of the Geoscience and Remote Sensing Symposium (IGARSS'90): Remote Sensing for the Nineties, 1990, IEEE Press, Piscataway, NJ, pp. 565–568.

[36] M. H. Neumann, *Multivariate wavelet thresholding in anisotropic function spaces*, Statist. Sinica, 10 (2000), pp. 399–431.
[37] R. Nowak and E. Kolaczyk, *A multiscale MAP estimation method for Poisson inverse problems*, in Proceedings of the 32nd Asilomar Conference on Signals, Systems, and Computing, Pacific Grove, CA, 1998, IEEE Press, Piscataway, NJ, Vol. 2, pp. 1682–1686.
[38] R. Nowak and E. Kolaczyk, *A statistical multiscale framework for Poisson inverse problems*, IEEE Trans. Inform. Theory, 46 (2000), pp. 1811–1825.
[39] W. Richardson, *Bayesian-based iterative method of image restoration*, J. Opt. Soc. Amer., 62 (1972), pp. 55–59.
[40] S. Sardy, A. Antoniadis, and P. Tseng, *Automatic smoothing with wavelets for a wide class of distributions*, J. Comput. Graph. Statist., 13 (2004), pp. 399–421.
[41] P. Scheunders, *Denoising of multispectral images using wavelet thresholding*, in Proceedings of the SPIE Conference on Image and Signal Processing for Remote Sensing IX, Barcelona, Spain, 2004, SPIE, Bellingham, WA, 2004, Vol. 5238, pp. 28–35.
[42] D. Takhar, J. N. Laska, M. B. Wakin, M. F. Duarte, D. Baron, S. Sarvotham, K. Kelly, and R. G. Baraniuk, *A new compressive imaging camera architecture using optical-domain compression*, in SPIE, 2006, Vol. 6065, paper 606509.
[43] J. Timlin, M. Sinclair, D. Haaland, M. Martinez, M. Manginell, S. Brozik, J. Guzowski, and M. Werner-Washburne, *Hyperspectral imaging of biological targets: The difference a high resolution spectral dimension and multivariate analysis can make*, in Proceedings of the IEEE International Symposium on Biomedical Imaging: From Nano to Macro, Arlington, VA, 2004, IEEE Press, Piscataway, NJ, 2004, pp. 1529–1532.
[44] K. Timmermann and R. Nowak, *Multiscale modeling and estimation of Poisson processes with application to photon-limited imaging*, IEEE Trans. Inform. Theory, 45 (1999), pp. 846–862.
[45] A. Wagadarikar, R. John, R. Willett, and D. Brady, *Single disperser design for coded aperture snapshot spectral imaging*, Appl. Optim., 47 (2008), pp. B44–B51.
[46] R. Willett, *Multiscale-analysis of photon-limited astronomical images*, in Statistical Challenges in Modern Astronomy IV, Astronom. Soc. Pacific Conf. Ser. 371, ASP, San Francisco, CA, 2007, pp. 247–264.
[47] R. Willett, *Multiscale intensity estimation for multi-photon microscopy*, in Proceedings of the IEEE International Symposium on Biomedical Imaging: From Nano to Macro, Washington, DC, 2007, IEEE Press, Piscataway, NJ, 2007, pp. 484–487.
[48] R. Willett, *Multiscale intensity estimation for marked Poisson processes*, in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Honolulu, HI, 2007, IEEE Press, Piscataway, NJ, 2007, Vol. 3, pp. 1249–1252.
[49] R. Willett and R. Nowak, *Fast multiresolution photon-limited image reconstruction*, in Proceedings of the IEEE International Symposium on Biomedical Imaging: From Nano to Macro, Arlington, VA, IEEE Press, Piscataway, NJ, 2004, Vol. 2, pp. 1192–1195.
[50] R. Willett and R. Nowak, *Platelets: A multiscale approach for recovering edges and surfaces in photon-limited medical imaging*, IEEE Trans. Med. Imaging, 22 (2003), pp. 332–350.
[51] R. Willett and R. Nowak, *Multiscale Poisson intensity and density estimation*, IEEE Trans. Inform. Theory, 53 (2007), pp. 3171–3187.
[52] R. M. Willett, M. E. Gehm, and D. J. Brady, *Multiscale reconstruction for computational spectral imaging*, in SPIE 2007, San Jose, CA, C. A. Bouman, E. L. Miller, and I. Pollak, eds., Vol. 6498, 2007, paper 64980L.
[53] F. Woolfe, M. Maggioni, G. Davis, F. Warner, R. Coifman, and S. Zucker, *Hyper-spectral Microscopic Discrimination between Normal and Cancerous Colon Biopsies*, manuscript, 2006.
[54] Y. Yang and A. Barron, *Information-theoretic determination of minimax rates of convergence*, Ann. Statist., 27 (1999), pp. 1564–1599.
[55] B. Zhang, J. Fadili, and J. Starck, *Wavelets, ridgelets, and curvelets for Poisson noise removal*, IEEE Trans. Image Process., 17 (2008), pp. 1093–1108.