A Statistical Multiscale Framework for Poisson Inverse Problems

Robert D. Nowak

Department of Electrical and Computer Engineering
Rice University, MS-380, P.O. Box 1892
Houston, Texas 77251-1892 USA

Fax: (+1 713) 737-6196
Email: nowak@rice.edu
Web: www.ece.rice.edu/~nowak

and *Eric D. Kolaczyk*

Department of Mathematics and Statistics
Boston University
Boston, MA 02215

Email: kolaczyk@math.bu.edu
Web: math.bu.edu/people/kolaczyk/

Accepted for the *Special Issue of the*
IEEE TRANSACTIONS ON INFORMATION THEORY
on Information-Theoretic Imaging

Abstract

This paper describes a statistical modeling and analysis method for linear inverse problems involving Poisson data based on a novel multiscale framework. The framework itself is founded upon a multiscale analysis associated with recursive partitioning of the underlying intensity, a corresponding multiscale factorization of the likelihood (induced by this analysis), and a choice of prior probability distribution made to match this factorization by modeling the "splits" in the underlying partition. The class of priors used here has the interesting feature that the "non-informative" member yields the traditional maximum likelihood solution; other choices are made to reflect prior belief as to the smoothness of the unknown intensity. Adopting the expectation-maximization (EM) algorithm for use in computing the MAP estimate corresponding to our model, we find that our model permits remarkably simple, closed-form expressions for the EM update equations. The behavior of our EM algorithm is examined, and it is shown that convergence to the global MAP estimate can be guaranteed. Applications in emission computed tomography and astronomical energy spectral analysis demonstrate the potential of the new approach.

Index Terms: Poisson processes, inverse problems, multiscale analysis, imaging

---

This work was supported by the National Science Foundation, grant no. MIP–9701692, the Office of Naval Research, award nos. N00014-99-1-0219 and N00014-00-1-0390, and the Army Research Office, grant no. DAAD19-99-1-0349.

I. Introduction

Many problems in science and engineering involve the recovery of an object (intensity) from indirect Poisson data (counts); that is, Poisson data are collected whose underlying intensity function is indirectly related to an object of interest through a linear system of equations. High-energy astronomical imaging [1] and emission computed tomographic imaging [2] are just two examples. We call all these problems *Poisson inverse problems.*

Poisson inverse problems can be very "ill-posed" [3], in the sense that small perturbations in the data can lead to dramatically different solutions to the recovery problem by a given method. Thus, solving these problems is especially challenging in low signal-to-noise ratio (SNR) situations when the total number of counts observed is limited, as is the situation in many photon imaging modalities. Maximum likelihood is one such method of estimation, and now a fairly standard one, that is not exempt from the effects of the ill-posed nature of the problem. As a result, more recent treatments of Poisson inverse problems have involved maximizing a criterion based on the likelihood equations augmented with an appropriate regularization or penalization term that stabilizes the otherwise ill-conditioned likelihood criterion. Often the regularization term takes the form of a Bayesian prior; the Maximum *a Posteriori* (MAP) estimator is then used in place of the Maximum Likelihood Estimator (MLE).

Wavelet and multiscale analysis and regularization methods have recently received considerable attention in the information theory and statistics literatures [4–7] (also see special issues of IEEE Information Theory [8, 9]). In particular, there have been many multiscale regularization methods proposed for inverse problems [10–15]. Most of the techniques developed to date are based on Gaussian noise models which are not directly applicable in Poisson inverse problems. More specifically, models developed for Gaussian problems do not capture the non-negativity of intensity functions and are not well-matched in functional form to the Poisson likelihood, rendering analysis, interpretation and implementation more difficult. In the low SNR cases of greatest practical interest, the data are not well modeled with standard Gaussian approximations to the Poisson likelihood, and hence many existing multiscale regularization methods are simply inappropriate.

Recently, some attempts have been made to wed the multiscale analysis paradigm with Poisson estimation problems. The Bayesian multiscale framework independently introduced in [6, 16] for problems involving direct Poisson observations is a key example. The approaches in [6, 16] begin with a multiscale factorization of the Poisson likelihood function, which in turn induces a re-parameterization of the underlying intensity in terms of canonical multiscale parameters. With the use of conjugate priors in the multiscale parameter space, this framework then admits a remarkably simple Bayesian multiscale analysis tool for Poisson data that is analogous to wavelet-based counterparts used in Gaussian denoising/estimation problems [17,18]. This work demonstrates that while conventional wavelet-based multiscale analysis is not necessarily as compatible with Poisson data, an equally appealing and closely related alternative is naturally suited to this case.

This paper introduces a novel Bayesian multiscale framework for Poisson inverse problems, and as such extends the authors' earlier work (described just above) in the case of directly observed Poisson data. This framework shares the characteristics and advantages just described, but in addition has several other desirable features that are germane to the specific context of Poisson inverse problems. First, this multiscale framework admits a simple EM algorithm for computing MAP reconstructions. EM has a strong information-theoretic motivation [19] and this work connects probabilistic photon-limited imaging models with multiscale analysis in a formal and precise fashion. The EM algorithm involves closed-form (analytic) steps at each iteration, making it computationally attractive. Second, under mild regularity assumptions on the multiscale prior density, which are readily verified by a simple algebraic check on hyperparameter settings, it can be proven that the

EM algorithm converges to a unique, global MAP estimate. Third, the effects of the multiscale prior density (and hyperparameter settings) are easily interpreted, which is important from a user's perspective in applications.

The paper is organized as follows. In Section II, we give the basic problem formulation and discuss existing estimation methods. In Section III, we review the Bayesian multiscale approach to modeling and analyzing Poisson data. In Section IV, we apply this framework to the Poisson inverse problem and derive a multiscale EM algorithm to compute the MAP estimate. In Section V, we study the convergence of the EM algorithm. In Section VI, we discuss the issue of selecting hyperparameters for the prior. In Section VII, we look at two applications of the new framework. We close in Section VIII with remarks and conclusions.

II. Problem Formulation

The following problem is addressed in this paper. Suppose that we observe Poisson distributed data (counts)

$$y_n \sim \mathcal{P}(\mu_n), \quad n = 0, \ldots, N-1, \tag{1}$$

where $\mathcal{P}(\mu_n)$ denotes a Poisson distribution with intensity parameter μ_n. The (unknown) intensities $\boldsymbol{\mu} = \{\mu_n\}_{n=0}^{N-1}$ are related to other (unknown) intensities, $\boldsymbol{\lambda} = \{\lambda_m\}_{m=0}^{M-1}$, of primary interest, via the relation $\boldsymbol{\mu} = \mathbf{P}\boldsymbol{\lambda}$, where $\mathbf{P} = \{p_{n,m}\}$ is an $N \times M$ matrix of known non-negative weights (usually transition probabilities). In the case where there exists known background information, in the form of a $N \times 1$ vector $\mathbf{b}$, the alternative model $\boldsymbol{\mu} = \mathbf{P}\boldsymbol{\lambda} + \mathbf{b}$ is used typically, but we shall assume here that $\mathbf{b} = 0$ for simplicity (our proposed method extends immediately to the case of $\mathbf{b} \neq 0$). We will also assume that the rows of $\mathbf{P}$ sum to unity, although this is not a necessary restriction. The problem is to estimate $\boldsymbol{\lambda}$ from the observed data $\mathbf{y} = \{y_n\}_{n=0}^{N-1}$. Throughout this paper, we will assume $M = 2^J$, for some integer $J > 0$, while N can be an arbitrary integer. Since M typically is chosen by the user, while N normally is predetermined by instrumental design constraints, this common condition should present no difficulties.

A classical application in which the above estimation problem arises is that of photon-limited imaging. Photons are emitted (from the emission space) according to an intensity $\boldsymbol{\lambda}$. Those photons emitted from location m are detected (in the detection space) at position n with transition probability $p_{n,m}$. From a conceptual standpoint, foreshadowing our later usage of an EM framework, it is useful to introduce a representation for these "unobservable" data [20], and denote the total number of $m \to n$ emission/detection events as $z_{n,m}$, in which case

$$z_{n,m} \sim \mathcal{P}(\lambda_m p_{n,m}) \ . \tag{2}$$

Hence the indirectly observed (and therefore "incomplete") data $\mathbf{y}$ in (1) are given by $y_n = \sum_m z_{n,m}$. Additionally, were we able to observe them, the direct emission data for each location m would be given by sums of the form $x_m \equiv \sum_n z_{n,m}$, from which it follows that $x_m \sim \mathcal{P}(\lambda_m)$. Therefore, if $\mathbf{z}$ were known, we could avoid the inverse problem altogether and simply deal with the issue of estimating a Poisson intensity given direct observations. Of course, this device is precisely what the well-known EM algorithm exploits in producing estimates of $\boldsymbol{\lambda}$ from the indirect data $\mathbf{y}$, a fact that will be fundamental to our own approach introduced herein.

A. Maximum Likelihood Estimation

The log-likelihood is

$$\log\, p(\mathbf{y}|\boldsymbol{\lambda}) \;=\; \sum_{n=0}^{N-1}\left(-\sum_{m=0}^{M-1} p_{n,m}\lambda_m + y_n \log\left(\sum_{m=0}^{M-1} p_{n,m}\lambda_m\right) - \log y_n!\right). \tag{3}$$

It is well-known that the maximizer of (3) cannot be expressed in closed-form and must be determined numerically. While in principle any numerical optimization method could be used, the iterative EM algorithm, as first proposed for this problem in [2], has a number of features that make it especially desirable, most notably its natural, probabilistic formulation, computationally straightforward calculations at each iteration step, and numerical stability [20]. Moreover, it can be shown that the EM algorithm monotonically increases the log-likelihood at each iteration and converges to a global (not necessarily unique) point of maximum for (3) [21].

Unfortunately, due to the ill-posed nature of the likelihood equations, the variance of the MLE can be quite high, particularly for applications involving very low counts. In fact, in many cases the MLE is practically useless. A popular remedy is to stop the EM algorithm prior to convergence (e.g., [22]). Stopping the algorithm acts implicitly as a smoothing operation and can produce acceptable results. However, it may be preferable to abandon the strictly likelihood-based perspective altogether, and approach the inverse problem with a different criterion, one that smooths through a well-defined optimal solution, while still providing useful and meaningful results.

B. Bayesian MAP Estimation and Penalized MLE

Several Bayesian (and/or Penalized Maximum Likelihood) procedures have been developed that use prior information (or regularizing/penalizing functionals) to produce MAP estimates that are more desirable than the MLE in many cases [1,23–27]. Here, the log-posterior, which is proportional to the log-likelihood plus the log-prior density, replaces the likelihood as the optimization criterion. The log-posterior is given by

$$\log\, p(\boldsymbol{\lambda}|\mathbf{y}) \;\propto\; \log\, p(\mathbf{y}|\boldsymbol{\lambda}) + \log\, p(\boldsymbol{\lambda}), \tag{4}$$

where $\log p(\mathbf{y}|\boldsymbol{\lambda})$ is given in (3) and the prior $p(\boldsymbol{\lambda})$ can be, for example, a *Markov Random Field* (MRF) [24]. The $\log\, p(\boldsymbol{\lambda})$ can also be interpreted as a penalizing functional, hence the terminology "penalized MLE." A MAP estimate is a value of $\boldsymbol{\lambda}$ that maximizes (4). As in the case of the MLE, the MAP estimate must be computed numerically, in general. The EM algorithm is again a popular choice for optimization, although, in general, the M steps do not have closed form expressions, as they do for the MLE case.

This more challenging nature of the optimization problem is a practical issue that has limited the widespread application of MAP methods in Poisson inverse problems, but there are other vexing issues as well. For example, the selection of useful hyperparameter (regularization parameter) settings has been the focus of considerable work (*e.g.,* [28]), and the difficulty of interpreting the effects of various user-selected parameters makes the application of these techniques somewhat of an art. In cases where the prior and its effects are more readily understood, as is the case for simple quadratic roughness penalties or Gaussian priors, one usually faces a direct trade-off between smoothness (resolution) and edge preservation. Furthermore, in many cases the priors do not reflect the strict non-negativity of Poisson intensity functions (*e.g.,* Gaussian prior or quadratic penalty), and the potential non-negativity of the resulting MAP estimator is either ignored or enforced using additional constraints. Since intensities are inherently non-negative, in a Bayesian context it is much more natural to employ a prior that supports this knowledge.

C. Multiscale Regularization Methods

Multiscale methods offer an alternative to conventional (spatial) regularization techniques in a host of inverse problems, most notably in tomographic image reconstruction [15,29,30], inverse problems involving (scale) homogeneous operators [11,31] and image deblurring [10,13,14]. Wavelet representations are often utilized in multiscale schemes. Wavelet-based methods are advantageous since they enable (nonlinear) estimation procedures that adapt to the local characteristics of the underlying object. Roughly speaking, one can recover edges and singularities that are well supported by the data, while simultaneously smoothing in other regions. Hence, such methods are a viable alternative to non-quadratic Markov random field priors.

Most attempts at multiscale regularization have been made in conjunction with a Gaussian observation model. Even in the contexts where Poisson observation models are especially appropriate, the tendency has been to use Gaussian approximations instead. This may be due to difficulty of formulating an appropriate multiscale analysis in the Poisson case. Wavelet methods have been developed by appropriately adapting schemes devised for Gaussian data as in [32] or by using cross-validation techniques [33]. However, as mentioned earlier in the introduction, our previous work has demonstrated that conventional wavelet analysis and Poisson data are somewhat incompatible, in terms of both theoretical tractability and algorithm implementation. However, the remarkable performance of wavelets and multiscale regularization in the Gaussian arena motivates the search for a similar treatment of Poisson data. Notably, our preliminary work which appeared in [34] and the multiresolution MRF for Bayesian tomography developed in [35] represent two very different steps in this direction. In this paper, we present a unified Bayesian multiscale framework in which the Poisson likelihood is complemented with a suitable and natural multiscale reparameterization and prior, based on the initial work in [34].

III. Multiscale Analysis of Poisson Problems — Direct Data

Let us suppose for the moment that the emission data

$$x_m \sim \mathcal{P}(\lambda_m), \quad m = 0, \ldots, M-1, \tag{5}$$

are available to us. Were we to observe this data directly, then we could employ the multiscale analysis and estimation techniques for Poisson data developed in [6] and [16]. Here, we briefly review the fundamental aspects of those techniques. We will use the same intensity parameterization and prior to tackle the more general inverse problem in Section IV. Moreover, in conjunction with the notion of the unobservable data $\mathbf{z}$, these results enable a very simple and natural EM algorithm for likelihood/posterior maximization.

Multiscale analysis refers to the study of behavior or structure in signals or data at various spatial and/or temporal resolutions [36]. Here we effect a multiscale analysis through simple recursive summation of the data, which is equivalent to processing with a Haar scale function. To illustrate, let us consider the one-dimensional case of this analysis, defined according to:

$$\begin{aligned} x_{J,m} &\equiv x_m, \quad m = 0, \ldots, 2^J - 1 \\ x_{j,m} &= x_{j+1,2m} + x_{j+1,2m+1}, \quad m = 0, \ldots, 2^j - 1, \ 0 \leq j \leq J-1. \end{aligned} \tag{6}$$

The index j refers to the resolution of the analysis, 2^j; $j = J$ being the index for the highest resolution (finest scale), and $j = 0$ corresponding to the lowest resolution (coarsest scale). The multiscale data $\{x_{j,m}\}$ are the (unnormalized) Haar scaling coefficients of $\mathbf{x}$, which can be organized and represented on a binary tree graph, as shown in Figure 1. This Haar multiscale analysis is especially well-suited to Poisson data, as the Poisson distribution reproduces under summation (i.e., the unweighted sum of independent Poisson variates is itself Poisson distributed). Analyses with more general wavelets result in arbitrary linear combinations of Poisson random variables, for which such nice distributional characteristics do not result. Hence, issues of mathematical tractability and interpretability quickly arise with more general wavelet analyses of Poisson data, contrary to the case when such wavelets are used instead with Gaussian data [17, 32].

Now, because the data $\mathbf{x}$ are independent (given the intensity $\boldsymbol{\lambda}$) and using standard conditional probability relationships, we can express the joint probability of the data in terms of the multiscale representation with the following factorized form:

$$\Pr(\mathbf{x}) \;=\; \Pr(x_{0,0}) \prod_{j=0}^{J-1} \prod_{m=0}^{2^j-1} \Pr(x_{j+1,2m}|x_{j,m}). \tag{7}$$

The expression in (7) actually holds more generally (i.e., not just for Poisson data), and may be viewed as a likelihood factorization with respect to a particular graphical model [37] – in this case, as a simple binary tree. See [17] for additional investigations along these lines.

This factorization captures the basic relationship between a "parent" (at a coarse scale, *e.g.*, $x_{j,m}$) and a "child" (at the next finer scale, *e.g.*, $x_{j+1,2m}$). To see this point in more detail, consider the specific distributional form of the conditional likelihood of the child given the parent, $p(x_{j+1,2m}\,|\,x_{j,m},\boldsymbol{\lambda})$. Specifically, first define a multiscale analysis of the intensity $\boldsymbol{\lambda}$, analogous to that defined for the data $\mathbf{x}$:

$$\begin{aligned}\lambda_{J,m} &\equiv \lambda_m, \quad m=0,\ldots,2^J-1\\ \lambda_{j,m} &= \lambda_{j+1,2m}+\lambda_{j+1,2m+1}, \quad m=0,\ldots,2^j-1,\ 0\le j\le J-1.\end{aligned} \tag{8}$$

The parameters $\{\lambda_{j,m}\}$ are the (unnormalized) Haar scaling coefficients of $\boldsymbol{\lambda}$ and can be represented on a binary tree graph, as also shown in Figure 1. With this definition in hand, we have the following expression for the parent-child conditional likelihood.

$$p(x_{j+1,2m}\,|\,x_{j,m},\boldsymbol{\lambda}) \;=\; \mathcal{B}\left(x_{j+1,2m}\;\middle|\;x_{j,m},\frac{\lambda_{j+1,2m}}{\lambda_{j,m}}\right), \tag{9}$$

where $\mathcal{B}(x\,|\,n,\rho)=\binom{n}{x}\rho^x(1-\rho)^{n-x}$, denotes the binomial distribution with parameters n and ρ. ¿From this expression, we identify the following *canonical* multiscale parameters associated with the Poisson observation model,

$$\rho_{j,m}=\frac{\lambda_{j+1,2m}}{\lambda_{j,m}}, \quad m=0,\ldots,2^j-1,\ 0\le j\le J-1, \tag{10}$$

which can be viewed as "splitting" factors that govern the multiscale refinement of the intensity $\boldsymbol{\lambda}$ [6, 16]. The splitting factors $\{\rho_{j,m}\}$ can be interpreted as multiplicative weights which are represented as edges (or links) of the binary tree graph depicted in Figure 1.

Hence, we can see that the factorization in (7) may be expressed as

$$p(\mathbf{x}\,|\,\boldsymbol{\lambda}) \;=\; \mathcal{P}(x_{0,0}|\lambda_{0,0}) \prod_{j=0}^{J-1} \prod_{m=0}^{2^j-1} \mathcal{B}\left(x_{j+1,2m}\;\middle|\;x_{j,m},\rho_{j,m}\right), \tag{11}$$

where $\mathcal{P}(x_{0,0}|\lambda_{0,0})$ denotes a Poisson probability mass function of $x_{0,0}$, with intensity $\lambda_{0,0}$.

A. Maximum Likelihood Estimation and Intensity Reconstruction

Inspection of the binomial conditional likelihood factors shows that the MLE of each split $\rho_{j,m}$ is given by

$$\widehat{\rho}_{j,m} \;=\; \frac{x_{j+1,2m}}{x_{j,m}}, \tag{12}$$

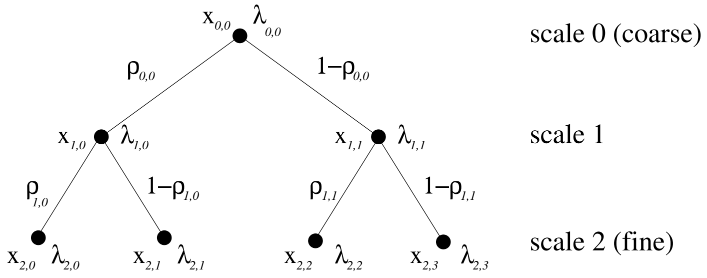

Fig. 1. Multiscale analysis and modeling represented on binary tree graph. A data coefficient $x_{j,k}$ and intensity coefficient $\lambda_{j,k}$ are associated with node j, k in the binary tree. Edges (or links) between nodes represent the multiplicative weights (splitting factors) $\{\rho_{j,k}\}$ governing the refinement of the intensity function.

which is simply the empirical splitting factor relating the data at one scale to another. Note that the condition $x_{j,m} = 0$ implies that $x_{j+1,2m} = 0$ as well. Thus, we adopt the convention that $\widehat{\rho}_{j,m} \equiv 0$ if $x_{j,m} = 0$. The MLE of the total intensity $\lambda_{0,0}$ is simply the total count $x_{0,0}$.

Because the mapping from $(\boldsymbol{\rho}, \lambda_{0,0})$ to $\boldsymbol{\lambda}$ is one-to-one, the MLE of intensity is computed by a simple $O(M)$ reconstruction algorithm based on the corresponding multiscale parameter estimates and using the multiscale synthesis equations

$$\begin{aligned} \widehat{\lambda}_{j+1,2m} &= \widehat{\lambda}_{j,m}\widehat{\rho}_{j,m}, \\ \widehat{\lambda}_{j+1,2m+1} &= \widehat{\lambda}_{j,m}(1-\widehat{\rho}_{j,m}), \quad m = 0, \ldots, 2^j - 1, \ 0 \leq j \leq J-1. \end{aligned} \tag{13}$$

Its easily seen that the MLE of each intensity element at the finest scale (highest resolution dictated by the data) is given by

$$\widehat{\lambda}_{J,m} = x_{J,m} \equiv x_m. \tag{14}$$

That is, the MLE simply returns the raw data $\mathbf{x}$ as our MLE intensity estimate, as expected. We have already discussed the shortcomings of the MLE, so we next consider a MAP estimation procedure.

B. Maximum A Posteriori Estimation

The crucial ingredient in any Bayesian procedure is the selection of a suitable prior. Ideally, the prior reflects known or assumed attributes of the intensity in question and, for computational purposes, it is also convenient if the prior is well matched, in functional form, to the Poisson (or Poisson-binomial) likelihood. Parametric *conjugate priors* are advantageous for computational reasons since the posterior distribution is obtained by simply "updating" the parameters of the prior based on the observations; see [38, pp. 97-111]. Moreover, we will see that conjugate priors can provide in the current setting very plausible models for the multiscale parameters. The family of gamma densities is a conjugate family to the Poisson likelihood and the beta family is conjugate to the binomial, and we adopt these priors here.

Begin by placing a gamma density prior on the total intensity parameter:

$$\lambda_{0,0} \ \sim \ \frac{\delta^\gamma}{\Gamma(\gamma)} \lambda_{0,0}^{\gamma-1} \exp\{-\delta\lambda_{0,0}\} \ \equiv \ \mathcal{G}(\lambda_{0,0}|\gamma, \delta),$$

with $\gamma > 0$ and $\delta > 0$. Next, we model each multiscale split parameter as an independent beta distributed random variable,

$$\rho \sim \frac{\rho^{\alpha-1}(1-\rho)^{\beta-1}}{B(\alpha, \beta)} \equiv \mathcal{B}e(\rho \,|\, \alpha, \beta),$$

$0 \le \rho \le 1$, where $B(\alpha, \beta)$ denotes the standard beta function. In this paper, we will only use symmetric beta priors of mean $1/2$, characterized by $\alpha = \beta$. Here, as in most related approaches, we do not have the parameters depend on the location m, since location dependent signal characteristics are usually not known *a priori*.

The prior density for the unknown parameters $\lambda_{0,0}$ and $\boldsymbol{\rho}$ is therefore

$$p(\lambda_{0,0}, \boldsymbol{\rho}) \;=\; \mathcal{G}(\lambda_{0,0}|\gamma, \delta) \prod_{j=0}^{J-1} \prod_{k=0}^{2^j-1} \mathcal{B}e(\rho_{j,k} \,|\, \alpha_j, \alpha_j). \tag{15}$$

The gamma prior on $\lambda_{0,0}$ can be tailored to reflect knowledge of the total intensity of the process under consideration. However, since the total count $x_{0,0}$ is typically quite large, with reasonable settings for the hyperparameters γ and δ the effect of the gamma prior is negligible. More important are the beta priors placed on the splits. Here, the hyperparameters $\{\alpha_j\}$ reflect our belief or prior knowledge regarding the regularity of the intensity (more on this in Section VI). To illustrate briefly, setting $\alpha_j = 1$, $j = 0, \ldots, J-1$, we have uniform (constant) prior densities on the splits, expressing absolute ignorance about the multiscale refinement of the intensity. With $\alpha_j \gg 1$, $j = 0, \ldots, J-1$, the beta prior densities are peaked about the point $1/2$, favoring a more even and regular refinement. Similar priors have been used for nonparametric probability density modeling and estimation under the name of *Polya trees* [39, 40].

Combining the prior (15) with the likelihood (11) and making use of the conjugacy of the beta and gamma priors produces a posterior density

$$p(\lambda_{0,0}, \boldsymbol{\rho} \,|\, \mathbf{x}) \;=\; p(\lambda_{0,0}|x_{0,0}) \prod_{j=0}^{J-1} \prod_{m=0}^{2^j-1} p(\rho_{j,m} \,|\, x_{j,m}, x_{j+1,2m}), \tag{16}$$

where

$$p(\lambda_{0,0}|x_{0,0}) = \mathcal{G}(\lambda_{0,0}|\gamma + x_{0,0}, \delta + 1)$$

$$p(\rho_{j,m} \,|\, x_{j,m}, x_{j+1,2m}) \;=\; \mathcal{B}e\left(\rho_{j,m} \,|\, \alpha_j + x_{j+1,2m}, \alpha_j + x_{j,m} - x_{j+1,2m}\right).$$

The factorization of the posterior shows that inferences can be made on each multiscale parameter individually, instead of requiring a complicated high dimensional analysis. MAP estimates of the ρ's and $\lambda_{0,0}$, $\widehat{\boldsymbol{\rho}} = \{\widehat{\rho}_{j,m}\}$ and $\widehat{\lambda}_{0,0}$, have simple closed-form expressions:

$$\widehat{\lambda}_{0,0} \;=\; \frac{\gamma + x_{0,0} - 1}{\delta + 1} \tag{17}$$

and

$$\widehat{\rho}_{j,m} \;=\; \frac{x_{j+1,2m} + \alpha_j - 1}{x_{j,m} + 2(\alpha_j - 1)}, \quad 0 \le m \le 2^j - 1,\; 0 \le j \le J-1. \tag{18}$$

Again, because the mapping from $(\boldsymbol{\rho}, \lambda_{0,0})$ to $\boldsymbol{\lambda}$ is one-to-one, the MAP estimate of intensity is computed using the synthesis equations (13) with the MAP estimates above (17-18) in place of the MLEs.

IV. A Bayesian Multiscale Approach to Poisson Inverse Problems

We now return attention to our original (and more formidable) Poisson inverse problem, as described in (1), in which the emission process is not directly observed. Our objective is to apply a Bayesian multiscale analysis, similar to that just described above, in this case. Specifically, we seek a MAP estimate that maximizes (4), using the intensity prior induced on $\boldsymbol{\lambda}$ by the multiscale prior (15) on $(\lambda_{0,0}, \boldsymbol{\rho})$. This is a well-defined MAP estimation problem, however the difficulty we face is the same as that faced in finding an MLE; maximizing the objective function is not straightforward and must be performed numerically.

As previously mentioned, the EM algorithm is a popular maximization tool, especially in the context of Poisson inverse problems [2, 20]. Maximization is facilitated within the EM framework through the introduction of a particular "unobservable" data space. For example, in the Poisson inverse problem, if the unobservable data $\mathbf{z}$ (as defined in (2)) were actually observed, then a closed-form maximizer of the accompanying complete (observed+unobserved) data likelihood function $p(\mathbf{z}|\boldsymbol{\lambda})$ would be available. [Note: Since the observed data are also determined from $\mathbf{z}$, according to $y_n = \sum_m z_{n,m}$, $n = 1, \ldots, N$, we will refer to $\mathbf{z}$ hereafter as the complete-data.] The EM algorithm is an iterative method that alternates between computing the conditional (i.e., given the observed data) expectation of the complete-data log-posterior

$$\ell_c(\boldsymbol{\lambda}) \;\equiv\; \log p(\mathbf{z}|\boldsymbol{\lambda}) \;+\; \log p(\boldsymbol{\lambda}), \tag{19}$$

and the maximizer of the resulting function, and it leads to a MAP estimate of the (observed data only) log-posterior. In this section we develop the details of the EM algorithm associated with our particular Bayesian multiscale framework.

A. Multiscale Framework

Recall from Section III, for the analysis of directly observed Poisson data, that a multiscale factorization of the data likelihood played a central role. In the context of indirectly observed data, through the EM algorithm one is led to consider the complete-data likelihood, as was just described. As we show below, the complete-data likelihood yields a multiscale factorization similar to that observed for the direct-data likelihood. In fact, somewhat remarkably, the former is actually proportional to the latter. Therefore, the same multiscale prior density given in (15) for the direct data case is just as appropriate in the case of indirectly observed data. Consequently, the corresponding complete-data log-posterior $\ell_c(\boldsymbol{\lambda})$ is easily maximized with respect to the multiscale parameters $\{\lambda_{0,0}, \boldsymbol{\rho}\}$.

In more detail, these results are derived as follows. Begin by recalling that the direct (unobserved) emission data are given by $x_m = \sum_n z_{n,m}$, $m = 0, \ldots, M-1$ (where $M = 2^J$). Define the multiscale analysis of these data according to (6), as we did in the direct observation case. For the unknown intensity parameter $\boldsymbol{\lambda}$, we again use the multiscale analysis given in (8). Calculations similar to those yielding equations (7) and (11) show that the complete-data likelihood can be factorized as follows.

$$\begin{aligned} p(\mathbf{z}\,|\,\boldsymbol{\lambda}) \;&=\; \mathcal{P}(x_{0,0}|\lambda_{0,0}) \;\times\; \prod_{j=0}^{J-1} \prod_{m=0}^{2^j-1} \mathcal{B}(x_{j+1,2m}\,|\,x_{j,m}, \rho_{j,m}) \;\times \\ & \prod_{m=0}^{M-1} \mathcal{M}(z_{0,m}, \ldots, z_{N-1,m}|p_{0,m}, \ldots, p_{N-1,m}, x_{J,m}). \end{aligned} \tag{20}$$

The first factor $\mathcal{P}$ is a Poisson mass function with intensity $\lambda_{0,0}$. The factors of the form $\mathcal{B}$ are binomial conditional likelihoods. Finally, the factors $\mathcal{M}$ are multinomial with parameters $p_{0,m}, \ldots, p_{N-1,m}$, the m-th column of the matrix $\mathbf{P}$ of transition probabilities, and $x_{J,m}$, the total counts at emission location m (we also make use of the assumption that the rows of $\mathbf{P}$ sum to unity here, although this does not play a critical role).

The first step in deriving the factorization in (20) is, for $m = 0, 1, \ldots, M-1$, to write

$$\Pr(\mathbf{z}_{\cdot,m}) = \Pr(x_{J,m}) \times \Pr(\mathbf{z}_{\cdot,m}|x_{J,m}) \ ,$$

where $x_{J,m}$ is just the summation of the elements in $\mathbf{z}_{\cdot,m} \equiv [z_{0,m}, \ldots, z_{N-1,m}]$. Using this step and the fact that $p(\mathbf{z}|\boldsymbol{\lambda}) = \prod_{m=0}^{M-1} p(\mathbf{z}_{\cdot,m}|\boldsymbol{\lambda})$, the product of factors $\mathcal{M}$ results. Left over from this step is a term of the form $\prod_m \Pr(x_{J,m})$. But this term is simply the direct-data likelihood, which therefore may be factorized as in equation (11). The impact of this deceptively simple set of steps becomes apparent when it is noted that the multinomial probability mass functions $\mathcal{M}$ depend only on the complete-data and the transition probabilities. In particular, they *do not* depend on the unknown intensity parameter $\boldsymbol{\lambda}$. Hence, for the purposes of studying $\boldsymbol{\lambda}$ we can ignore these factors and simply write

$$p(\mathbf{z}\,|\,\boldsymbol{\lambda}) \ \propto \ \mathcal{P}(x_{0,0}|\lambda_{0,0}) \ \times \ \prod_{j=0}^{J-1} \prod_{m=0}^{2^j-1} \mathcal{B}(x_{j+1,2m}\,|\,x_{j,m}, \rho_{j,m}). \tag{21}$$

In other words, the *complete data likelihood* is proportional to the *direct data likelihood.* The fact that such a result is special can be confirmed easily, for example by noting that a similar result does *not* follow in the case of a Gaussian inverse problem, with $y_n \sim \text{Normal}(\mu_n, \sigma^2)$, in analogy to (1). In that case, the term analogous to the second line in (20) is a product of conditional multivariate Gaussian factors, involving $\mathbf{z}$, $\mathbf{P}$, *and* $\boldsymbol{\lambda}$.

To finish our discussion of the Poisson case, we note that in combination with the logarithm of the prior in (15), the result in (21) allows us to express the log complete-data posterior distribution as

$$\begin{aligned}
\ell_c(\boldsymbol{\lambda}) \ &\equiv \ \log p(\lambda_{0,0}, \boldsymbol{\rho}|\mathbf{z}) \\
&= \ \log \mathcal{P}(x_{0,0}|\lambda_{0,0}) + \log \mathcal{G}(\lambda_{0,0}|\gamma,\delta) \ + \\
&\quad \sum_{j=0}^{J-1} \sum_{m=0}^{2^j-1} \log \mathcal{B}(x_{j+1,2m}\,|\,x_{j,m}, \rho_{j,m}) + \log \mathcal{B}e(\rho_{j,m}\,|\,\alpha_j, \alpha_j) + C,
\end{aligned} \tag{22}$$

where C is a constant that does not depend on the parameters $(\lambda_{0,0}, \boldsymbol{\rho})$. So, we have two equivalent expressions for the complete-data log-posterior; (19) in the spatial domain and (22) in the multiscale parameterization. Due to its simple form, maximizing (22) with respect to the splits and total intensity is trivial; simply differentiate the expression to obtain the MAP estimates, also given by expressions (17) and (18). We take advantage of this in our formulation of an EM algorithm next.

B. EM Algorithm for Multiscale MAP Estimation

In general, the main difficulty encountered in the MAP-EM algorithm is that the M-Step typically does not admit a closed-form solution, as it does in the case of the MLE. Hence, much of the simplicity of the EM algorithm is often lost in MAP estimation problems; one exception is obtained by taking a quadratic (Gaussian) prior, but this can lead to oversmoothing and the parameter space must be restricted to insure non-negative solutions (quadratic priors are defined over all real-valued images, not just non-negative intensities) [41]. One of the strengths of our multiscale approach, in addition to insuring a non-negative estimate, is that we do have closed-form M-Steps.

The steps of our EM algorithm take the following forms. The initial iterate $\boldsymbol{\lambda}^{(0)}$ can be chosen as a constant intensity or another suitable starting point (such as the reconstruction obtained by a simple, filtered back-projection reconstruction). At the k+1-st iteration the E-Step and M-Step are:

E-Step: Compute the expectation of the log posterior, conditioned on y, and under the Poisson law induced by $\boldsymbol{\lambda}^{(k)}$:

$$Q(\boldsymbol{\lambda}, \boldsymbol{\lambda}^{(k)}) \ \equiv\ E_{\boldsymbol{\lambda}^{(k)}} [\log p(\mathbf{z}|\boldsymbol{\lambda})|\mathbf{y}] \ +\ \log p(\boldsymbol{\lambda}) \ . \tag{23}$$

Note that since $\log p(\mathbf{z}|\boldsymbol{\lambda})$ is linear in $\mathbf{z}$ (up to a constant term depending only on data), this step reduces to computing $\mathbf{z}^{(k)} \equiv E_{\boldsymbol{\lambda}^{(k)}} [\mathbf{z}|\mathbf{y}]$ and, since $\mathbf{z}|\mathbf{y}$ is multinomial, we have

$$z^{(k)}(n,m) \ =\ \frac{y_n \lambda_m^{(k)} p_{n,m}}{\sum_{l=0}^{2^J-1} \lambda_l^{(k)} p_{n,l}}. \tag{24}$$

M-Step: Maximize the expected complete-data log-posterior, (19), after transforming into the multiscale representation (22). This reduces to a two-step process.

(i.) Generate $\mathbf{x}^{(k)}$ from $\mathbf{z}^{(k)}$.

(ii.) Calculate $(\lambda_{0,0}^{(k+1)}, \boldsymbol{\rho}^{(k+1)})$ according to

$$\widehat{\lambda}_{0,0}^{(k+1)} \ =\ \frac{\gamma + x_{0,0}^{(k)} - 1}{\delta + 1}$$

and

$$\widehat{\rho}_{j,m}^{(k+1)} \ =\ \frac{x_{j+1,2m}^{(k)} + \alpha_j - 1}{x_{j,m}^{(k)} + 2(\alpha_j - 1)}, \quad 0 \leq m \leq 2^j - 1,\ 0 \leq j \leq J-1.$$

Then reconstruct $\boldsymbol{\lambda}^{(k+1)}$ via

$$\begin{aligned} \widehat{\lambda}_{j+1,2m}^{(k+1)} &= \widehat{\lambda}_{j,m}^{(k+1)} \widehat{\rho}_{j,m}^{(k+1)}, \\ \widehat{\lambda}_{j+1,2m+1}^{(k+1)} &= \widehat{\lambda}_{j,m}^{(k+1)} (1 - \widehat{\rho}_{j,m}^{(k+1)}), \quad m = 0, \ldots, 2^j - 1,\ 0 \leq j \leq J-1. \end{aligned}$$

This algorithm has several desirable properties. First, as a standard property of the EM algorithm, the posterior probability is non-decreasing as we iterate. Second, it is easily verified that, by construction, the resulting estimate is non-negative. Third, if we set $\alpha_j = 1,\ j = 0, \ldots, J-1$, in which case the beta densities for the $\{\rho_{j,m}\}$ coincide with the uniform density on $[0,1]$ (a non-informative case of our split prior), and set $\gamma = 1,\ \delta = 0$ in the gamma prior on $\lambda_{0,0}$, a limiting (improper) form of the gamma density, then we recover the classical MLE method [2].

On a final note, we mention that a perhaps surprising feature of our model is the computational simplicity of the accompanying implementation of the EM algorithm. In fact, this implementation is no more demanding than that proposed originally for the simple likelihood-based model [2]. Most other MAP criteria proposed for this problem do not admit such a simple implementation; usually the maximization step does not have a closed-form expression and must be computed numerically or approximately.

V. Convergence of Multiscale EM Algorithm

In this section, we establish the convergence properties of the multiscale EM algorithm developed above, in the form of two key results. First, the multiscale prior in (15) induces a prior on $\boldsymbol{\lambda}$ which, under certain conditions on the hyperparameters, is strictly concave, which implies a unique maximizer exists. Second, under these same conditions, our EM algorithm converges to the unique maximum point of the log-posterior of $\boldsymbol{\lambda}$.

We begin with the following result.

Lemma 1 *For $\boldsymbol{\lambda}$ with non-negative components in $\mathbb{R}^M$, the multiscale prior defined on $(\lambda_{0,0}, \boldsymbol{\rho})$ induces a prior distribution with the following density.*

$$\begin{aligned} h(\boldsymbol{\lambda}) &= \frac{\delta^\gamma}{\Gamma(\gamma)} \lambda_{0,0}^{\gamma - M_J} \exp\{-\delta \lambda_{0,0}\} \\ & \prod_{j=0}^{J-1} \prod_{m=0}^{M_j - 1} \frac{1}{B(\alpha_j, \alpha_j)} \left(\frac{\lambda_{j+1,2m} \lambda_{j+1,2m+1}}{\lambda_{j,m}^2} \right)^{\alpha_j - M_{J-j-1}} , \end{aligned} \tag{25}$$

where $M_j \equiv 2^j$.

The proof of Lemma 1 follows by induction on $J = \log_2(M)$, and may be found in the appendix (as may the proofs of all other results stated in this section). Considering the manner in which $h(\boldsymbol{\lambda})$ was defined (i.e., with respect to the multiscale prior in 15), it is not at all clear that $h(\boldsymbol{\lambda})$ should necessarily change in a well-behaved manner in $\boldsymbol{\lambda}$. In fact, it is not difficult to show (e.g., consider the simple case of $J = 2$) that the multiscale log-prior is not concave in $(\lambda_{0,0}, \boldsymbol{\rho})$. However, as the following lemma describes, there exists an interesting condition under which the induced log-prior $\log h(\cdot)$ is strictly concave in $\boldsymbol{\lambda}$.

Lemma 2 *The log-prior density function, $\log h(\boldsymbol{\lambda})$, is strictly concave in $\boldsymbol{\lambda}$ if and only if the hyperparameters $\gamma, \alpha_0, \ldots, \alpha_{J-1}$ satisfy*

$$\alpha_{J-1} - 1 + \sum_{j=1}^{J-1} (\alpha_{j-1} - 2\alpha_j) r_j + (\gamma - 2\alpha_0) r_0 \ > \ 0 \tag{26}$$

for every set of positive numbers $0 \le r_o \le r_1 \le \cdots \le r_{J-1} \le 1$.

Two points are worth noting in light of Lemma 2. First, the hyperparameter δ in the prior distribution on $\lambda_{0,0}$ plays no role. Second, the conditions $\gamma \ge 2\alpha_0$, $\alpha_0 \ge 2\alpha_1$, …, $\alpha_{J-2} \ge 2\alpha_{J-1}$, and $\alpha_{J-1} \ge 1$, with strict inequality holding for at least one pair, are sufficient. In other words, concavity can be achieve essentially through a doubling of the hyperparameters α_j, as j decreases (moving from fine to coarse scale).

We defer discussion of the practical impact of these hyperparameter constraints until Section VI. Continuing here with our convergence analysis, we write the log-posterior density function as

$$\ell(\boldsymbol{\lambda}) \ \equiv \ \log p(\mathbf{y}|\boldsymbol{\lambda}) \ + \ \log h(\boldsymbol{\lambda}), \tag{27}$$

which is simply (4) re-expressed to emphasize that the prior density is that defined in Lemma 1. It is known from [21] that the (incomplete) data log-likelihood is concave in $\boldsymbol{\lambda}$, though not strictly concave (except under unlikely conditions on $\mathbf{P}$). Hence the log-posterior density is the sum of a concave function and a strictly concave function, under the condition of Lemma 2, and therefore strictly concave itself. As a result, we have the following.

Theorem 1 *Under the conditions of Lemma 2, the iterates $\{\boldsymbol{\lambda}^{(k)}\}$ of the EM algorithm defined in Section IV converge to a limit point $\boldsymbol{\lambda}^{(\infty)}$, and $\boldsymbol{\lambda}^{(\infty)}$ is the maximum point of $\ell(\boldsymbol{\lambda})$.*

Our analysis of convergence, given in the Appendix, essentially follows the form of that given in [42] (also [43]), which was seminal for similar analyses in [23,44]. It is of some interest perhaps to note that this particular method of proof differs from that used in [21] to prove convergence of the EM algorithm in the MLE context. In that case, because the data log-likelihood is not strictly concave, there may exist multiple global maxima. This additional complexity of the optimization function necessitates the use of deeper technical conditions than we do in the Appendix, drawing on results of [45]. Ultimately, of course, all of the above-referenced methods rely on certain fundamental conditions laid out in [46]. However, unlike [46], in all of these models for the Poisson inverse problem (as with our model, as well), it is possible that an optimal solution occur at the boundary of the parameter space. This additional complication leads to a non-trivial amount of technical changes in the method of proof, beyond those in [46], as can be seen from our own proof in the Appendix.

VI. Hyperparameter Selection

As with any Bayesian technique, it is important to consider the effect of the values of the hyperparameters of our prior density, $\{\delta, \gamma, \alpha_0, \ldots, \alpha_{J-1}\}$, on the quality of the final MAP estimate. Ideally, which choice of values to make should be influenced by prior information available to the scientist regarding the potential structure of $\boldsymbol{\lambda}$ at various scales j of aggregation. We argued earlier that the hyperparameters δ and γ of the gamma prior placed on the total intensity, $\lambda_{0,0}$, are not critical to the estimation process since in most practical problems of interest, the total number of counts is fairly large and hence the data will dominate the gamma prior. In fact, we saw that δ is not even involved in the convergence analysis, and in practice (and throughout this paper) we set this to a very small positive constant $\delta \ll 1$, which effectively eliminates its role in the MAP estimation process (see (17)). More crucial is the choice of the beta density hyperparameters which control the regularity and smoothness of the estimate. The hyperparameters $\{\alpha_j\}$ may be interpreted as regularization parameters. In Section IV, we noted that if $\alpha_j = 1,\ j = 0, \ldots, J-1$, then the MAP estimates of the splits $\{\rho_{j,m}\}$ coincide with the MLEs $\frac{x_{j+1,2m}}{x_{j,m}}$ (as is readily apparent from (18)). Setting $\alpha_j > 1$ tends to stabilize the estimates in low-count situations, pushing each MAP estimate (of $\rho_{j,m}$) away from the MLE and closer to $1/2$ (an even split indicative of smoothness or regularity in the intensity at that scale and position). Large settings for $\{\alpha_j\}$ tend to produce more smoothing.

Some additional insight into the role of the hyperparameters may be obtained by considering $\boldsymbol{\lambda}$ as a stochastic process, and examining its autocorrelation function. The handful of examples below demonstrate the usefulness of this approach in illustrating the richness of the class of possible models for $\boldsymbol{\lambda}$. Formally, consider the M-length vector $\boldsymbol{\lambda}$ as a stochastic process on the finite lattice $\{0, 1, \ldots, M-1\}$. For technical reasons, it is useful to extend this process to a shift-invariant analogue on the discrete, M-point circle, which may be accomplished formally by placing a discrete uniform prior on the set of possible shifts $s \in \{0, 1, \ldots, M-1\}$ [6,16]. In this context, [16] proves that $\boldsymbol{\lambda}$ is a stationary process. The autocorrelation function, say $r(\tau)$, of $\boldsymbol{\lambda}$ may be expressed in a non-trivial but closed-form expression, and calculated in a straightforward manner, for fixed choice of M and the hyperparameters [6].

Of course, there is a tremendous degree of flexibility in choosing which combination of hyperparameters to examine graphically, through $r(\tau)$. Here the result of Lemma 2 suggests an interesting point of departure. Figure 2 shows plots of $r(\tau)$ for three choices from the space of hyperparameters – one from that region in which the conditions of Lemma 2 are satisfied, one from that region in which they are not, and one at the boundary of these two regions. In the first case a process with short-range negative correlation is induced, in the second case, a process with short-range positive correlations, and in the last case, a process with zero correlation across all lags. Although the results of Figure 2 are only illustrative, they suggest that the condition for the log-prior density function to be concave in $\boldsymbol{\lambda}$ actually delineates a local separation of the space of hyperparameters $\{\gamma, \alpha_0, \ldots, \alpha_{J-1}\}$ by a hyperplane into two sections, corresponding to $\boldsymbol{\lambda}$ processes with positive or negative short-range dependencies. This issue is currently being studied in greater depth by the authors.

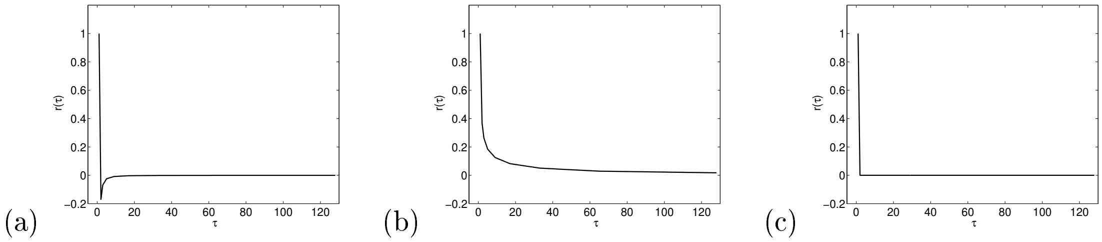

Fig. 2. Autocorrelation $r(\tau)$ for shift-invariant stochastic process $\boldsymbol{\lambda}$. Three cases shown are examples corresponding to hyperparemeter choices inside, outside, and on the boundary of the region in which the conditions of Lemma 2 are satisfied. (a) $\alpha_j = 3^{J-j-1}$, for $j = 0, \ldots, J-1$, and $\gamma = 3\alpha_0$ (left-hand side of (26) greater than zero). (b) $\alpha_j = 1^{J-j-1}$, and $\gamma = 1\alpha_0$ (left-hand side of (26) less than zero). (c) $\alpha_j = 2^{J-j-1}$, and $\gamma = 2\alpha_0$ (left-hand side of (26) equal to zero).

¿From a practical perspective, the boundary case illustrated in Figure 2(c) is quite useful. If existing prior knowledge is insufficient for knowing whether positive and/or negative autocorrelation in $\boldsymbol{\lambda}$ is likely, a zero-correlation model may be an acceptable option. Along these lines, we have obtained very satisfactory results using the following approach (e.g., see Section VII). First, set α_{J-1} in proportion to $A = (\sum_n y_n)/M$, the average counts per (reconstructed) intensity (also, this setting is restricted so as to be greater than or equal to 1). Then set $\alpha_j = 2\alpha_{j+1}$, $j = 0, \ldots, J-2$ and $\gamma = 2\alpha_0$, This particular scheme for selecting the hyperparameters reduces the J free hyperparameters to just one key parameter α_{J-1}, which gives adequate control over the behavior of the MAP estimator while maintaining the concavitity we desire.

On a final note, we mention that previous results in [6] suggest another interesting class of priors, obtained by setting $\alpha_j = C$, $C \geq 1$, $j = 0, \ldots, J-1$. With this choice, the induced intensity prior has $1/f$ spectral characteristics. Assume that $\alpha_j = C \geq 1$, and define $\nu \equiv -1 - \log_2 M_2(C)$, where $M_2(C) = \int \rho^2 \mathcal{B}e(\rho \,|\, C, C) d\rho$, the second moment of the beta prior density. Note that $1/4 < M_2(C) < 1/2$, where the lower and upper bounds corresponds to the two extreme limits of the beta density: a point mass at $1/2$ or two point masses at 0 and 1, respectively. This implies $0 < \nu < 1$. It can then be shown that the autocorrelation function of the intensity prior induced by our multiscale prior is approximated as follows.

$$r(\tau) \approx C_1 |\tau|^{(\nu-1)} + C_2 |\tau| 2^{(\nu-2)J}, \tag{28}$$

where C_1 and C_2 are constants. For large J and $\nu < 1$, the term $C_2|\tau|2^{(\nu-2)J}$ is negligible, and hence the correlation function behaves like $|\tau|^{(\nu-1)}$ and the power spectrum decays like $\frac{1}{|f|^\nu}$ (see [47] for the relationship between autocorrelation functions and power spectrums of $1/f$ processes). This may be very relevant to intensity analysis since there is convincing empirical evidence that natural intensity functions such as images have similar spectral characteristics; see the comprehensive study in [48]. Also, this suggests another guideline for hyperparameter selection. If we have some prior knowledge of the spectral decay rate of the underlying intensity function we seek, then C may be set accordingly. However, as indicated by the convergence analysis in Section V, setting all beta hyperparameters to a common constant value will not, in general, produce a concave log-posterior. Conversely, in general, hyperparameter settings that satisfy the necessary and sufficient conditions of Lemma 2 do not generate a prior density with $1/f$ spectral decay.

VII. Applications

A. Astronomical Energy Spectra

The analysis of energy spectra is a standard task in high-energy astrophysics. The ultimate goal is to identify and label spectral lines observed in association with some object of interest. In astrophysical spectroscopy stellar objects are studied and the data take the form of photon counts at different energy levels (in units of electron-volts (eV)). In high-energy astrophysics it is necessary to escape the earth's atmosphere to obtain proper measurements, a task that therefore falls upon satellite instruments. Due to the geometry of the instruments, among other things, a "blurring" is introduced into the measurement process. Since the arrival of high-energy photons, say at the X-ray and gamma-ray levels, typically is well-represented by a Poisson process, the "de-blurring" and estimation of the underlying spectra may be viewed as a Poisson inverse problem. Calibration of the blurring effect may be accomplished on the ground, prior to the launching of a satellite, from which knowledge of the transition matrix $\mathbf{P}$ is established.

Figure 3(a) depicts a theoretical energy spectrum, corresponding to the production of gamma rays by energetic particles interacting with the ambient solar atmosphere [49]. Figure 3(b) shows a collection of Poisson counts corresponding to this spectrum, simulated as if having been observed by the COMPTEL instruments [50] on board the Compton Gamma Ray Observatory (CGRO). Due to the underlying physics of the measurement devices on COMPTEL, the true energy of a photon entering the instruments has a good chance of actually being recorded at some lower energy level. This is immediately apparent from comparison of Figures 3(a) and (b). Recovery of the spectra ($\boldsymbol{\lambda}$, in the notation of this paper) thus corresponds, in a sense, to redistribution of the counts across higher energy levels, according to the matrix $\mathbf{P}$.

Figure 3(c) shows the estimate of $\boldsymbol{\lambda}$ recovered from the data by the multiscale method proposed in the preceding sections. Because it is difficult to specify explicitly, from the underlying physics, to what degree and in what manner positive and/or negative correlations might exist in $\boldsymbol{\lambda}$, we use a set of hyperparameter settings that yield a zero-correlation model. Since ideally such spectra are viewed simply as a collection of "lines," this is perhaps not an entirely unrealistic model to use here. In examining the estimate obtained, note that most of the locations of the spectral peaks, as well as their relative heights and widths, are reasonably well-captured.

(a) (b) (c)

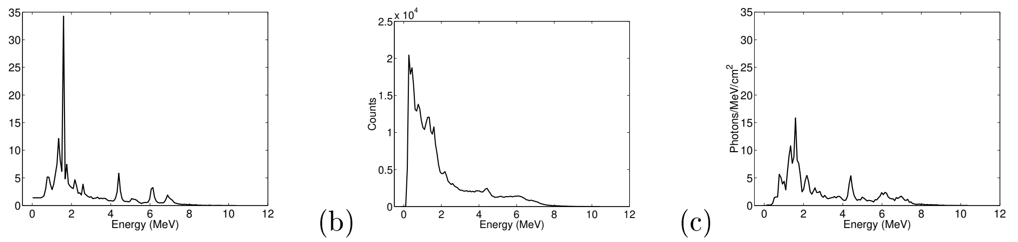

Fig. 3. Bayesian multiscale MAP estimation in astronomical spectral analysis. (a) Theoretical energy spectrum, in the 0.6 - 11 MeV energy range, corresponding to the production of gamma-rays during a solar flare ($M = 128$). (b) Counts simulated from the spectrum in part (a), as they might be observed by the COMPTEL instruments on NASA's CGRO ($N = 128$). (c) MAP estimate of underlying spectrum, produced from observed counts and the hyperparameter settings $\alpha_{J-1} = 1.01$, $\alpha_j = 2\alpha_{j-1}$, $j = 0, \ldots, J-2$, and $\gamma = 2\alpha_0$. Convergence was declared when $||\boldsymbol{\lambda}^{(k+1)} - \boldsymbol{\lambda}^{(k)}||_2 / ||\boldsymbol{\lambda}^{(k)}||_2 < 10^{-6}$, which required just over 100 iterations. [The authors thank Dr. Alex Young, UNH, for simulating the data in this example.]

B. Nuclear Medicine Imaging

Here we consider the application of our multiscale framework to emission computed tomography (ECT). In medical ECT, a human subject is injected with a radioactive pharmaceutical specifically designed for absorption in certain bodily organs or tissues. The distribution of this pharmaceutical within the subject can provide functional and/or anatomical diagnostic information. To obtain a mapping of pharmaceutical uptake, data are collected by detecting gamma-ray photons that are emitted from within the subject as the pharmaceutical decays. From these *projection* data (the indirect data $\mathbf{y}$ in our problem), we wish to estimate the underlying pharmaceutical distribution (intensity $\boldsymbol{\lambda}$). The probability transition matrix $\mathbf{P}$ is derived from the physics and geometry of the detection device and data collection process [21].

In ECT, the intensity of interest is usually a two or three dimensional object, and our basic multiscale framework can be easily extended to multidimensional settings like this. The following simple extension was proposed in [6]. To illustrate the extension, we focus on two-dimensional problems, but the same ideas can be used in higher dimensions as well. In two dimensions (2-d) the Haar multiscale data analysis of the emission is as follows. We begin with 2-d data $\{x_{k,l}\}$, $k,l = 0,\ldots,2^J-1$, and define

$$\begin{aligned} x_{J,k,l} &\equiv x_{k,l}, \quad k,l=0,\ldots,2^J-1 \\ x_{j,k,l} &= x_{j+1,2k,2l}+x_{j+1,2k+1,2l}+x_{j+1,2k,2l+1}+x_{j+1,2k+1,2l+1}, \\ &\qquad\qquad k,l=0,\ldots,2^j-1,\ 0\le j\le J-1. \end{aligned}$$

Again, the index j refers to the resolution of the analysis, 2^j; $j = J$ and $j = 0$ correspond to the highest (finest) and lowest (coarsest) resolutions (scales), respectively. We take the 2-d multiscale splits to be the factors corresponding to the *multiplicative* refinement of a coarse intensity into four finer intensities by first splitting it horizontally (vertically) into two halves, then next vertically (horizontally) splitting each half into two quarters as described by [6]. That is, take

$$\begin{aligned} \lambda_{j-1,k,l} &= \lambda_{j,2k,2l}+\lambda_{j,2k,2l+1}+\lambda_{j,2k+1,2l}+\lambda_{j,2k+1,2l+1}, \\ \rho^1_{j-1,k,l} &= \frac{\lambda_{j,2k,2l}+\lambda_{j,2k,2l+1}}{\lambda_{j,2k,2l}+\lambda_{j,2k,2l+1}+\lambda_{j,2k+1,2l}+\lambda_{j,2k+1,2l+1}}, \\ \rho^2_{j-1,k,l} &= \frac{\lambda_{j,2k,2l}}{\lambda_{j,2k,2l}+\lambda_{j,2k,2l+1}}, \\ \rho^3_{j-1,k,l} &= \frac{\lambda_{j,2k+1,2l}}{\lambda_{j,2k+1,2l}+\lambda_{j,2k+1,2l+1}}. \end{aligned} \tag{29}$$

Alternating vertical and horizontal splitting in this fashion effectively maps the 2-d problem into a 1-d multiscale representation which can be handled by Bayesian framework above. Alternatively, it is possible to consider a fully 2-d refinement process in which we simultaneously split a coarse scaling coefficient into four finer coefficients. In this case the conditional parent-child likelihoods would be multinomially instead of binomial, and the natural conjugate prior would be the Dirichlet rather than the beta density, but otherwise the multiscale framework would be essentially the same.

In Figure 4 we illustrate the application of our multiscale framework to a simulated ECT problem. The underlying 2-d intensity in our simulation is the common Shepp-Logan model, a standard benchmark in ECT. The intensity $\boldsymbol{\lambda}$ is a 64×64 square image shown in Figure 4(a). The transition probability matrix $\mathbf{P}$, corresponding to a *parallel strip-integral geometry* with 80 radial samples and 60 angular samples distributed uniformly over 180°, was generated by the *ASPIRE* software system [51]. $\mathbf{P}$ was applied to $\boldsymbol{\lambda}$ to obtain $\boldsymbol{\mu}$, and we used a standard Poisson random number generator to synthesize the projection data $\mathbf{y}$. Several multiscale MAP reconstructions based on our multiscale prior are shown in Figures 4(c)-(d). For comparison, in Figure 4(b) we also show the very best likelihood-based reconstruction obtained by stopping the likelihood-based EM algorithm at the very best reconstruction; that is, the reconstruction having the smallest squared error, which is impossible to determine in practice since the true intensity is, of course, unknown. The multiscale MAP EM algorithms converge to satisfactory reconstructions, comparable in quality to that obtained by the stopped likelihood-based EM algorithm. One potential advantage of our multiscale approach is that the effect of hyperparameter settings on the reconstruction quality is fairly interpretable, whereas stopping rules for the classical EM approach are notoriously difficult to analyze.

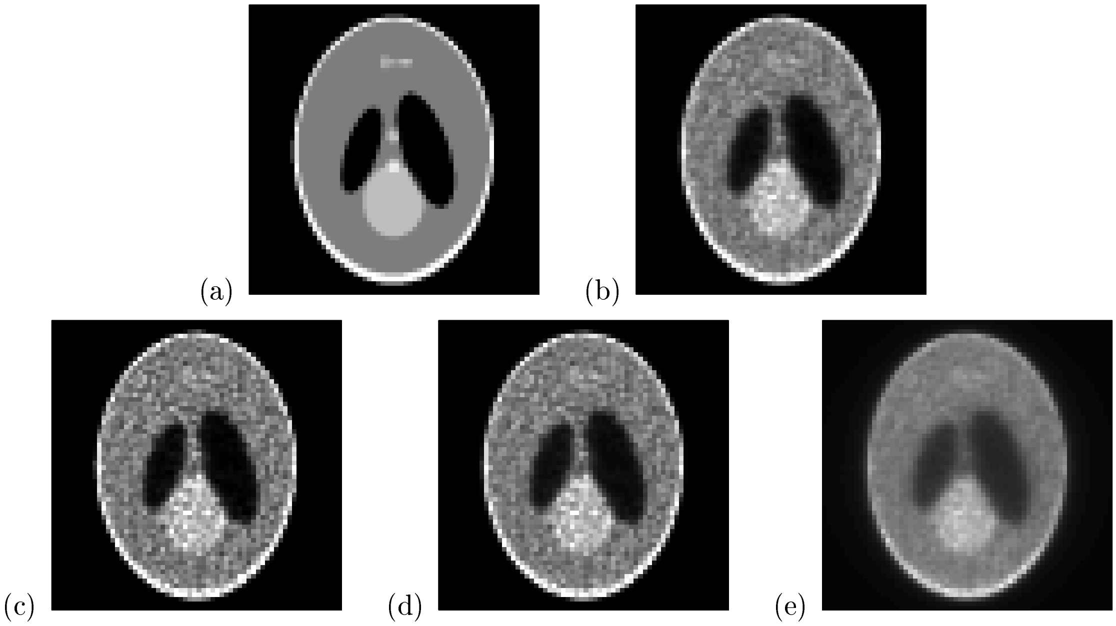

Fig. 4. *Bayesian multiscale MAP reconstruction in emission computed tomography. (a) Shepp-Logan software phantom intensity; total counts in simulated projections* $= 1.6 \times 10^6$. *(b) Likelihood-based EM reconstruction (stopped after 44 iterations (optimal clairvoyant stopping at minimum squared error reconstruction); average squared pixel error at present reconstruction = 0.78). (c) Multiscale MAP EM reconstruction* ($\alpha_{J-1} = 1.01$ *(average count per pixel/340) and* $\alpha_j = 2\alpha_{j-1},\ j = 0,\ldots,J-2, \gamma = 2\alpha_0$ *(essentially the weakest prior that insures concavity of the log-posterior); convergence in 97 iterations; average squared pixel error = 1.13). (d) Multiscale MAP EM reconstruction* ($\alpha_{J-1} = 3.44$ *(average count per pixel/100) and* $\alpha_j = 2\alpha_{j-1},\ j = 0,\ldots,J-2$, $\gamma = 2\alpha_0$*; convergence in 81 iterations; average squared pixel error = 1.02). (e) Multiscale MAP EM reconstruction* ($\alpha_{J-1} = 34.4$ *(average count per pixel/10) and* $\alpha_j = 2\alpha_{j-1},\ j = 0,\ldots,J-2$, $\gamma = 2\alpha_0$*; convergence in 34 iterations; average squared pixel error = 2.69). In all cases, convergence was declared when* $||\boldsymbol{\lambda}^{(k+1)} - \boldsymbol{\lambda}^{(k)}||_2/||\boldsymbol{\lambda}^{(k)}||_2 < 10^{-6}$.

VIII. Conclusions and Extensions

This paper introduced a new Bayesian multiscale framework for linear inverse problems involving Poisson data. The foundation of our framework is a multiscale factorization of the Poisson likelihood function, induced by recursive aggregation (or partitioning) of the data space, and the resulting re-parameterization of the underlying intensity function, which directly captures how the intensity is to be "split" at each location-scale combination. Conjugate priors are used in the multiscale parameter space to constrain the manner in which these "splits" may occur. The hyperparameters of the prior can be selected to insure that the log-posterior is strictly concave, which allowed us to develop an EM algorithm that is guaranteed to converge to the global MAP estimate. Furthermore, the prior has a simple interpretation and it can be easily tailored to reflect prior belief as to the smoothness of the unknown intensity. This class of priors also has the interesting feature that a "non-informative" member yields the traditional maximum likelihood solution. Moreover, unlike many other Bayesian approaches (*e.g.,* based on Gauss-Markov random field priors), our MAP solution is guaranteed to be non-negative. From a computational perspective, our EM algorithm is comparable to the most efficient techniques currently available since our EM update equations have simple, closed-form expressions. The potential of the new framework was examined in astronomical energy spectral analysis and emission tomography applications.

Extensions of the general framework outlined in this paper include translation-invariant (TI) implementations and wider classes of prior densities. TI implementations partially overcome the strict dyadic partioning underlying the basic multiscale model. Additionally, TI estimates are more regular (approximately piecewise linear) than the piecewise constant estimates associated with traditional Haar multiscale analysis. In terms of prior models, it should be possible to incorporate more sophisticated information including known boundaries of regions and the fusion of information from other imaging modalities. For example, structural information from a magnetic resonance image (MRI) could be built into the multiscale prior by selecting more informative beta priors (supporting either smoothness or an edge at each position and scale) according to a complementary multiscale decomposition of the MRI.

As another extension, one can replace the beta priors with beta mixtures. Beta mixture densities provide a mechanism for more sophisticated modeling of singularities. For example, a two-component beta mixture consisting of a uniform density ($\alpha = 1$) and a beta density sharply peaked at $1/2$ ($\alpha = 1000$) could represent the possibility of an "edge" or very homogeneous structure, respectively. It has also been demonstrated that the MAP criterion associated with certain beta mixture priors coincides with a Minimum Description Length complexity regularization [52]. Essentially the same EM algorithm can be used to find a (local) MAP estimate in conjunction with such priors, but, of course, the log-posterior is non-concave in these cases. It is also possible to account for correlations between nearby splits using the hidden Markov model we recently proposed for Poisson imaging problems [18, 53]. Such models may provide more effective modeling of continuous boundaries within images.

We focused on a standard EM algorithm in this paper, but it is quite possible to implement our framework using less expensive variants of EM that may provide even faster convergence such as SAGE [54] or OSEM [55]. For example, in our recent work in SPECT [56] we employed the OSEM algorithm to optimize our multiscale MAP criterion.

Finally, it remains to study the practical performance of our multiscale framework in greater depth than we have attempted to here in section VII. Comparative studies with, for example, standard Gibbs priors, including analysis of trade-offs between bias-variance and resolution-noise, are needed to further explore the potential of the new approach.

Appendix A
Convergence Analysis

Proof of Lemma 1: Recall that we assume $\boldsymbol{\mu}$ is $N \times 1$, $\boldsymbol{\lambda}$ is $M \times 1$, and $\mathbf{P}$ is $N \times M$, where $M = 2^J$ for some $J > 0$ (note that no such restriction is placed on N).

The multiscale prior density defined in (15) has the specific form

$$p(\lambda_{0,0}, \boldsymbol{\rho}) = \frac{\delta^\gamma}{\Gamma(\gamma)} \lambda_{0,0}^{\gamma-1} \exp\{-\delta\lambda_{0,0}\} \prod_{j=0}^{J-1} \prod_{m=0}^{M_j-1} \frac{1}{B(\alpha_j,\alpha_j)} (\rho_{j,m})^{\alpha_j-1}(1-\rho_{j,m})^{\alpha_j-1} .$$

By the standard change-of-variables formula, we have

$$h(\boldsymbol{\lambda}) = p\left(\lambda_{0,0}, \boldsymbol{\rho}(\boldsymbol{\lambda})\right) \cdot \mathcal{J}^{-1},$$

where $\mathcal{J}$ is the Jacobian of the transformation $(\lambda_{0,0}\boldsymbol{\rho}) \rightarrow \boldsymbol{\lambda}$. It may be shown that

$$\mathcal{J} \equiv \lambda_{0,0}^{M_J-1} \prod_{j=0}^{J-2} \prod_{m=0}^{M_j-1} \left[\rho_{j,m}(1-\rho_{j,m})\right]^{M_{J-j-1}-1} , \tag{30}$$

and use of (30) and the relation $\rho_{j,m} \equiv \lambda_{j+1,2m}/\lambda_{j,m}$ can be seen to yield the result in equation (25). Therefore, proof of Lemma 1 is reduced to proof of the expression in equation (30).

To show this we use proof by induction, starting with the case $J = 2$. In this case we have

$$\begin{aligned} \mathcal{J}_2 &= \det\left(\frac{\partial}{\partial(\lambda_{0,0},\boldsymbol{\rho})}\boldsymbol{\lambda}(\lambda_{0,0},\boldsymbol{\rho})\right) \\ &= \left|\begin{array}{cccc} \rho_{0,0}\rho_{1,0} & \lambda_{0,0}\rho_{1,0} & \lambda_{0,0}\rho_{0,0} & 0 \\ \rho_{0,0}(1-\rho_{1,0}) & \lambda_{0,0}(1-\rho_{1,0}) & -\lambda_{0,0}\rho_{0,0} & 0 \\ (1-\rho_{0,0})\rho_{1,1} & -\lambda_{0,0}\rho_{1,1} & 0 & \lambda_{0,0}(1-\rho_{0,0}) \\ (1-\rho_{0,0})(1-\rho_{1,1}) & -\lambda_{0,0}(1-\rho_{1,1}) & 0 & -\lambda_{0,0}(1-\rho_{0,0}) \end{array}\right| . \end{aligned}$$

Using the fact that a matrix and its transpose share the same determinant, and applying the method of cofactors to the third and fourth columns, we find that

$$\begin{aligned} \mathcal{J}_2 &= -\lambda_{0,0}^2\rho_{0,0}(1-\rho_{0,0}) \left[\left|\begin{array}{cc} \rho_{0,0}(1-\rho_{1,0}) & (1-\rho_{0,0})(1-\rho_{1,1}) \\ \lambda_{0,0}(1-\rho_{1,0}) & -\lambda_{0,0}(1-\rho_{1,1}) \end{array}\right| \right. \\ &\quad + \left|\begin{array}{cc} \rho_{0,0}\rho_{1,0} & (1-\rho_{0,0})(1-\rho_{1,1}) \\ \lambda_{0,0}\rho_{1,0} & -\lambda_{0,0}(1-\rho_{1,1}) \end{array}\right| + \left|\begin{array}{cc} \rho_{0,0}(1-\rho_{1,0}) & (1-\rho_{0,0})\rho_{1,1} \\ \lambda_{0,0}(1-\rho_{1,0}) & -\lambda_{0,0}\rho_{1,1} \end{array}\right| \\ &\quad \left. + \left|\begin{array}{cc} \rho_{0,0}\rho_{1,0} & (1-\rho_{0,0})\rho_{1,1} \\ \lambda_{0,0}\rho_{1,0} & -\lambda_{0,0}\rho_{1,1} \end{array}\right| \right] . \end{aligned}$$

In other words, the existence of only four non-zero terms in the third and fourth columns has been used to reduce the determinant calculation for a 4×4 matrix to that of four 2×2 matrices. Some algebra and repeated use of the equality $\rho_{j,m} + (1-\rho_{j,m}) = 1$ for terms at scale $j = 1$ yields the result $\mathcal{J}_2 = \lambda_{0,0}^3\rho_{0,0}(1-\rho_{0,0})$.

More generally, we assume that (30) holds for $J = L-1$ and show that this implies the case $J = L$ holds as well. Define $\mathcal{I}_L \equiv \partial\boldsymbol{\lambda}/\partial(\lambda_{0,0},\boldsymbol{\rho})$ to be the matrix of partial derivatives. Next, note that the rows of the first 2^{L-1} columns of $\mathcal{I}_L$ can be obtained from the rows of $\mathcal{I}_{L-1}$ (i.e., a $2^{L-1} \times 2^{L-1}$ matrix) in the following fashion. For $n = 0, 1, \ldots, 2^{L-1}$, the first 2^{L-1} elements in rows $2n$ and $2n+1$ are equal to the elements of the n-th row of $\mathcal{I}_{L-1}$ multiplied by $\rho_{L-1,n}$ and $1-\rho_{L-1,n}$, respectively. This follows from the fact that $\boldsymbol{\lambda}$ at scale L is a refinement of $\boldsymbol{\lambda}$ at scale $L-1$, obtained simply by splitting each component λ_n of the latter into two pieces $\lambda_n\rho_{L-1,n}$ and $\lambda_n(1-\rho_{L-1,n})$.

Furthermore, since there are exactly 2^{L-1} of these new splitting factors in moving from $J = L-1$ to $J = L$, it is differentiation of $\boldsymbol{\lambda}$ (at scale L) with respect to these factors that yields the last 2^{L-1} columns of $\mathcal{I}_L$. However, for the reasons just mentioned, each $\rho_{L-1,n}$ is a variable in two and only two components of $\boldsymbol{\lambda}$. Therefore, the last 2^{L-1} columns of $\mathcal{I}_L$ each contain 2 nonzero terms and $2^L - 2$ zero terms. The nonzero terms at rows $2n$ and $2n+1$ are given by

$$\lambda_{0,0} \prod_{j=0}^{L-2} \prod_{m(n)} \rho_{j,m(n)}(1-\rho_{j,m(n)}) \ ,$$

preceded by $+1$ or -1, respectively, where the rightmost product is over all ancestral locations $m(n)$ of n.[1]

Now we adopt the same approach as in the proof for $J = 2$, using the transpose of $\mathcal{I}_L$ and applying the method of cofactors. The result is to reduce the calculation of the determinant for our $2^L \times 2^L$ matrix $\mathcal{I}_L$ to one of calculating a weighted sum of $2^{2^{L-1}}$ determinants of matrices of size 2^{L-1}. However, the weights for each of these determinants is the same, and can be shown to be equal to

$$\lambda_{0,0}^{M_{L-1}} \prod_{j=0}^{L-2} \prod_{m=0}^{M_j - 1} \left[\rho_{j,m}(1-\rho_{j,m})\right]^{M_{L-j-2}} \ .$$

Moreover, each of the determinants can be expressed in the form

$$\kappa_{i_0}\kappa_{i_1}\cdots\kappa_{i_{2^{L-1}-1}} \cdot |\mathcal{I}_{L-1}| \ ,$$

which we write as $\boldsymbol{\kappa}_i \mathcal{J}_{L-1}$, where the elements of $\boldsymbol{\kappa}_i$ are $\kappa_{i_m} = \rho_{L-1,m}$ or $1-\rho_{L-1,m}$. Finally, there is a certain symmetry in the $2^{2^{L-1}}$ possible values of $\boldsymbol{\kappa}_i$ in that each value with, say, $\kappa_{i_m} = \rho_{L-1,m}$ in the m-th location will have one and only one mate that contains the same components in all locations but the m-th, in which it has $\kappa_{i_m} = 1-\rho_{L-1,m}$ instead.

Combining the elements of the above discussion, we see that we can write

$$\begin{aligned}
\mathcal{J}_L &= \left[\lambda_{0,0}^{M_{L-1}} \prod_{j=0}^{L-2} \prod_{m=0}^{M_j-1} \left[\rho_{j,m}(1-\rho_{j,m})\right]^{M_{L-j-2}}\right] \cdot \left[\sum_{i=1}^{2^{2^{L-1}}} \boldsymbol{\kappa}_i \mathcal{J}_{L-1}\right] \\
&= \left[\lambda_{0,0}^{M_{L-1}} \prod_{j=0}^{L-2} \prod_{m=0}^{M_j-1} \left[\rho_{j,m}(1-\rho_{j,m})\right]^{M_{L-j-2}}\right] \cdot \mathcal{J}_{L-1} \cdot \left[\sum_{i=1}^{2^{2^{L-1}}} \boldsymbol{\kappa}_i\right] \\
&= \left[\lambda_{0,0}^{M_{L-1}} \prod_{j=0}^{L-2} \prod_{m=0}^{M_j-1} \left[\rho_{j,m}(1-\rho_{j,m})\right]^{M_{L-j-2}}\right] \cdot \mathcal{J}_{L-1} \\
&= \lambda_{0,0}^{M_{L-1}} \prod_{j=0}^{L-2} \prod_{m=0}^{M_j-1} \left[\rho_{j,m}(1-\rho_{j,m})\right]^{M_{L-j-1}-1} \ .
\end{aligned} \tag{31}$$

The transition from the second to the third lines above is accomplished by exploiting the above mentioned symmetry in the $\boldsymbol{\kappa}_i$'s to show that their sum is 1. This completes our proof. □

[1]The term *ancestral* refers to the binary tree graphical representation of the multiscale parameterization. A location j, m is an ancestor of $L-1, n$ if $m2^j \in \{n2^{L-1}, \ldots, (n+1)2^{L-1} - 1\}$.

Proof of Lemma 2: Begin by noting that in (25) each term to the right of the double product involves summations of components of $\boldsymbol{\lambda}$ at both scales j and $j+1$. Separating out terms strictly by scale yields the expression

$$\begin{aligned}\log h(\boldsymbol{\lambda}) &= -\delta\lambda_{0,0} + [\gamma - M_J - 2(\alpha_0 - M_{J-1})]\log\lambda_{0,0} \\ &+ \sum_{j=1}^{J-1}[\alpha_{j-1} - M_{J-j} - 2(\alpha_j - M_{J-j-1})]\cdot\sum_{m=0}^{M_j-1}\log\lambda_{j,m} \\ &+ (\alpha_{J-1}-1)\cdot\sum_{m=0}^{M_J-1}\log\lambda_{J,m} \;+\; C \;, \qquad (32)\end{aligned}$$

where C is a constant not depending on $\boldsymbol{\lambda}$. ¿From (32) it follows that the Hessian matrix of the log-density can be expressed in the form $\partial^2\log h(\boldsymbol{\lambda})/\partial\boldsymbol{\lambda}^2 = -\sum_{j=0}^{J}\eta_j A_j$, where

$$\eta_j = \begin{cases}\gamma - 2\alpha_0, & \text{if } j=0\\ \alpha_{j-1}-2\alpha_j, & \text{if } j=1,\ldots,J-1\\ \alpha_{J-1}-1, & \text{if } j=J\end{cases} \qquad (33)$$

and A_j is a block diagonal matrix, containing M_j blocks of size $M_{J-j}\times M_{J-j}$. Within each A_j, the m-th block is fully composed of terms identically equal to $1/\lambda_{j,m}^2$, for $m=0,1,\ldots,M_j-1$.

As a result of the above, for any non-zero $\mathbf{x}\in\mathbb{R}^M$ we have

$$\mathbf{x}^T\,\frac{\partial^2\log h(\boldsymbol{\lambda})}{\partial\boldsymbol{\lambda}^2}\,\mathbf{x} = -\sum_{j=0}^{J}\eta_j\sum_{m=0}^{M_j-1}\left(\frac{x_{j,m}}{\lambda_{j,m}}\right)^2 \quad ,$$

where the $x_{j,m}$ are defined in analogy to the $\lambda_{j,m}$. Letting $c_j \equiv \sum_{m=0}^{M_j-1}\left(\frac{x_{j,m}}{\lambda_{j,m}}\right)^2$, we see that the necessary and sufficient condition for strict concavity is

$$\sum_{j=0}^{J}\eta_j c_j \;>\; 0.$$

Because $\boldsymbol{\lambda}\geq 0$, the $\{c_j\}$ satisfy the following inequalities:

$$c_0 \;\leq\; c_1 \;\leq\; \cdots \;\leq\; c_J.$$

To see this, it suffices to show that for $x_1, x_2\in\Re$ and $\lambda_1,\lambda_2\geq 0$

$$\frac{(x_1+x_2)^2}{(\lambda_1+\lambda_2)^2} \;\leq\; \frac{x_1^2}{\lambda_1^2}+\frac{x_2^2}{\lambda_2^2},$$

which is equivalent to the inequality

$$(x_1+x_2)^2 \;\leq\; (1+a)^2x_1^2 + (1+a^{-1})^2x_2^2,$$

with $a\geq 0$. Working with the right hand side,

$$\begin{aligned}(1+a)^2x_1^2+(1+a^{-1})^2x_2^2 &= x_1^2+x_2^2+a^2x_1^2+\frac{1}{a^2}x_2^2+2ax_1^2+\frac{2}{a}x_2^2,\\ &\geq x_1^2+x_2^2+2ax_1\frac{1}{a}x_2+2ax_1^2+\frac{2}{a}x_2^2,\\ &= (x_1+x_2)^2+2ax_1^2+\frac{2}{a}x_2^2,\\ &\geq (x_1+x_2)^2,\end{aligned}$$

where the step from the first to second line uses the fact that $y^2 + z^2 \geq 2yz$.

Now, setting $r_j \equiv \frac{c_j}{c_J}$, we write the necessary and sufficient condition for strict concavity as

$$\eta_J \ + \sum_{j=0}^{J-1} \eta_j r_j \ > \ 0,$$

from which the statement of the lemma immediately follows. □

Proof of Theorem 1: The proof follows as the result of a handful of smaller lemmas, which we establish in sequence.

Lemma 3 *(Monotonicity) Let the k-th iterate of the EM algorithm be denoted by $\boldsymbol{\lambda}^{(k)}$. Then $\ell(\boldsymbol{\lambda}^{(k+1)}) \geq \ell(\boldsymbol{\lambda}^{(k)})$, for all k.*

Proof of Lemma 3: Follows as in the case of the standard EM algorithm with an unpenalized likelihood (e.g., see [20, p. 82]). □

Lemma 4 *The iterates $\boldsymbol{\lambda}^{(k)}$ all belong to the same compact, convex set.*

Proof of Lemma 4: As in [42], *i.e.,* the result of Lemma 3 insures that all iterates belong to the set $\{\boldsymbol{\lambda} : \ell(\boldsymbol{\lambda}) \geq \ell(\boldsymbol{\lambda}^{(0)})\}$, which is compact because of the continuity of ℓ and the behavior of ℓ as $||\boldsymbol{\lambda}|| \to \infty$. Convexity follows from the concavity of ℓ.

Lemma 5 *The Euclidean distance $||\boldsymbol{\lambda}^{(k+1)} - \boldsymbol{\lambda}^{(k)}||$ tends to zero as $k \to \infty$.*

Proof of Lemma 5: The method of proof is nearly identical to that of Lemma 3 of [42]. Define

$$Q(\boldsymbol{\lambda}|\boldsymbol{\lambda}^{(k)}) \equiv E_{\boldsymbol{\lambda}^{(k)}} [\log\, g(\mathbf{z}|\boldsymbol{\lambda})|\mathbf{y}] + \log\, h(\boldsymbol{\lambda}) \ \ ,$$

where $g(\mathbf{z}|\boldsymbol{\lambda})$ is the complete data likelihood function. Then by a standard argument we have

$$\ell(\boldsymbol{\rho}^{(k+1)}) - \ell(\boldsymbol{\rho}^{(k)}) \geq Q(\boldsymbol{\rho}^{(k+1)}|\boldsymbol{\rho}^{(k)}) - Q(\boldsymbol{\rho}^{(k)}|\boldsymbol{\rho}^{(k)}) \ \ . \tag{34}$$

Using a second order Taylor series expansion, the right hand side of (34) may be written as

$$-\frac{1}{2}\left(\boldsymbol{\rho}^{(k+1)} - \boldsymbol{\rho}^{(k)}\right)^T \left[\frac{\partial^2}{\partial \boldsymbol{\rho}^2} Q(\boldsymbol{\rho}|\boldsymbol{\rho}^{(k)})_{|\boldsymbol{\rho}=\bar{\boldsymbol{\rho}}}\right] \left(\boldsymbol{\rho}^{(k+1)} - \boldsymbol{\rho}^{(k)}\right) \ \ , \tag{35}$$

where $\bar{\boldsymbol{\rho}}$ is on the line segment connecting $\boldsymbol{\rho}^{(k)}$ and $\boldsymbol{\rho}^{(k+1)}$.

Some algebra shows that the expression in (35) simplifies to 1/2 times

$$\sum_{m=0}^{M-1} \left(\lambda_m^{(k+1)} - \lambda_m^{(k)}\right)^2 \frac{z_{\cdot,m}^{(k)}}{\bar{\lambda}_m^2} \quad + \quad \sum_{j=0}^{J} \eta_j \left[\sum_{m=0}^{M_j-1} \left(\frac{\lambda_{j,m}^{(k+1)} - \lambda_{j,m}^{(k)}}{\bar{\lambda}_{j,m}}\right)^2\right] \ \ , \tag{36}$$

where $z_{\cdot,m}^{(k)} \equiv E_{\boldsymbol{\lambda}^{(k)}} [\sum_n z_{n,m}|\mathbf{y}]$. The first component is due to the conditional expectation of the complete-data log-likelihood, as in the standard maximum likelihood context, which was shown in [42] to be greater than or equal to $c||\boldsymbol{\lambda}^{(k+1)} - \boldsymbol{\lambda}^{(k)}||_2^2$, for some constant c strictly greater than zero independent of k. The second component in (36) is due to the log prior, and is non-negative under the conditions of Lemma 2. Therefore, it follows that

$$||\boldsymbol{\lambda}^{(k+1)} - \boldsymbol{\lambda}^{(k)}||^2 \leq c^{-1} \left[\ell(\boldsymbol{\lambda}^{(k+1)}) - \ell(\boldsymbol{\lambda}^{(k)})\right] \ .$$

By Lemma 3 and the boundedness of $\ell(\boldsymbol{\lambda})$ on $\{\boldsymbol{\lambda} : \ell(\boldsymbol{\lambda}) \geq \ell(\boldsymbol{\lambda}^{(0)})\}$, the right hand side above goes to zero as $k \to \infty$, and our result follows. □

Lemma 6 *Let $\boldsymbol{\lambda}^*$ be a limit point of the sequence $\{\boldsymbol{\lambda}^{(k)}\}$. Then for $\lambda_m^* > 0$,*

$$\frac{\partial \ell(\boldsymbol{\lambda}^*)}{\partial \lambda_m} = 0 \ .$$

Furthermore, there are only a finite number of such limit points.

Proof of Lemma 6: Argued analogous to the proof of Lemma 3 in [42]. The first part requires that we establish the continuity of $\partial Q(\boldsymbol{\lambda}|\boldsymbol{\lambda}')/\partial \lambda_m$ in $(\boldsymbol{\lambda}, \boldsymbol{\lambda}')$, for $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda}'$ with positive components in $\mathbb{R}^M$, which is straightforward. For the second part it may be argued that the number of limit points is bounded above by 2^M i.e., by the number of possible subsets of the M parameters in $\boldsymbol{\lambda}$. □

Lemma 7 *The set of limit points $\boldsymbol{\lambda}^*$ is compact and connected, and therefore consists of a single member.*

Proof of Lemma 7: Identical to that of Lemma 9 in [44]. Specifically, the result of our Lemma 5 implies that the set of limit points is compact and connected. Since by the second part of Lemma 6 above the elements of this set are also finite in number, they must consist of a single member in order to be connected. □

Finally, with the above lemmas established, Theorem 1 follows by showing that $\boldsymbol{\lambda}^{(\infty)}$ satisfies the Kuhn-Tucker conditions for our optimization problem. Specifically, for each component $\lambda_m^{(\infty)}$ we require that

$$\frac{\partial \ell\left(\boldsymbol{\lambda}^{(\infty)}\right)}{\partial \lambda_m} \begin{cases} = 0, & \text{if } \ \lambda_m^{(\infty)} > 0 \\ \leq 0, & \text{if } \ \lambda_m^{(\infty)} = 0 \end{cases} \tag{37}$$

The case of $\lambda_m^{(\infty)} > 0$ is already established in the first part of Lemma 6. The other case is argued using proof by contradiction, as in the proof of the analogous theorem in [42]. This completes our proof of the theorem. □

References

[1] D. L. Snyder and M. I. Miller, *Random Point Processes in Time and Space.* New York: Springer-Verlag, 1991.

[2] L. A. Shepp and Y. Vardi, "Maximum likelihood reconstruction for emission tomography," *IEEE Trans. Med. Imaging*, vol. 1, pp. 113–122, 1982.

[3] F. O'Sullivan, "A statistical perspective on ill-posed inverse problems," *Statistical Science.*, vol. 1, no. 4, pp. 502–527, 1986.

[4] M. Basseville, A. Benveniste, K. C. Chou, S. A. Golden, R. Nikoukhah, and A. S. Willsky, "Modeling and estimation of multiresolution stochastic processes," *IEEE Trans. Inform. Theory*, vol. 38, pp. 766–784, Mar. 1992.

[5] D. Donoho and I. Johnstone, "Ideal adaptation via wavelet shrinkage," *Biometrika*, vol. 81, pp. 425–455, 1994.

[6] K. Timmermann and R. Nowak, "Multiscale modeling and estimation of Poisson processes with application to photon-limited imaging," *IEEE Transactions on Information Theory*, vol. 45, no. 3, pp. 846–862, April, 1999.

[7] H. Krim and I. Schick, "Minmax description length for signal denoising and optimal representation," *IEEE Trans. on Information Theory*, vol. 45, no. 3, April, 1999.

[8] Special Issue on Wavelet Transforms and Multiresolution Signal Analysis, *IEEE Trans. Inform. Theory*, vol. 38, no. 2, March 1992.

[9] Special Issue on Multiscale Statistical Signal Analysis and its Applications, *IEEE Trans. Inform. Theory*, vol. 45, no. 3, April 1999.

[10] F. M. J.-L Starck and A. Bijaoui, *Multiscale Image Processing and Data Analysis.* Cambridge Univeristy Press, 1998.

[11] D. L. Donoho, "Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition," *App. and Comp. Harmonic Analysis*, vol. 2, pp. 101–126, 1995.

[12] F. Abramovich and B. W. Silverman, "Wavelet decomposition approaches to statistical inverse problems," *Biometrika*, vol. 85, pp. 115–129, 1998.

[13] M. R. Banham and A. K. Katsaggelos, "Spatially adaptive wavelet-based multiscale image restoration," *IEEE Trans. Image Processing*, vol. 5, no. 4, pp. 619–634, 1996.

[14] J. Z. G. Wang and G.-W. Pan, "Solution of inverse problem in image processing by wavelet expansion," *IEEE Trans. Image Processing*, vol. 4, no. 5, pp. 579–593, 1995.

[15] E. Kolaczyk, "A wavelet shrinkage approach to tomographic image reconstruction," *J. Amer. Statist. Assoc.*, vol. 91, pp. 1079–1990, 1996.

[16] E. Kolaczyk, "Bayesian multi-scale models for Poisson processes," *J. Amer. Statist. Assoc.*, vol. 94, pp. 920–933, 1999.

[17] E. Kolaczyk, "Some observations on the tractability of certain multi-scale models," in *Bayesian Inference in Wavelet-Based Models*, pp. 51–66, Springer-Verlag, 1999. Editors P. Müller and B. Vidakovic.

[18] R. Nowak, "Multiscale hidden Markov models for Bayesian image analysis," in *Bayesian Inference in Wavelet Based Models*, pp. 243–266, Springer-Verlag, 1999. Editors P. Müller and B. Vidakovic.

[19] J. A. O'Sullivan, R. E. Blahut, and D. L. Snyder, "Information-theoretic image formation," *IEEE Trans. Info. Theory*, vol. 44, pp. 2094–2123, 1998.

[20] G. McLachlan and T. Krishnan, *The EM Algorithm and Extensions.* New York: Wiley, 1997.

[21] Y. Vardi, L. A. Shepp, and L. Kaufman, "A statistical model for positron emission tomography," *J. Amer. Statist. Assoc.*, vol. 80, pp. 8–37, 1985.

[22] J. Llacer and E. Veklerov, "Feasible images and practical stopping rules for iterative algorithms in emission tomography," *IEEE Trans. Med. Imaging*, vol. 8, pp. 186–193, 1989.

[23] K. Lange, M. Bahn, and R. Little, "A theoretical study of some maximum likelihood algorithms for emission and transmission tomography," *IEEE Trans. Med. Imaging*, vol. 6, pp. 106–114, 1987.

[24] T. Hebert and R. Leahy, "A generalized EM algorithm for 3-d Bayesian reconstruction from Poisson data using Gibbs priors," *IEEE Trans. Med. Imaging*, vol. 8, no. 2, pp. 194–202, 1989.

[25] P. J. Green, "Bayesian reconstruction from emission tomography data using a modified EM algorithm," *IEEE Trans. Med. Imaging*, vol. 9, no. 1, pp. 84–93, 1990.

[26] J. A. Fessler and A. O. Hero, "Penalized maximum-likelihood image reconstruction using space-alternating generalized EM algorithms," *IEEE Trans. Image Processing*, vol. 4, no. 10, pp. 1417–1429, 1995.

[27] C. A. Bouman and K. Sauer, "A unified approach to statistical tomography," *IEEE Trans. Image Processing*, vol. 5, pp. 480–492, 1996.

[28] D. S. Lalush and B. M. W. Tsui, "Simulation evaluation of Gibbs prior distributions for use in maximum a posteriori SPECT reconstructions," *IEEE Trans. Med. Imaging*, vol. 11, pp. 267–275, 1992.

[29] M. Bhatia, W. C. Karl, and A. S. Willsky, "A wavelet-based method for multiscale tomographic reconstruction," *IEEE Trans. Med. Imaging*, vol. 15, no. 1, pp. 92–101, 1996.

[30] A. H. Delaney and Y. Bresler, "Multiresolution tomographic reconstruction using wavelets," *IEEE Trans. Image Processing*, vol. 4, pp. 799–813, 1995.

[31] F. Abramovich, T. Sapatinas, and B. W. Silverman, "Wavelet thresholding via a Bayesian approach," *J. Roy. Statist. Soc. Ser. B., 60, 725-749*, vol. 60, pp. 725–749, 1998.

[32] E. D. Kolaczyk, "Wavelet shrinkage estimation of certain Poisson intensity signals using corrected thresholds," *Statistica Sinica*, vol. 9, pp. 119–135, 1999.

[33] R. D. Nowak and R. G. Baraniuk, "Wavelet-based filtering for photon imaging systems," *IEEE Trans. Image Processing*, vol. 8, no. 5, pp. 666–678, 1999.

[34] R. Nowak and E. Kolaczyk, "A multiscale MAP estimation method for Poisson inverse problems," in *Proc. 32nd Asilomar Conf. Signals, Systems, and Comp.,* Pacific Grove, CA, pp. 1682–1686, IEEE Computer Society Press, 1998.

[35] S. S. Saquib, C. A. Bouman, and K. Sauer, "A multiresolution non-homogeneous MRF model for bayesian tomography," Preprint. Submitted to *IEEE Transactions on Image Processing*, 1999.

[36] S. Mallat, *A Wavelet Tour of Signal Processing.* San Diego, CA: Academic Press, 1998.

[37] S. L. Lauritzen, *Graphical Models.* Clarendon Press, 1996.

[38] C. Robert, *The Bayesian Choice: A Decision Theoretic Motivation.* New York: Springer-Verlag, 1994.

[39] R. D. Mauldin, W. D. Sudderth, and S. C. Williams, “Polya trees and random distributions,” *Ann. Stat.*, vol. 20, pp. 1203–1221, 1992.

[40] M. Lavine, “Some aspects of Polya tree distributions for statistical modelling,” *Ann. Stat.*, vol. 20, pp. 1222–1235, 1992.

[41] A. R. Depierro, “A modified expectation maximization algorithm for penalized likelihood estimation in emission tomography,” *IEEE Trans. Med. Imaging*, pp. 132–137, 1995.

[42] K. Lange and R. Carson, “EM reconstruction algorithms for emission and transmission tomography,” *J. Comp. Assist. Tomo.*, vol. 8, pp. 302–316, 1984.

[43] T. M. Cover, “An algorithm for maximizing expected log investment return,” *IEEE Trans. Inform. Theory*, vol. 30, pp. 369–373, 1984.

[44] K. Lange, “Convergence of em image reconstruction algorithm with Gibbs smoothing,” *IEEE Trans. Med. Imaging*, vol. 9, pp. 439–446, 1990.

[45] I. Csiszár and G. Tusnády, “Information geometry and alternating minimization procedures,” *Statistics and Decisions, Supplementary Issue No. 1*, pp. 205–237, 1984.

[46] C. F. J. Wu, “On the convergence properties of the EM algorithm,” *Annals of Statistics*, vol. 11, pp. 95–103, 1984.

[47] G. Wornell, *Signal Processing with Fractals. A Wavelet-Based Approach.* Englewood Cliffs, New Jersey: Prentice Hall, 1996.

[48] A. van der Schaaf and J. van Hateren, “Modelling the power spectra of natural images,” *Vision Research*, vol. 36, no. 17, pp. 2759–2770, 1996.

[49] R. J. Murphy, R. Ramaty, D. V. Reames, and B. Kozlovsky, “Solar abundances from gamma-ray spectroscopy - comparisons with energetic particle, photospheric, and coronal abundances,” *Astrophysical Journal*, vol. 371, pp. 793–803, 1991.

[50] V. Schoenfelder and et al., “Instrument description and performance of the imaging gamma-ray telescope COMPTEL aboard the compton gamma-ray observatory,” *Astrophysical Journal Supplement Series*, vol. 86, pp. 657–692, 1993.

[51] J. A. Fessler, “Aspire 3.0 user's guide: A sparse iterative reconstruction library,” Communication & Signal Processing Laboratory Technical Report No. 293, Department of Electrical and Computer Engineering, University of Michigan, Ann Arbor, 1998.

[52] R. Nowak and M. Figueiredo, “Unsupervised progressive parsing of Poisson fields using minimum description length criteria,” in *Proceedings of IEEE Conference on Image Processing*, Kobe, Japan, October, 1999.

[53] R. Nowak, “Multiscale hidden Markov models for photon-limited imaging,” *Proceedings of SPIE Conf. 3816 — Mathematical Modeling, Bayesian Estimation, and Inverse Problems*, Denver CO, July, 1999.

[54] J. A. Fessler and A. O. Hero, "Space alternating generalized expectation-maximization algorithm," *IEEE Trans. Signal Processing*, vol. 42, no. 10, pp. 2664–2677, 1994.

[55] H. M. Hudson and R. S. Larkin, "Accelerated image reconstruction using ordered subsets of projection data," *IEEE Trans. Med. Im.*, vol. 13, pp. 601–609, 1994.

[56] R. Nowak, E. Kolaczyk, D. Lalush, and B. Tsui, "A Bayesian multiscale framework for SPECT," in *Proceedings of IEEE Medical Imaging Conference*, (Seattle, WA, October), 1999.