A BAYESIAN MULTISCALE FRAMEWORK FOR POISSON INVERSE PROBLEMS

Robert Nowak

ECE Department, Michigan State University
East Lansing, MI 48824-1226
nowak@egr.msu.edu

Eric D. Kolaczyk

Department of Math & Stat, Boston University
Boston, MA 02215
kolaczyk@math.bu.edu

ABSTRACT

This paper describes a maximum *a posteriori* (MAP) estimation method for linear inverse problems involving Poisson data based on a novel multiscale framework. The framework itself is founded on a carefully designed multiscale prior probability distribution placed on the "splits" in the multiscale partition of the underlying intensity, and it admits a remarkably simple MAP estimation procedure using an expectation-maximization (EM) algorithm. Unlike many other approaches to this problem, the EM update equations for our algorithm have simple, closed-form expressions. Additionally, our class of priors 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.

1. INTRODUCTION

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

Bayesian methods have become increasingly popular for Poisson inverse problems because they enable the incorporation of prior knowledge about the underlying intensity to be recovered, *e.g.*, [3, 4]. The crucial element of Bayesian techniques is the choice of the prior probability model for the underlying intensity of interest. Many approaches are based on Markov random field (MRF) models [5], especially in imaging applications. The results obtained using classical MRF models are encouraging. However, good intensity models should be capable of representing discontinuities (*e.g.*, edges in images) and other inhomogeneous behavior. While this is possible within the MRF framework [5], inference based on inhomogeneous MRF models usually requires computationally intensive stochastic sampling algorithms.

To address these limitations, in this paper we develop a new Bayesian approach to Poisson inverse problems based on a novel multiscale prior probability model devised specifically for Poisson data. Our prior is capable of representing inhomogeneous behavior in a very natural and succinct manner. In fact, in certain configurations this prior is itself a non-Gaussian $1/f$ process, making it especially well suited to modeling a wide class of intensities. Moreover, our particular use of conjugacy in building the prior allows for simple and computationally efficient maximum *a posteriori* (MAP) estimates of the intensity to be computed using an expectation-maximization (EM) algorithm.

The paper is organized as follows. In Section 2, we give the basic problem statement. In Section 3, we re-formulate the basic problem within a new Bayesian multiscale framework. In section 4, we present an EM algorithm for computing a MAP estimate of the intensity. In Section 5, we study two numerical experiments, and some concluding remarks are made in Section 6.

This work was supported by the National Science Foundation, grant no. MIP-9701692

2. PROBLEM STATEMENT

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

$$c_n \sim \text{Poisson}(\mu_n), \quad n = 0, \ldots, N-1, \tag{1}$$

where Poisson(μ_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 probabilities). The problem is to estimate $\boldsymbol{\lambda}$ from the so-called *indirect data* $\mathbf{c} = \{c_n\}_{n=0}^{N-1}$. A classical application in which this problem arises is photon-limited imaging: photons are emitted (from the emission space) according to an intensity $\boldsymbol{\lambda}$; photons emitted from location m are detected (in the detection space) at position n with probability $p_{n,m}$. Such a scenario is faced routinely, for example, with satellite imaging in high-energy astrophysics.

It is well known that a maximum likelihood estimate (MLE) of $\boldsymbol{\lambda}$ can be obtained using the expectation-maximization (EM) algorithm [2]. However, because the variance of this MLE can be quite high, it is an unsatisfactory solution in many situations, particularly those involving very low counts. To mitigate this problem, several Bayesian procedures have been developed that use prior information and produce MAP estimates that are better than the MLE in many cases [6, 3, 4]. However, as mentioned in the introduction, most of these methods are based on classical MRF models that suffer from certain limitations.

3. BAYESIAN MULTISCALE FORMULATION

The development of our work is motivated in part by two observations. First, classical MRFs are not well matched to Poisson data problems. Second, in practice it is generally rather difficult to deal with standard non-homogeneous MRFs. Here we propose a new approach based on an extension of the multiscale models independently developed in [7, 8] for modeling non-negative Poisson intensity functions in the context of *direct* data (in contrast to the indirect problem described above).

3.1. Complete Data

In anticipation of an EM algorithm, suppose that we have the *unobservable* data [9]

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

The count $z_{n,m}$ is precisely the number of counts originating from location m in emission space that are detected at location n in detection space. The observed (indirect) data (1) are obtained as $c_n = \sum_m z_{n,m}$. Also note that the (direct) emission data, the counts emitted from each location m, are given by the sums $\{\sum_n z_{n,m}\}_m$, and that $\sum_n z_{n,m} \sim \text{Poisson}(\lambda_m)$.

Our objective is to estimate $\boldsymbol{\lambda}$ using a multiscale approach. Taking $N = M = 2^J$, for some $J > 0$, we define the following multiscale analysis of the (unobserved) direct or "hidden" data:

$$\begin{aligned} z_m^{(J-1)} &= \sum_{n=0}^{2^J-1} z_{n,2m} + z_{n,2m+1} \\ z_m^{(j)} &= z_{2m}^{(j+1)} + z_{2m+1}^{(j+1)} \quad j = J-2, \ldots, 0. \end{aligned} \tag{3}$$

The $\{z_m^{(j)}\}$ are called the multiscale direct data coefficients.[1]

3.2. Likelihood Function

It is natural with Poisson data to adopt a likelihood-based framework. Fundamental to our approach here is a multiscale factorization of the so-called complete-data likelihood function, in which both data and parameters are passed through a multiscale transformation. The transformation of the data was just described. For the intensity parameter $\boldsymbol{\lambda}$, set

$$\lambda_m^{(J)} \equiv \lambda_m \ , \tag{4}$$

and define, for $j = 0, 1, \ldots, J-1$,

$$\lambda_m^{(j)} = \lambda_{2m}^{(j+1)} + \lambda_{2m+1}^{(j+1)}, \;\; m = 0, \ldots, 2^j - 1 \ . \tag{5}$$

That is, J denotes the finest scale of analysis (resolution of $\boldsymbol{\lambda}$), and each "parent" intensity $\lambda_m^{(j)}$ is simply the sum of the intensities of its two children.

The *canonical* multiscale parameters in this problem are the intensity ratios

$$\rho_m^{(j)} = \frac{\lambda_{2m}^{(j+1)}}{\lambda_m^{(j)}}, \;\; j = 0, \ldots, J-1 \tag{6}$$

The $\boldsymbol{\rho} = \{\rho_m^{(j)}\}$ can be interpreted as factors governing the multiscale refinement of the intensities underlying the hidden data. The complete-data likelihood (a function of the unobservable data $\mathbf{z} = \{z_{n,m}\}_{n,m}$), can now be factorized as follows.

$$\begin{aligned} p(\mathbf{z}|\lambda_0^{(0)}, \boldsymbol{\rho}) \;\; = \;\; & \Pr(z_0^{(0)}) \;\times \\ & \prod_{j=0}^{J-2} \prod_{m=0}^{2^j-1} \Pr(z_{2m}^{(j+1)}, z_{2m+1}^{(j+1)} | z_m^{(j)}) \;\times \\ & \prod_{m=0}^{2^{(J-1)}-1} \Pr(z_{0,2m}, \ldots, z_{2^J-1,2m+1} | z_m^{(J-1)}). \end{aligned} \tag{7}$$

The first factor $\Pr(z_0^{(0)})$ is a Poisson mass function with parameter $\lambda_0^{(0)}$, the single intensity parameter at the coarsest analysis scale. The factors of the form $\Pr(z_{2m}^{(j+1)}, z_{2m+1}^{(j+1)} | z_m^{(j)})$ are binomial distributed with parameters $z_m^{(j)}$ and $\rho_m^{(j)}$. The factors of the form

$$\Pr(z_{0,2m}, \ldots, z_{2^J-1,2m+1} | z_m^{(J-1)})$$

are distributed

$$\mathcal{M}\left(p_{0,2m}\rho_m^{(J-1)}, \ldots, p_{2^J-1,2m+1}(1-\rho_m^{(J-1)})\right),$$

where $\mathcal{M}(\theta_1, \ldots, \theta_q)$ denotes the multinomial distribution with parameters $\theta_1, \ldots, \theta_q$. In deriving this result, we use the standard convention [9] that the rows of our our matrix $\mathbf{P}$ sum to 1 (e.g., interpreting $\mathbf{P}_{\cdot,m}$ as the density of detected counts from bin m).

3.3. Multiscale Intensity Prior Probability Model

We induce a prior probability model on the unknown intensity $\boldsymbol{\lambda}$ by specifying priors for the canonical multiscale parameters $\{\rho_m^{(j)}\}$. The $\rho_m^{(j)}$ are modeled as independent random variables distributed according to a mixture of symmetric beta densities

$$\rho_m^{(j)} \sim \sum_{i=1}^{q} p_i \, \mathcal{B}e(\rho_m^{(j)} | \alpha_i, \alpha_i), \tag{8}$$

where $p_i \geq 0$, $\sum_{i=1}^q p_i = 1$, and

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

denotes a symmetric beta density with parameter α, with $B(\cdot,\cdot)$ the standard beta function. This model was independently introduced in [7, 8] for modeling Poisson intensities.[2] In practice, we have found a mixture of two or three beta densities to provide a sufficiently rich model. A simple two-component mixture ($q = 2$) consists of a low variance component, *e.g.*, point mass at 1/2 ($\equiv \mathcal{B}e(\rho|\alpha,\alpha)$ as $\alpha \to \infty$), and a high variance component, *e.g.*, uniform density on $[0,1]$ ($\equiv \mathcal{B}e(\rho|1,1)$). By assigning a high prior probability, *e.g.*, $p_1 \approx 1$, to the low variance component, we can reflect a prior belief that the intensity is generally homogeneous, but may contain isolated singularities (corresponding to the high variance component). This simple multiscale intensity

[1] In fact, these coefficients can be seen to be simply the unnormalized Haar scaling coefficients of the direct data.

[2] We do not address the issue of modeling or estimating the single intensity $\lambda_0^{(0)}$ at the coarsest analysis scale, although this could be dealt with easily using a (conjugate) gamma prior or improper non-informative prior [10]. In practice, we use the maximum likelihood estimate $\sum_{m,n} z_{n,m}$, which is generally quite reliable; essentially the same estimate is obtained from a MAP procedure using a non-informative Bayesian prior.

prior has been applied to intensity estimation problems involving *direct* (in contrast to the indirect problem considered here) observations with great success [7, 11, 8]. Moreover, in certain cases this multiscale prior has $1/f$ spectral characteristics [11], making it a reasonable model for natural intensities.

3.4. MAP Estimation from Complete Data

The log posterior probability density, $\log p(\boldsymbol{\rho}|\mathbf{z})$, is (up to a constant) given by

$$\log p(\boldsymbol{\rho}|\mathbf{z}) \;=\; \log p(\mathbf{z}|\boldsymbol{\rho}) + \log p(\boldsymbol{\rho}) \qquad (9)$$

For certain priors $p(\boldsymbol{\rho})$, this expression is easily maximized with respect to $\boldsymbol{\rho}$.

First, consider the special case in which the $\boldsymbol{\rho} = \{\rho_m^{(j)}\}$ are modeled as independent beta distributed random variables (*i.e.*, single beta component, no mixture). In this case, the prior takes the form

$$p(\boldsymbol{\rho}) = \prod_{j=0}^{J-1}\prod_{m=0}^{2^j-1} \frac{1}{B(\alpha,\alpha)} \left(\rho_m^{(j)}\right)^{\alpha-1} \left(1-\rho_m^{(j)}\right)^{\alpha-1} \qquad (10)$$

The multiscale factorization of the likelihood (7) and the assumption of independence among the $\{\rho_m^{(j)}\}$ makes it possible to decouple the joint maximization into a separate maximizations over each individual parameter. Taking the log of (7) and maximizing (9) with respect to $\{\rho_m^{(j)}\}$ produces the following MAP estimates:

$$\begin{aligned} \widehat{\rho}_m^{(j)} &= \frac{z_{2m}^{(j+1)} + \alpha - 1}{z_m^{(j)} + 2(\alpha-1)}, \qquad 0 \le j < J-1 \\ \widehat{\rho}_m^{(J-1)} &= \frac{\sum_{n=0}^{2^J-1} z_{n,2m} + \alpha - 1}{z_m^{(J-1)} + 2(\alpha-1)} \end{aligned} \qquad (11)$$

To model inhomogeneous behavior more flexibly, consider a slightly more complex prior consisting of mixtures of a beta density and a point mass at $1/2$.

$$\begin{aligned} p(\boldsymbol{\rho}) \;=\; & \prod_{j=0}^{J-1}\prod_{m=0}^{2^j-1} \frac{p}{B(\alpha,\alpha)} \left(\rho_m^{(j)}\right)^{\alpha-1} \left(1-\rho_m^{(j)}\right)^{\alpha-1} \\ & + (1-p)\,\delta(\rho_m^{(j)} - 1/2) \end{aligned} \qquad (12)$$

where $0 < p < 1$ is the mixing probability, and δ denotes the point mass function. In this case, the maximization is only slightly more complicated. Again, exploiting the fact that the joint maximization can be factored into individual maximizations over each parameter, let us consider only the terms in the log posterior involving a single parameter $\rho_m^{(j)}$. We denote the corresponding function to be maximized by $l_{j,m}(\rho_m^{(j)})$.

$$\begin{aligned} l_{j,m}(\rho_m^{(j)}) \;=\; & z_{2m}^{(j+1)} \log(\rho_m^{(j)}) + \\ & (z_m^j - z_{2m}^{(j+1)}) \log(1-\rho_m^{(j)}) + \\ & \log\Big(\frac{p}{Be(\alpha,\alpha)} \left(\rho_m^{(j)}\right)^{\alpha-1} \left(1-\rho_m^{(j)}\right)^{\alpha-1} + \\ & (1-p)\,\delta(\rho_m^{(j)} - 1/2)\Big) \end{aligned} \qquad (13)$$

There are two cases to consider: $\rho_m^{(j)} = 1/2$ and $\rho_m^{(j)} \neq 1/2$. If $\rho_m^{(j)} \neq 1/2$, then the point mass term drops out from (13) and the maximizer is given by $\widehat{\rho}_m^{(j)}$ in (11). The overall maximizer, denoted $\widetilde{\rho}_m^{(j)}$, is thus determined by comparing the values $l_{j,m}(1/2)$ and $l_{j,m}(\widehat{\rho}_m^{(j)})$. Specifically,

$$\widetilde{\rho}_m^{(j)} = \begin{cases} 1/2, & \text{if } l_{j,m}(1/2) > l_{j,m}(\widehat{\rho}_m^{(j)}) \\ \widehat{\rho}_m^{(j)}, & \text{otherwise.} \end{cases} \qquad (14)$$

4. MAP ESTIMATION FROM OBSERVED DATA

In the development above, we operated as if the unobservable data $\mathbf{z}$ were available. Although untrue in practice, the EM algorithm may be used to iterate to the estimates just derived. Given the observed data $\mathbf{c}$ and an estimate $\widehat{\boldsymbol{\lambda}}$ (equivalently $\widehat{\boldsymbol{\rho}}$), we can easily compute the (conditional) expected value

$$\widehat{z}_{n,m} = E\left[z_{n,m} \mid \mathbf{c}, \widehat{\boldsymbol{\lambda}}\right] = \frac{c_m \widehat{\lambda}_m p_{n,m}}{\sum_{k=0}^{2^J-1} \widehat{\lambda}_k p_{k,m}} \qquad (15)$$

This expression enables the following EM algorithm to find a local MAP estimate of $\boldsymbol{\lambda}$.

E-Step: Given the data $\mathbf{c}$ and an estimate $\widehat{\boldsymbol{\lambda}}$, compute $\widehat{\mathbf{z}}$ according to (15).

M-Step: Given an estimate of the unobservable data $\widehat{\mathbf{z}}$, compute the MAP estimate $\widehat{\boldsymbol{\rho}}$ according to (11) or (14), depending on the chosen prior.

Using standard results from the theory of EM algorithms [9], it can be shown that iteration of the E-Step and M-Step leads to a local maximum of the posterior distribution of $\boldsymbol{\lambda}$ given the data $\mathbf{c}$. The first E-Step can be computed with an arbitrary positive initialization of $\widehat{\boldsymbol{\lambda}}$ (*e.g.*, $\widehat{\boldsymbol{\lambda}} \equiv 1$). Each subsequent E-Step is then calculated with $\widehat{\boldsymbol{\lambda}}$ constructed from the $\widehat{\boldsymbol{\rho}}$ found in the previous M-Step.[3]

The MAP estimation procedure described above has two other desirable properties. First, it is easily verified that, by construction, the resulting estimate is non-negative. Second, if we take the $\{\rho_m^{(j)}\}$ to be i.i.d. uniform on $[0,1]$, a special (non-informative) case of our prior, then we recover the classical MLE method [2].

5. EXPERIMENTAL RESULTS

To demonstrate the effectiveness of our multiscale Bayesian method, let us consider the following simulated Poisson inverse problems. Consider the intensity functions $\boldsymbol{\lambda}_1$ ('Blocks') and $\boldsymbol{\lambda}_2$ ('Bumps') in Figures 1 and 2 (a), respectively. Distorted intensities ($\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$) were generated by circularly convolving each intensity with a lowpass filter (5-point Hamming window) whose weights were [0.036, 0.241, 0.446, 0.241, 0.036]. Next, a realization of counts was generated in each case, $\mathbf{c}_i \sim \text{Poisson}(\boldsymbol{\mu}_i)$, $i = 1, 2$, as shown in Figures 1 and 2 (b), respectively. All quantities are 256×1 dimensional (*i.e.*, the dimensions of the emission and detection spaces are $M = N = 256$). These examples were designed to be representative of the type of data encountered in various photon-limited estimation problems. Figures 1 and 2 (c) depict the MLEs in each case, obtained using the classical EM approach [2]. Figures 1 and 2 (d) depict the MAP estimate[4] obtained using our new multiscale framework with a prior

$$\rho_m^{(j)} \sim 0.5\, Be(\rho_m^{(j)}|5,5) \;+\; 0.5\, \delta(\rho_m^{(j)} - 1/2), \qquad (16)$$

[3] Note that since the mapping $\{\lambda_0^{(0)}, \boldsymbol{\rho}\} \mapsto \boldsymbol{\lambda}$ is one-to-one, the MAP estimate of $\boldsymbol{\rho}$ generates the MAP estimate of $\boldsymbol{\lambda}$.

[4] The EM algorithm was initialized with $\widehat{\boldsymbol{\lambda}} \equiv 1$

for all j, m. The experiments depicted in Figures 1 and 2 were repeated in 50 independent trials; the results shown here were typical. In the MAP case, the EM algorithm typically converged in fewer than 25 iterations. Table 1 shows the averaged mean squared error of the data itself ('Counts'), the MLE, and the MAP estimator.[5] Remarkably, the MAP estimator performed very well despite the notable difference in the shapes of the intensity functions in these two cases.

Table 1: *MSE results for two test intensities.*

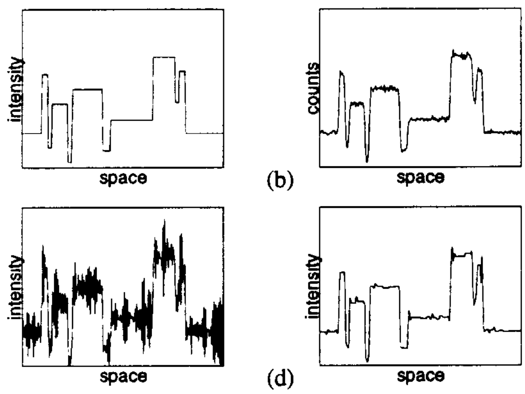
Intensity	*Counts*	*MLE*	*MAP*
λ_1 'Blocks'	0.0055	0.0882	0.0022
λ_2 'Bumps'	0.0528	0.6815	0.0270

6. CONCLUSIONS

In marked contrast to other Bayesian approaches to this problem based on classical MRFs, *e.g.*, [3, 4], or multiscale MRFs [12], our new multiscale framework admits a remarkably simple MAP estimation algorithm (while still producing estimates with impressive performance results). In fact, it is no more computationally intensive than the classical MLE approach pioneered in [2]. The simplicity of our method owes to two facts.

1. The complete-data likelihood function *factors* with respect to our particular choice of multiscale analysis, with a single element of the canonical multiscale parameter ρ accompanying each likelihood component.
2. The beta priors on the parameters are *conjugate* priors for the likelihood components, which leads to the simple closed-form expressions for the MAP estimates.[6]

Our method can also be easily extended to higher dimensions by using a multidimensional multiscale prior similar to that proposed in [11].

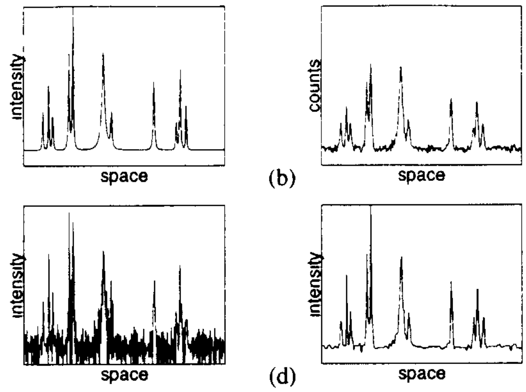


(a) (b) (c) (d)

Figure 1: *(a) Underlying intensity* λ_1, *'Blocks'. (b) Observations* c_1. *(c) MLE of* λ_1. *(d) MAP estimate of* λ_1.

[5] All MSEs are normalized by the squared norm of the underlying intensity function.

[6] The beta density is the *natural* conjugate density of the binomial distribution. Moreover, the marginal components of the multinomial are also binomial; hence, the beta prior plays a similar conjugate role. For more information on conjugate priors see [10].



(a) (b) (c) (d)

Figure 2: *(a) Underlying intensity* λ_2, *'Bumps'. (b) Observations* c_2. *(c) MLE of* λ_2. *(d) MAP estimate of* λ_2.

7. 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] 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.

[4] 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.

[5] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 6, no. 6, pp. 712–741, 1994.

[6] S. Geman and D. McClure, "Bayesian image analysis: An application to single photon emission tomography," in *Proc. Statist. Comput. Sect. Amer. Stat. Assoc.*, (Washington, DC), pp. 12–18, 1985.

[7] K. E. Timmermann and R. D. Nowak, "Multiscale Bayesian estimation of Poisson intensities," in *Proc. Asilomar Conf. Signals, Systems, and Comp.*, Pacific Grove, CA, IEEE Computer Society Press, 1997.

[8] E. D. Kolaczyk, "Bayesian multi-scale models for Poisson processes," *Technical Report 468, Department of Statistics, University of Chicago*, 1998.

[9] G. McLachlan and T. Krishnan, *The EM algorithm and extensions*. New York: Wiley, 1997.

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

[11] R. D. Nowak and K. E. Timmermann, "Stationary wavelet-based intensity models for photon-limited imaging," in *Proc. IEEE Int. Conf. on Image Proc.*, Chicago, 1998.

[12] S. S. Saquib, C. A. Bouman, and K. Sauer, "A non-homogeneous mrf model for multiresolution Bayesian estimation," *IEEE Int'l Conf. on Image Proc.*, 1996.