

Review

Algorithms for Non-Negatively Constrained Maximum Penalized Likelihood Reconstruction in Tomographic Imaging

Jun Ma

Department of Statistics, Macquarie University, North Ryde, New South Wales 2109, Australia;
E-Mail: jun.ma@mq.edu.au; Tel.: +61-2-9850-8548; Fax: +61-2-9850-7669

*Received: 28 November 2012; in revised form: 18 February 2013 / Accepted: 19 February 2013 /
Published: 12 March 2013*

Abstract: Image reconstruction is a key component in many medical imaging modalities. The problem of image reconstruction can be viewed as a special inverse problem where the unknown image pixel intensities are estimated from the observed measurements. Since the measurements are usually noise contaminated, statistical reconstruction methods are preferred. In this paper we review some non-negatively constrained simultaneous iterative algorithms for maximum penalized likelihood reconstructions, where all measurements are used to estimate all pixel intensities in each iteration.

Keywords: tomographic imaging; penalized likelihood; algorithms; constrained optimization

1. Introduction

Image reconstruction in medical imaging, in general, considers estimating pixel intensities or attenuations from measurements obtained from an imaging system. For example, for positron emission tomography (PET), the measurements are obtained according to the procedure summarized below; see [1,2] for more details. A type of radioactive isotope is introduced into the body of a patient and, due to the decay of radioisotope, it emits positrons. Each positron moves in the body for a small distance (usually less than 1 mm) and then interacts with an electron to produce a pair of gamma photons that travel in almost opposite directions. The scanning device in the imaging system can detect each pair of gamma photons with a certain probability and all such detections form the measurements that can appear in a histogram or a list form [3]. It is usually assumed that the detection probabilities are known and they can be pre-computed and stored or computed on-the-fly.

Note that a special feature of measurements is that they are contaminated by noises, which can be a severe problem particularly if each measurement is small in value due to dose safety limit. It is possible that, if the noises are not properly addressed, the reconstructed image can be distorted by excessive noises. For example, for low dose X-ray CT (a type of transmission tomography), the metal streak artifact (e.g., [4]) can be a severe problem for the traditional filtered backprojection method. Statistical iterative reconstruction methods, due to their ability to model the physics and measurements more accurately, are capable to reduce metal streak artifacts [5].

To deal with the noise contamination problem, statistical image reconstruction methods in emission, transmission, X-ray CT, *etc.* have been developed based on specified probability models for measurements. For example, for single photon emission computed tomography (SPECT), possible options include: weighted least squares (equivalent to variable variance Gaussian) [6], fixed variance Gaussian [7] and Poisson [8] models. These models can also be used for transmission scans. Since accidental coincidences are the main source of background noise in PET, most PET scans are precorrected for accidental coincidences by real-time subtraction of the coincidences in the delayed window [9]. For randoms-precorrected PET scans, possible measurement models are Gaussian, ordinary Poisson and shifted Poisson [9], and all of these are just approximations as the true probability density function (pdf) for the measurements is difficult. Shifted Poisson is also used to model X-ray CT measurements [10].

Different algorithms have been proposed to maximize their corresponding objective functions. For example, for emission tomography, the expectation-maximization (EM) algorithm [8] is designed to maximize the log-likelihood formulated from Poisson distributed measurements, or the iterative space reconstruction algorithm (ISRA) [7] for maximizing the log-likelihood formulated from Gaussian (with fixed variances) distributed measurements. An attractive aspect of both EM and ISRA is that they are very easy to implement and both respect the non-negativity constraint on the reconstructions. However, if the objective function contains a penalty term, which is normally used to smooth the reconstruction, then both EM and ISRA become impractical as they involve, in each iteration, a non-linear system of equations that is tedious to solve exactly due to the large number of unknowns in these equations. Moreover, the penalty function also adds an extra inconvenience when searching for a non-negative solution is desirable.

To simplify notations, both the measurements and the unknown image are lexicographically ordered into vectors. More specifically, we use $\mathbf{y} = (y_1, \dots, y_n)^T$ to present the measurement vector and $\mathbf{x} = (x_1, \dots, x_n)^T$ to denote the unknown image vector, where superscript T denotes matrix transpose. Note although the notations are unified for different reconstruction problems in this paper, the meaning of these notations, such as $\mathbf{x}$ and $\mathbf{y}$, can be different for different imaging modalities. Vectors $\mathbf{y}$ and $\mathbf{x}$ are related through a system matrix A ; see Equation (4) below for some examples. For tomographic reconstruction problems, matrix A is usually assumed known so its estimation is not covered by this paper. Rather, we focus on how to estimate $\mathbf{x}$ from the observed $\mathbf{y}$ and the known system matrix A . We denote the estimate of $\mathbf{x}$ by $\hat{\mathbf{x}}$.

Statistical reconstruction $\hat{\mathbf{x}}$ obtained by maximum penalized likelihood (MPL) (also known as maximum a posteriori (MAP)) is defined by

$$\hat{\mathbf{x}} = \arg \max_{\mathbf{x} \geq 0} \Psi(\mathbf{x}) \quad (1)$$

where $\Psi(\mathbf{x})$ is an objective function derived from the probability distribution for measurements and the penalty function. When the y_i 's are assumed independent (given $\mathbf{x}$), the penalized likelihood objective function is

$$\Psi(\mathbf{x}) = l(\mathbf{x}) - hJ(\mathbf{x}) \quad (2)$$

where $l(\mathbf{x})$ is the log-likelihood function given by

$$l(\mathbf{x}) = \sum_{i=1}^n l_i(\mu_i(\mathbf{x}); y_i) \quad (3)$$

Here $h > 0$ is the smoothing parameter and $J(\mathbf{x})$ is the penalty function used to smooth $\hat{\mathbf{x}}$. In Equation (3), l_i denotes the log-density function for measurement y_i , and μ_i is a function of $\mathbf{x} \in \mathbb{R}_+^p$ (here $\mathbb{R}_+^p$ denotes the non-negative orthant of $\mathbb{R}^p$) representing the mean measurement of camera bin i . Examples of μ_i include

$$\mu_i(\mathbf{x}) = \begin{cases} \eta_i(\mathbf{x}) + r_i & \text{emission} \\ b_i e^{-\eta_i(\mathbf{x})} + r_i & \text{transmission} \end{cases} \quad (4)$$

where $\eta_i(\mathbf{x}) = A_i \mathbf{x}$ with A_i being the i th row of matrix A , b_i is the known blank scan counts of the i th detector and r_i the known mean background counts. Another example is polyenergetic transmission scans (such as X-ray CT) where

$$\mu_i(\mathbf{x}) = \sum_{m=1}^M b_{im} e^{-A_i \mathbf{x}_m} + r_i \quad (5)$$

and here $\mathbf{x}_m = (x_{1m}, \dots, x_{pm})^T$ denotes the attenuation map corresponding to the m -th energy spectrum, $\mathbf{x}$ is a vector formed by the $\mathbf{x}_m$'s and b_{im} is the blank scan count from energy spectrum m .

In Equation (3) the notation $l_i(\mu_i; y_i)$ is used to emphasize that l_i is a function of μ_i and it also involves measurement y_i . We can also write this function as $l_i(\eta_i)$ or $l_i(\mathbf{x})$ in different contexts when there is no ambiguity. However, the functional properties of l_i may change with respect to its different arguments. For example, if assuming y_i follows a Poisson distribution for either emission or transmission scans, then

$$l_i(\mu_i) = -\mu_i + y_i \log \mu_i \quad (6)$$

This is clearly a concave function of μ_i for both emission and transmission cases. However, for $l_i(\mathbf{x})$ (treated as a function of $\mathbf{x}$), it may be no longer concave for transmission but still concave for emission scans. Concavity is an important property exploited by the optimization transfer algorithms.

Let $\boldsymbol{\mu}$ be an n -vector of all μ_i . The first term of Equation (2), *i.e.*, $l(\mathbf{x})$, measures similarity between $\mathbf{y}$ and $\boldsymbol{\mu}$. Different probability distributions have been used to model y_i even under the same imaging modality. For example, for emission tomography, if assuming the Poisson model for y_i (*i.e.*, $y_i \sim \text{Poisson}(\mu_i)$) then l_i is given by Equation (6), or if considering the weighted least squares then

$$l_i = -(y_i - \mu_i)^2 / w_i \quad (7)$$

where w_i is the weight. When $w_i = \mu_i$ we have the weighted least squares model as suggested in [11]. Another example in emission (or transmission) tomography is the randoms-precorrected PET scan (assume no scattering to simplify). In this context, the observed measurements are $y_i = y_i^{\text{Prompt}} - y_i^{\text{Delay}}$,

where y_i^{Prompt} and y_i^{Delay} (both unavailable directly) denotes the number of coincidences of the prompt and delayed windows respectively. Although we can assume $y_i^{\text{Prompt}} \sim \text{Poisson}(A_i \mathbf{x} + r_i)$ and $y_i^{\text{Delay}} \sim \text{Poisson}(r_i)$ and that they are independent, the exact distribution of y_i cannot be derived directly (e.g., [9]). An approximate probability model suggested in [9] is the shifted Poisson distribution, namely $y_i + 2r_i \sim \text{Poisson}(A_i \mathbf{x} + 2r_i)$, which gives

$$l_i = -(A_i \mathbf{x} + 2r_i) + (y_i + 2r_i) \log(A_i \mathbf{x} + 2r_i) \quad (8)$$

or the weighted least squares given by

$$l_i = -(y_i - A_i \mathbf{x}_i)^2 / (A_i \mathbf{x} + 2r_i) \quad (9)$$

Note that the shifted Poisson approximation matches the first two moments with the true probability model for $y_i + 2r_i$ when both the prompt and delayed measurements are assumed independent and follow Poisson distributions.

In this paper, we present and discuss several important non-negatively constrained penalized likelihood reconstruction algorithms. When designing a reconstruction algorithm in tomographic imaging, one considers the following important issues: (i) the algorithm is computationally efficient, and ideally it involves only forward-projection (e.g., $A\mathbf{x}$) and back-projection (e.g., $A^T \mathbf{y}$) operations; (ii) the algorithm can be easily applied to different measurement probability models and imaging modalities; (iii) the algorithm can impose the non-negativity constraint; (iv) the algorithm converges fast. Our discussions on the algorithms in this paper will mainly focus on these points.

In tomographic imaging, it is important to produce smoothed reconstructions as severe noise in a reconstruction can cause false diagnoses. Smoothing can generally be achieved by one of the following five practices: (i) early termination of the iterations (e.g., [12]); (ii) MPL reconstructions with an appropriate smoothing parameter (e.g., [13]); (iii) functional representation of the unknown image by a set of smooth basis functions (e.g., [14]); (iv) post smoothing of the reconstruction within each iteration (e.g., [15]) or after all iterations ([16]); and (v) pre-smoothing of the camera data (*i.e.*, sinogram) followed by filter backprojection (FBP) (e.g., [17,18]). We focus on the penalized likelihood approach to smoothing in this paper. In Equation (2), the smoothing parameter h balances two conflicting targets: fidelity of the μ_i s to the y_i s and smoothness of $\mathbf{x}$. Although an appropriate choice of h is important for achieving a reconstruction with *balanced* fidelity and smoothness, we will not consider how to estimate h in this paper. A penalty function $J(\mathbf{x})$ is used to smooth or regulate the estimate $\hat{\mathbf{x}}$. Usually, $J(\mathbf{x})$ takes the form of

$$J(\mathbf{x}) = \sum_{j=1}^p \rho(C_j \mathbf{x}) \quad (10)$$

where $C_j \mathbf{x}$ represents a neighborhood operation (such as the first or second order difference) on pixel j , and function $\rho(\cdot)$ measures the magnitude of $C_j \mathbf{x}$. A common choice of ρ is the quadratic function: $\rho(v) = \frac{1}{2}v^2$. Generally, a quadratic penalty tends to produce images with over-smoothed edges. Possible edge preserving penalties include total variation (TV) (e.g., [19]) Huber [20] and hyperbolic functions (e.g., [21]). Note that $\rho(\cdot)$ is convex for all these options.

The optimal choice of the penalty function J and the smoothing parameter h are unsolved problems in image processing and will not be further elaborated in this paper. We emphasize that smoothing by

MPL indeed produce visually improved reconstructions over the tradition filtered-backprojection method particularly in dose-limited tomography such as low dose X-ray CT. The edge preserving penalties are extremely useful, such as TV and Huber penalties; see [22–24]. However, the MPL reconstructions can have unnatural noise textures very different from the familiar filtered-backprojection method. Its impact on diagnostic tasks is still unknown and this is an active research area; see [25] for examples and discussions.

We adopt the following notations throughout this paper. Let $\mathbf{x}^{(k)}$ be the estimate of $\mathbf{x}$ obtained at iteration k of an algorithm. The notation $\nabla b(\cdot)$ indicates the derivative of function b with respect to the variable in the brackets. For example, $\nabla b(A_i \mathbf{x})$ represents the derivative of b with respect to $A_i \mathbf{x}$ and $\nabla b(\mathbf{x}; \mathbf{x}^{(k)})$ the derivative of b with respect to $\mathbf{x}$. We use $\nabla_j b(\mathbf{x})$ to denote the derivative of b with respect to x_j , the j -th element of vector $\mathbf{x}$. We also let $\nabla b(\mathbf{x}^{(k)})$ and $\nabla_j b(\mathbf{x}^{(k)})$ represent, respectively, $\nabla b(\mathbf{x})$ and $\nabla_j b(\mathbf{x})$ evaluated at $\mathbf{x} = \mathbf{x}^{(k)}$.

Non-negatively constrained MPL image reconstruction algorithms can be classified into simultaneous and block-iterative (a.k.a. ordered subset (OS)) algorithms. For simultaneous algorithms, all elements in $\mathbf{y}$ are used to update $\mathbf{x}$ in each iteration, and for block-iterative algorithms, distinct portions of $\mathbf{y}$ are used in turn to update $\mathbf{x}$. We discuss in this paper some simultaneous algorithms for non-negatively constrained MPL reconstructions, and the block-iterative algorithms are not included in our discussions. The rest of this paper is arranged as follows. The expectation-maximization algorithm for emission tomography is discussed in Section 2. Section 3 explains the alternating minimization algorithm designed specifically for transmission tomography. Section 4 contains explanations on the optimization transfer algorithms and their applications to tomographic reconstructions. The multiplicative iterative (MI) algorithms for tomographic imaging are provided in Section 5 and the Fisher scoring based Jacobi or Gauss–Seidel over-relaxation algorithms are presented in Section 6. Section 7 explains another Gauss–Seidel method named the iterative coordinate ascent algorithm. Finally, Section 8 includes discussions and remarks about this paper.

In this paper we focus on explaining and summarizing different non-negatively constrained tomographic imaging algorithms. Numerical comparisons of some of these algorithms are available in [26], and therefore will not be given in this paper.

2. EM Algorithm for Maximum Likelihood Reconstruction in Emission Tomography

The expectation-maximization (EM) algorithm [27] is a statistical algorithm for iteratively computing maximum likelihood estimates when data contain random missing values. Here “random” means these missing values do not provide extra information about the parameters we wish to estimate. We first give a brief summary of the EM algorithm below.

Since there exist the missing and the observed (or incomplete) components, we can define the complete data set as a combination of the incomplete and the missing data. Note, however, that our aim is to estimate the unknown parameters by maximizing the log-likelihood of the incomplete data. The rationale for the EM algorithm is that if maximizing the incomplete data likelihood is difficult while maximizing the complete data likelihood is easy, then EM can be used to compute iteratively

the maximum of the incomplete data likelihood by maximizing the complete data likelihood in each iteration.

Let $\mathcal{C}$ be the complete data set given by $\mathcal{C} = [\mathcal{Y}, \mathcal{M}]$, where $\mathcal{Y}$ denotes the incomplete data and $\mathcal{M}$ the missing data. Let $l_{\mathcal{C}}(\mathbf{x})$ be the log-likelihood based on the complete data $\mathcal{C}$ and $l(\mathbf{x})$ the log-likelihood of the incomplete data $\mathcal{Y}$, where $\mathbf{x}$ is a p -vector for the unknown parameters. Let $\hat{\mathbf{x}}$ be the maximum likelihood (ML) estimate of $\mathbf{x}$. Then iteration $k + 1$ of the EM algorithm comprises two steps:

1. **E-Step:** Compute the conditional expectation of the complete data log-likelihood given the incomplete data and $\mathbf{x}^{(k)}$, and denote this function by

$$Q(\mathbf{x}; \mathbf{x}^{(k)}) = E(l_{\mathcal{C}}(\mathbf{x}) \mid \mathcal{Y}, \mathbf{x}^{(k)}) \quad (11)$$

2. **M-Step:** Update the $\mathbf{x}$ estimate by maximizing the Q function, namely

$$\mathbf{x}^{(k+1)} = \underset{\mathbf{x}}{\operatorname{argmax}} Q(\mathbf{x}; \mathbf{x}^{(k)}) \quad (12)$$

One major advantage of EM is that it guarantees, under certain regularity conditions, that the incomplete data log-likelihood $l(\mathbf{x})$ increases in consecutive iterations before convergence. Note that EM requires availability of the Q function in a closed form; otherwise, a Monte-Carlo E-step can be used to replace the E-step [28].

The EM algorithm was first applied to emission tomograph by Shepp and Vardi [8] and Lange and Carson [29]. Both papers adopt the Poisson model for emission counts, namely y_i are independent Poisson random variables with mean $\mu_i = A_i \mathbf{x}$. This model assumes $r_i = 0$; otherwise, we can depict y_i as the value after subtracting r_i from the bin i measurement. From this Poisson model, we can formulate the complete data as $\mathcal{C} = \{y_{ij} : y_i = \sum_{j=1}^p y_{ij}\}$, where y_{ij} follows the Poisson distribution with mean $\mu_{ij} = a_{ij}x_j$. Clearly, each y_{ij} represents the unknown portion of measurement on camera bin i attributed to image pixel j . The corresponding complete data log-likelihood is

$$l_{\mathcal{C}}(\mathbf{x}) = \sum_{i=1}^n \sum_{j=1}^p \{-\mu_{ij} + y_{ij} \log \mu_{ij}\} \quad (13)$$

and the corresponding Q function is

$$Q(\mathbf{x}; \mathbf{x}^{(k)}) = \sum_{i=1}^n \sum_{j=1}^p \left\{ -\mu_{ij} + y_{ij}^{(k)} \log \mu_{ij} \right\} \quad (14)$$

where $y_{ij}^{(k)} = E(y_{ij} \mid y_i, \mathbf{x}^{(k)})$. Since the conditional distribution of $y_{ij} \mid y_i$ is $\text{Binomial}(y_i; \mu_{ij}/\mu_i)$, we have $y_{ij}^{(k)} = y_i a_{ij} x_j^{(k)} / \sum_{t=1}^p a_{it} x_t^{(k)}$. Thus after solving $\nabla_j Q(\mathbf{x}; \mathbf{x}^{(k)}) = 0$, the M-step of the EM algorithm gives the following updating formula for $\mathbf{x}$:

$$x_j^{(k+1)} = \frac{x_j^{(k)}}{\sum_{i=1}^n a_{ij}} \sum_{i=1}^n \frac{a_{ij} y_i}{\sum_{t=1}^p a_{it} x_t^{(k)}} \quad (15)$$

for $j = 1, \dots, p$. It has been pointed out in [23,30] that formula (15) can also be explained by the Bayes conditional probability formula. This EM algorithm possesses the following properties making it attractive for emission tomography; they are:

1. If the initial $\mathbf{x}^{(0)} \geq 0$ then $\mathbf{x}^{(k)} \geq 0$ for all $k \geq 1$; *i.e.*, it automatically satisfies the non-negativity constraint on $\mathbf{x}$.
2. The algorithm is easy to implement as it only involves forward- and back-projections.
3. The updating formula in Equation (15) increases the incomplete data log-likelihood: $l(\mathbf{x}^{(k+1)}) \geq l(\mathbf{x}^{(k)})$, where equality holds only when the iteration has converged.
4. $\mathbf{x}^{(k)}$ satisfies $\sum_i \mu_i^{(k)} = \sum_i y_i$, where $\mu_i^{(k)}$ is μ_i with $\mathbf{x} = \mathbf{x}^{(k)}$. Thus the $\mathbf{x}$ estimate at any iteration satisfies that the total expected and the total observed counts are equal.

The above EM is easy to implement and possesses some attractive properties on the reconstructions. This algorithm, however, is restricted only to emission tomography with Poisson distributed measurements. It cannot be easily extended to other reconstruction tasks. For example, application of the EM algorithm to transmission tomography does not lead to an exact updating formula due to the fact that its M-step does not produce a closed-form solution; see [29]. Another limitation is that this EM algorithm can only be used for maximum likelihood reconstructions, and its application to the MPL reconstruction will not in general result in closed-form updating formula. To rectify this problem, Green [31] developed a one-step-late (OSL) algorithm for the MPL reconstruction by replacing $\mathbf{x}$ in the derivative of the penalty function by its current estimate $\mathbf{x}^{(k)}$, and therefore an “exact” solution can still be accomplished. But this method suffers from the deficiencies that (i) the algorithm may be non-convergent; and (ii) some estimates may be negative.

De Pierro [32] reproduced the EM updating formula using a totally different argument. In his derivation, there is no missing data and hence no E-step. Although the algorithm is named “modified EM”, it is not a real EM. In fact, this algorithm belongs to a more general class called the optimization transfer algorithms, since the Poisson log-likelihood optimization problem is transferred to a simpler optimization in each iteration. We will summarize the optimization transfer algorithms in the Section 4.

3. Alternating Minimization Algorithms for Transmission Tomography

We have explained in Section 2 that the EM algorithm is not directly suitable for transmission scans as its M-step cannot be computed exactly. In this section, we summarize an alternating minimization algorithm designed to solve the transmission tomographic problem, including X-Ray CT. This algorithm is a generalization to the EM algorithm [33] and its application to transmission tomography can be found in [34].

Following [34], we explain this algorithm using the polyenergetic transmission tomography example. In this context, if assuming transmission scans follow Poisson distributions, the corresponding log-likelihood is

$$l(\mathbf{z}) = \sum_{i=1}^n \{y_i \log \mu_i(\mathbf{z}) - \mu_i(\mathbf{z})\} \quad (16)$$

where y_i is the scan count of detector i and μ_i (now expressed as a function of vector $\mathbf{z}$, which will be defined below) is given by Equation (5). Moreover, elements of the attenuation map associated with spectrum m , namely elements of $\mathbf{x}_m$ in Equation (5), are further modeled by

$$x_{mj} = \sum_{r=1}^a u_{mr} z_{rj} \quad (17)$$

where j indexes pixels, r represents different types of materials, u_{mr} are known linear attenuation coefficients and z_{rj} are the unknown partial densities (e.g., [34]) we wish to estimate. In Equation (16), $\mathbf{z}$ is a vector of size $pa \times 1$ formed by column-wise stacking the vectors $\mathbf{z}_j = (z_{1j}, \dots, z_{aj})^T$.

Define set

$$\mathcal{E} = \{q_{im}; i = 1, \dots, n \text{ and } m = 0, 1, \dots, M\} \quad (18)$$

where

$$q_{im} = b_{im} e^{-\sum_{j=1}^p a_{ij} \sum_{r=1}^a u_{mr} z_{rj}} \quad (19)$$

for $m = 1, \dots, M$ and q_{im} equals the background noise r_i for $m = 0$. Clearly, μ_i given in Equation (5) can now be expressed as $\mu_i = \sum_{m=0}^M q_{im}$. Define another set

$$\mathcal{L} = \{p_{im} : p_{im} \geq 0 \text{ and } \sum_m p_{im} = y_i; i = 1, \dots, n \text{ and } m = 0, 1, \dots, M\} \quad (20)$$

In [34], $\mathcal{E}$ is called the exponential family and $\mathcal{L}$ the linear family. Let $\mathbf{p}$ and $\mathbf{q}$ be the vectors created from p_{im} and q_{im} respectively. It can be shown that the problem of maximizing the log-likelihood Equation (16) can be re-written as

$$\max_{\mathbf{z}} l(\mathbf{z}) = \min_{\mathbf{q} \in \mathcal{E}} \min_{\mathbf{p} \in \mathcal{L}} \{I(\mathbf{p} \parallel \mathbf{q})\} \quad (21)$$

subject to $z_{rj} \geq 0$, where $I(\mathbf{p} \parallel \mathbf{q})$ is the I -divergence [35] given by

$$I(\mathbf{p} \parallel \mathbf{q}) = \sum_{i=1}^n \sum_{m=0}^M \left(p_{im} \log \frac{p_{im}}{q_{im}} - p_{im} + q_{im} \right) \quad (22)$$

Thus, maximizing the log-likelihood in Equation (16) can be achieved iteratively. Assuming the estimates $\mathbf{p}^{(k)}$, $\mathbf{q}^{(k)}$ and $\mathbf{z}^{(k)}$ are obtained at iteration k , then iteration $k+1$ contains two steps:

- (i) compute $\mathbf{p}^{(k+1)}$ by minimizing $I(\mathbf{p} \parallel \mathbf{q}^{(k)})$ subject to $\mathbf{p} \in \mathcal{L}$;
- (ii) compute $\mathbf{q}^{(k+1)}$ by minimizing $I(\mathbf{p}^{(k+1)} \parallel \mathbf{q})$ subject to $\mathbf{q} \in \mathcal{E}$.

Note that the second step is equivalent to minimizing $I(\mathbf{p}^{(k+1)} \parallel \mathbf{q})$ over $z_{rj} \geq 0$ with q_{im} being given by the expression in Equation (19).

Minimizing $I(\mathbf{p} \parallel \mathbf{q}^{(k)})$ over $\mathbf{p} \in \mathcal{L}$ is easily achieved using the Lagrange multiplier, and the result is

$$p_{im}^{(k+1)} = q_{im}^{(k)} \frac{y_i}{\sum_{m'=0}^M q_{im'}^{(k)}} \quad (23)$$

On the other hand, direct optimization of $I(\mathbf{p}^{(k+1)} \parallel \mathbf{q})$ over $z_{rj} \geq 0$ is an unmanageable task as the z_{rj} 's are mixed (i.e., not decoupled or separated from each other) within the objective function. One approach to overcome this problem is by using a decoupled objective function representing an upper bound of the original objective function. In fact, it can be shown that for q_{im} given by Equation (19),

$$I(\mathbf{p}^{(k+1)} \parallel \mathbf{q}) \leq \sum_{r=1}^a \sum_{j=1}^p \sum_{i=1}^n \sum_{m=0}^M \left(p_{im}^{(k+1)} a_{ij} u_{mr} z_{rj} + \hat{q}_{im} a_{ij} u_{mr} \frac{1}{v_0} e^{v_0(\hat{z}_{rj} - z_{rj})} \right) + \text{terms independent of } z_{rj} \quad (24)$$

where $v_0 = \max_{(i,m)} \sum_j \sum_r a_{ij} u_{mr}$ and $\hat{q}_{im}$ is an estimate of q_{im} corresponding to the estimate $\hat{z}_{rj} \geq 0$ of z_{rj} . This inequality is obtained from the fact that $I(\mathbf{p}^{(k+1)} \parallel \mathbf{q})$ is a convex function of z_{rj} . Clearly, z_{rj} on the right hand side of Equation (24) are decoupled and thus their non-negatively constrained optimizations will result in closed-form solutions. When we take $\hat{z}_{rj} = z_{rj}^{(k)}$, the optimal solution to z_{rj} is

$$z_{rj}^{(k+1)} = \max \left\{ 0, z_{rj}^{(k)} - \frac{1}{v_0} \log \left(\frac{\tilde{w}_r^{(k+1)}}{\hat{w}_r^{(k)}} \right) \right\} \quad (25)$$

where $\tilde{w}_r^{(k+1)} = \sum_i \sum_m a_{ij} u_{rm} p_{im}^{(k+1)}$ and $\hat{w}_r^{(k)} = \sum_i \sum_m a_{ij} u_{rm} q_{im}^{(k)}$. We give some remarks about this algorithm below.

Remarks

- (1) This algorithm is designed for maximum likelihood estimation. However, it can be easily extended to MPL where the penalty function must be convex and therefore can also be decoupled.
- (2) This algorithm is developed for the likelihood function derived from the simple Poisson measurement noise. Note that the alternating minimization algorithm was also developed for a compound Poisson noise model in [36] and its comparison with the simple Poisson alternating minimization was provided in [37]. For other measurement distributions, however, the corresponding algorithms have to be completely re-developed.
- (3) The convergence properties of the alternating maximization algorithm have been studied in [34]. Particularly, it is monotonically convergent under certain conditions.
- (4) It will become clear in Section 5 (Example 5.3) that the multiplicative-iterative algorithm can be derived more easily for this transmission reconstruction problem.
- (5) The trick of decoupling the objective function using its convex (or concave) property is also the key technique of the optimization transfer algorithms discussed in Section 4.

4. Optimization Transfer Algorithms

Details of the optimization transfer (OT) algorithm (also called the minorization–maximization (MM) algorithm for maximizations) can be found in, for example, [38]. In this section we present this algorithm briefly and explain its application in emission and transmission tomography.

The fundamental idea of the OT algorithm is that it employs a surrogate function to minorize (see the definition below) the objective function $\Psi(\mathbf{x})$ in each iteration, and then update the parameter estimate by maximizing this surrogate function.

More specifically, a function $\Phi(\mathbf{x}; \mathbf{x}^{(k)})$ is said to minorize $\Psi(\mathbf{x})$ at $\mathbf{x}^{(k)}$ if it satisfies the following “minorization” conditions:

- (i) $\Psi(\mathbf{x}^{(k)}) = \Phi(\mathbf{x}^{(k)}; \mathbf{x}^{(k)})$, and
- (ii) $\Psi(\mathbf{x}) \geq \Phi(\mathbf{x}; \mathbf{x}^{(k)})$ for all $\mathbf{x}$.

Then at iteration $k + 1$, $\mathbf{x}$ is estimated by maximizing $\Phi(\mathbf{x}; \mathbf{x}^{(k)})$, i.e.,

$$\mathbf{x}^{(k+1)} = \arg \max_{\mathbf{x} \geq 0} \Phi(\mathbf{x}; \mathbf{x}^{(k)}) \quad (26)$$

If the exact maximum is not easy to obtain, we can find an $\mathbf{x}^{(k+1)}$ by simply increasing $\Phi(\mathbf{x}; \mathbf{x}^{(k)})$, as this will also guarantee that the monotonic condition stated below remains for $\Psi(\mathbf{x})$.

An attractive property when using this surrogate function is that $\mathbf{x}^{(k+1)}$ satisfies the monotonic condition, namely

$$\Psi(\mathbf{x}^{(k+1)}) \geq \Psi(\mathbf{x}^{(k)}) \quad (27)$$

where equality holds only when the iteration has converged. This monotonic property can be easily verified by the minorization conditions since

$$\begin{aligned} \Psi(\mathbf{x}^{(k+1)}) &= \Phi(\mathbf{x}^{(k+1)}; \mathbf{x}^{(k)}) + \Psi(\mathbf{x}^{(k+1)}) - \Phi(\mathbf{x}^{(k+1)}; \mathbf{x}^{(k)}) \geq \Phi(\mathbf{x}^{(k)}; \mathbf{x}^{(k)}) + \Psi(\mathbf{x}^{(k)}) - \Phi(\mathbf{x}^{(k)}; \mathbf{x}^{(k)}) \\ &= \Psi(\mathbf{x}^{(k)}) \end{aligned}$$

For implementation of the OT algorithm to medical imaging, a surrogate function $\Phi(\mathbf{x}; \mathbf{x}^{(k)})$ must be determined. There exist different ways of choosing the surrogate function, such as those listed in [38]. We mainly consider two approaches in this paper: (i) the method based on the inequality on concave functions (called the concave inequality hereafter); and (ii) the method based on quadratic lower bounds (also known as paraboloidal surrogates [39]). These ideas are summarized below.

Let $G(\mathbf{x}) = \sum_{i=1}^n g_i(A_i \mathbf{x})$ be the objective function we wish to maximize, where A_i is the i -th row of matrix $A_{n \times p}$ and $\mathbf{x}$ is a p -vector. For matrix A , we assume its elements a_{ij} are non-negative and $\sum_j a_{ij} \neq 0$. We also assume that all $g_i(\cdot)$ are concave functions. Let $\pi_{ij} \geq 0$ be weights satisfying $\sum_{j=1}^p \pi_{ij} = 1$. Then according to the concave inequality we have

$$g_i(A_i \mathbf{x}) = g_i\left(\sum_{j=1}^p \pi_{ij} \frac{a_{ij} x_j}{\pi_{ij}}\right) \geq \sum_{j=1}^p \pi_{ij} g_i\left(\frac{a_{ij} x_j}{\pi_{ij}}\right) \quad (28)$$

There are different ways of choosing weights π_{ij} . For example, we can use $\pi_{ij} = a_{ij} x_j / A_i \mathbf{x}$, which is also adopted in [32]. In this case since each π_{ij} is a function of $\mathbf{x}$, the surrogate function corresponding to Equation (28) is

$$\Phi(\mathbf{x}; \mathbf{x}^{(k)}) = \sum_{j=1}^p \sum_{i=1}^n \frac{a_{ij} x_j^{(k)}}{A_i \mathbf{x}^{(k)}} g_i\left(\frac{A_i \mathbf{x}^{(k)}}{x_j^{(k)}} x_j\right) \quad (29)$$

and it is easy to verify that this surrogate satisfies the minorization conditions. The right hand side of Equation (29) is a weighted summation of functions g_i , each involving a single x_j only (*i.e.*, decoupled), and therefore maximization with respect to $\mathbf{x}$ of $\Phi(\mathbf{x}; \mathbf{x}^{(k)})$ can be achieved by a sequence of 1-D optimizations. Another trick, due to De Pierro [32], uses the following concave inequality:

$$g_i(A_i \mathbf{x}) = g_i\left(\sum_{j=1}^p \pi_{ij} \left[\frac{1}{\pi_{ij}} a_{ij} (x_j - x_j^{(k)}) + A_i \mathbf{x}^{(k)}\right]\right) \geq \sum_{j=1}^p \pi_{ij} g_i\left(\frac{1}{\pi_{ij}} a_{ij} (x_j - x_j^{(k)}) + A_i \mathbf{x}^{(k)}\right) \quad (30)$$

If the weights π_{ij} do not depend on x_j , then Equation (30) leads to the surrogate function of

$$\Phi(\mathbf{x}; \mathbf{x}^{(k)}) = \sum_{j=1}^p \sum_{i=1}^n \pi_{ij} g_i\left(\frac{a_{ij}}{\pi_{ij}} (x_j - x_j^{(k)}) + A_i \mathbf{x}^{(k)}\right) \quad (31)$$

which clearly also meets the minorization conditions. In Equation (31), the choice of π_{ij} is again flexible, and one popular option is to use $\pi_{ij} = a_{ij} / \sum_r a_{ir}$.

The above two surrogates are developed based on the concave inequality. Another useful approach is to employ a quadratic lower bound (e.g., [40]). Assume g_i is twice differentiable with its second derivative denoted by $\nabla^2 g_i$. Let $d_i^{(k)}$ be a number such that $d_i^{(k)} \leq \nabla^2 g_i(A_i \mathbf{x})$ for all $A_i \mathbf{x} > 0$, then

$$g_i(A_i \mathbf{x}) \geq g_i(A_i \mathbf{x}^{(k)}) + (\mathbf{x} - \mathbf{x}^{(k)})^T A_i^T \nabla g_i(A_i \mathbf{x}^{(k)}) + \frac{1}{2} (\mathbf{x} - \mathbf{x}^{(k)})^T A_i^T d_i^{(k)} A_i (\mathbf{x} - \mathbf{x}^{(k)}) \quad (32)$$

The right hand side of Equation (32) is a parabola surrogate of g_i and the condition on $d_i^{(k)}$ guarantees that this function lies below g_i . Unlike the previous surrogate functions, this surrogate is not separable in $\mathbf{x}$, and therefore its maximization with respect to $\mathbf{x}$ cannot be reduced to a series of 1-D problems. To overcome this problem we can find another function surrogating the above parabola surrogate but with separable $\mathbf{x}$. Towards this, we denote the right hand side quadratic function of Equation (32) by $q_i^{(k)}(A_i \mathbf{x})$. Since $q_i^{(k)}$ is concave in $A_i \mathbf{x}$, we can use either Equations (29) or (31) to find a surrogate to $q_i^{(k)}$ and the resulting algorithm is called the separable paraboloidal surrogate (SPS) algorithm [39]. For example, corresponding to Equation (31), a separable parabola surrogate of $q^{(k)}$ is

$$\Phi(\mathbf{x}; \mathbf{x}^{(k)}) = \sum_{j=1}^p \sum_{i=1}^n \pi_{ij} q_i^{(k)} \left(\frac{a_{ij}}{\pi_{ij}} (x_j - x_j^{(k)}) + A_i \mathbf{x}^{(k)} \right) \quad (33)$$

A careful selection of the curvature $b_i^{(k)}$ in Equation (32) can lead to fast convergence of the SPS algorithm. Erdoğlan and Fessler [39] derived the optimal curvature for the SPS algorithm in transmission tomography.

Next, we present two examples explaining how to implement the OT algorithm to emission and transmission tomography.

Example 4.1 (OT for emission scans with Poisson noise).

In this example we explain the application of OT for MPL reconstruction in emission tomography, where measurements are assumed to follow Poisson distributions. De Pierro's modified EM (MEM) [32] coincides with the method discussed below when $r_i = 0$. Firstly, under the Poisson model for emission scans, the penalized log-likelihood function is

$$\Psi(\mathbf{x}) = \sum_{i=1}^n \{-(A_i \mathbf{x} + r_i) + y_i \log(A_i \mathbf{x} + r_i)\} - h \sum_{t=1}^p \rho(C_t \mathbf{x}) \quad (34)$$

where ρ is assumed a convex function. Let

$$l_i(\eta_i) = -(\eta_i + r_i) + y_i \log(\eta_i + r_i) \quad (35)$$

where $\eta_i = A_i \mathbf{x}$. It is easy to verify that l_i is concave with respect to η_i , so we can use Equation (28) to define its surrogate function. On the other hand, for the penalty function in Equation (34), $-\rho$ is concave, so we can use Equation (31) to construct its surrogate. Combining them together we have the following surrogate for $\Psi(\mathbf{x})$:

$$\Phi(\mathbf{x}; \mathbf{x}^{(k)}) = \sum_{j=1}^p \left[\sum_{i=1}^n \frac{a_{ij} x_j^{(k)}}{\eta_i^{(k)}} l_i \left(\frac{\eta_i^{(k)}}{x_j^{(k)}} x_j \right) - h \sum_{t=1}^p \pi_{tj} \rho \left(\frac{c_{tj}}{\pi_{tj}} (x_j - x_j^{(k)}) + C_t \mathbf{x}^{(k)} \right) \right] \quad (36)$$

where $\pi_{tj} = c_{tj} / \sum_r c_{tr}$. Now

$$\nabla_j \Phi(\mathbf{x}; \mathbf{x}^{(k)}) = \sum_{i=1}^n a_{ij} \left(-1 + \frac{y_i}{x_j \eta_i^{(k)} / x_j^{(k)} + r_i} \right) - h \sum_{t=1}^p c_{tj} \nabla \rho \left(\frac{c_{tj}}{\pi_{tj}} (x_j - x_j^{(k)}) + C_t \mathbf{x}^{(k)} \right) \quad (37)$$

The equation $\nabla_j \Phi(\mathbf{x}; \mathbf{x}^{(k)}) = 0$ has a closed-form solution for x_j when $\rho(v) = v^2/2$ and $r_i = 0$ for all i . In this context, Equation (37) reduces to a quadratic function so we wish to solve for x_j from

$$\left(h \sum_{t=1}^p \frac{c_{tj}^2}{\pi_{tj}} \right) x_j^2 + \left[\sum_{i=1}^n a_{ij} + h \sum_{t=1}^p \left(c_{tj} C_t \mathbf{x}^{(k)} - \frac{c_{tj}^2}{\pi_{tj}} x_j^{(k)} \right) \right] x_j - x_j^{(k)} \sum_{i=1}^n a_{ij} \frac{y_i}{\eta_i^{(k)}} = 0 \quad (38)$$

subject to $x_j \geq 0$, and its analytic solution is readily available. If $r_i \neq 0$ or ρ is not quadratic, the analytic solution to Equation (37) does not exist. In this case, one can use an 1-D optimization method to solve it, or alternatively, one may use a separable parabola surrogate rather than Equation (36). An example of the latter is explained in the next example where the reconstruction problem is for transmission tomography.

Example 4.2 (OT for transmission scans with Poisson noise).

This example considers the application of OT to MPL reconstruction in transmission tomography. Our explanations follow [39] closely. For transmission scans with Poisson noise, the penalized log-likelihood is given by

$$\Psi(\mathbf{x}) = \sum_{i=1}^n \left\{ -(b_i e^{-A_i \mathbf{x}} + r_i) + y_i \log(b_i e^{-A_i \mathbf{x}} + r_i) \right\} - h \sum_{t=1}^p \rho(C_t \mathbf{x}) \quad (39)$$

where ρ is convex. Let $\eta_i = A_i \mathbf{x}$ and

$$l_i(\eta_i) = -(b_i e^{-\eta_i} + r_i) + y_i \log(b_i e^{-\eta_i} + r_i) \quad (40)$$

Since $l_i(\eta_i)$ is concave with respect to η_i , a separable parabola surrogate can be defined according to Equation (33). For the first term of Equation (39) (i.e., the log-likelihood part), a separable parabola is given by

$$\Phi_1(\mathbf{x}; \mathbf{x}^{(k)}) = \sum_{j=1}^p \sum_{i=1}^n \pi_{ij} q_i^{(k)} \left(\frac{a_{ij}}{\pi_{ij}} (x_j - x_j^{(k)}) + A_i \mathbf{x}^{(k)} \right) \quad (41)$$

where

$$q_i^{(k)}(\eta_i) = l_i(\eta_i^{(k)}) + \nabla l_i(\eta_i^{(k)}) (\eta_i - \eta_i^{(k)}) + \frac{1}{2} d_i^{(k)} (\eta_i - \eta_i^{(k)})^2 \quad (42)$$

and here $d_i^{(k)}$ satisfies $d_i^{(k)} \leq \nabla^2 l_i(\eta_i)$ for all $\eta_i \geq 0$. For the second term of Equation (39) (i.e., the penalty part), let $\gamma_t = C_t \mathbf{x}$ and let the weights $\xi_{tj} = c_{tj} / \sum_r c_{tr}$. Its separable parabola surrogate is

$$\Phi_2(\mathbf{x}; \mathbf{x}^{(k)}) = \sum_{j=1}^p \sum_{t=1}^p \xi_{tj} w_t^{(k)} \left(\frac{c_{tj}}{\xi_{tj}} (x_j - x_j^{(k)}) + C_t \mathbf{x}^{(k)} \right) \quad (43)$$

where

$$w_t^{(k)}(\gamma_t) = \rho(\gamma_t^{(k)}) + \nabla \rho(\gamma_t^{(k)}) (\gamma_t - \gamma_t^{(k)}) + \frac{1}{2} e_t^{(k)} (\gamma_t - \gamma_t^{(k)})^2 \quad (44)$$

Here $e_t^{(k)}$ is chosen such that $e_t^{(k)} \geq \nabla^2 \rho(\gamma_t)$ for all γ_t in its range; this curvature $e_t^{(k)}$ ensures that $w_t^{(k)}(\gamma_t)$ lies above $\rho(\gamma_t)$. Aggregating Equations (41) and (43) we obtain a separable parabola surrogate for $\Psi(\mathbf{x})$:

$$\Phi(\mathbf{x}; \mathbf{x}^{(k)}) = \Phi_1(\mathbf{x}; \mathbf{x}^{(k)}) - h\Phi_2(\mathbf{x}; \mathbf{x}^{(k)}) \quad (45)$$

We have

$$\begin{aligned} \nabla_j \Phi(\mathbf{x}; \mathbf{x}^{(k)}) = & \sum_{i=1}^n a_{ij} \left[\nabla l_i(\eta_i^{(k)}) + \frac{d_i^{(k)} a_{ij}}{\pi_{ij}} (x_j - x_j^{(k)}) \right] \\ & - h \sum_{t=1}^p c_{tj} \left[\nabla \rho(\gamma_t^{(k)}) + \frac{e_t^{(k)} c_{tj}}{\xi_{tj}} (x_j - x_j^{(k)}) \right] \end{aligned} \quad (46)$$

and for this example

$$\nabla l_i(\eta_i^{(k)}) = b_i e^{-\eta_i^{(k)}} \left(-\frac{y_i}{b_i e^{-\eta_i^{(k)}} + r_i} + 1 \right) \quad (47)$$

Let $a_{i.} = \sum_r a_{ir}$ and $c_{t.} = \sum_r c_{tr}$. The solution of $\nabla_j \Phi(\mathbf{x}; \mathbf{x}^{(k)}) = 0$, subject to $x_j \geq 0$, is given by $x_j^{(k+1)} = \max\{0, \tilde{x}_j^{(k+1)}\}$, where

$$\tilde{x}_j^{(k+1)} = x_j^{(k)} - \frac{\sum_{i=1}^n a_{ij} \nabla l_i(\eta_i^{(k)}) - h \sum_{t=1}^p c_{tj} \nabla \rho(\gamma_t^{(k)})}{\sum_{i=1}^n a_{ij} (a_{i.} d_i^{(k)}) - h \sum_{t=1}^p c_{tj} (c_{t.} e_t^{(k)})} \quad (48)$$

This is in fact a special gradient algorithm with a diagonal preconditioning matrix.

5. Multiplicative Iterative Algorithms

The OT algorithms presented in the last section have the following important achievements: (1) they manage to transform a high dimensional optimization problem into a series of 1-D optimizations; (2) due to 1-D optimizations, the non-negativity constraints can be easily enforced by simply resetting negative estimates to zero in each iteration; (3) the surrogate given by the separable parabola approach is general enough to be applicable to different tomographic reconstructions. A limitation of OT is that it requires all $l_i(\cdot)$ (log-density) and $-J(\cdot)$ (negative penalty) to be concave functions.

In this section we discuss a competitive alternative to the OT method called the multiplicative iterative (MI) algorithm; its application to tomographic imaging can be found in [26] and to box-constrained image processing in [41].

The main motivation of the MI algorithm is that it can be easily derived under different imaging modalities and different measurement noise models. Moreover, for some difficult penalties, such as TV, or even non-convex penalties [42], MI can be easily implemented to solve the corresponding optimization problems.

A general MI updating formula can be developed suitable for all tomographic reconstruction problems regardless of the mean function model, measurement probability distribution and penalty function. The simulation study reported in [26] reveals that MI has competitive convergence speed when compared with OT and other reconstruction algorithms. The MI algorithm does not require concavity of the functions l_i and $-J$ and therefore is more general than the OT algorithm. It requires existence of the first derivatives of $l_i(\cdot)$ and $J(\cdot)$. It is possible that the objective function $\Psi(\mathbf{x})$ in Equation (2) has multiple

local maxima. In this case, MI finds one of the local non-negative maxima, depending on the starting value of the algorithm.

Here are some notations needed to explain the MI algorithm. For a function $b(z)$, let $b(z)^+$ be the positive component of $b(z)$ and $b(z)^-$ the negative component so that $b(z) = b(z)^+ + b(z)^-$. For a number b , Let $[b]^+ = \max(0, b)$ and $[b]^- = \min(0, b)$ so that $b = [b]^+ + [b]^-$. Thus, for the numerical value of function $b(\cdot)$ at point z^* , we can also write $b(z^*) = [b(z^*)]^+ + [b(z^*)]^-$.

We develop the MI algorithm from the Karush–Kuhn–Tucker (KKT) necessary conditions for the non-negatively constrained optimization of $\Psi(x)$. They are:

$$\nabla_j \Psi(\mathbf{x}) = 0 \text{ if } x_j > 0 \text{ and} \quad (49)$$

$$\nabla_j \Psi(\mathbf{x}) \leq 0 \text{ if } x_j = 0 \quad (50)$$

for $j = 1, \dots, p$. Therefore, we aim to solve for $\mathbf{x}$ from

$$x_j \left(\sum_{i=1}^n \nabla l_i(\mu_i) \nabla_j \mu_i(\mathbf{x}) - h \nabla_j J(\mathbf{x}) \right) = 0 \quad (51)$$

Note that the expression inside the brackets of Equation (51) represents $\nabla_j \Psi(x)$, and x_j is included in Equation (51) to reflect the conditions in Equations (49) and (50).

The key step in developing the MI algorithm is to rearrange Equation (51) such that its positive and negative terms appear on different sides of the Equation (51). Hence we rewrite Equation (51) as

$$\begin{aligned} & x_j \left\{ - \sum_{i=1}^n (\nabla l_i(\mu_i)^+ \nabla_j \mu_i(\mathbf{x})^- + \nabla l_i(\mu_i)^- \nabla_j \mu_i(\mathbf{x})^+) + h [\nabla_j J(\mathbf{x})]^+ \right\} \\ &= x_j \left\{ \sum_{i=1}^n (\nabla l_i(\mu_i)^+ \nabla_j \mu_i(\mathbf{x})^+ + \nabla l_i(\mu_i)^- \nabla_j \mu_i(\mathbf{x})^-) - h [\nabla_j J(\mathbf{x})]^- \right\} \end{aligned} \quad (52)$$

This equation naturally suggests the following fixed point algorithm to update $\mathbf{x}$:

$$x_j^{(k+1/2)} = x_j^{(k)} \frac{\delta_{j1}^{(k)} + \epsilon}{\delta_{j2}^{(k)} + \epsilon} \quad (53)$$

where $\delta_{j1}^{(k)}$ and $\delta_{j2}^{(k)}$ denote respectively the right and left hand side of Equation (52), namely,

$$\delta_{j1}^{(k)} = \sum_{i=1}^n \{ \nabla l_i(\mu_i^{(k)})^+ \nabla_j \mu_i(\mathbf{x}^{(k)})^+ + \nabla l_i(\mu_i^{(k)})^- \nabla_j \mu_i(\mathbf{x}^{(k)})^- \} - h [\nabla_j J(\mathbf{x}^{(k)})]^- \quad (54)$$

and

$$\delta_{j2}^{(k)} = - \sum_{i=1}^n \{ \nabla l_i(\mu_i^{(k)})^+ \nabla_j \mu_i(\mathbf{x}^{(k)})^- + \nabla l_i(\mu_i^{(k)})^- \nabla_j \mu_i(\mathbf{x}^{(k)})^+ \} + h [\nabla_j J(\mathbf{x}^{(k)})]^+ \quad (55)$$

and ϵ is a small positive constant, such as $\epsilon = 10^{-5}$, used to avoid zero denominator of Equation (53). Note that the ϵ value does not affect where the algorithm converges to. As both numerator and denominator of Equation (53) are positive, $x_j^{(k+1/2)} \geq 0$ whenever $x_j^{(k)} \geq 0$.

In Equation (53) the updated x_j is denoted by $x_j^{(k+1/2)}$ indicating this is not the final estimate for iteration $k + 1$. In fact, this update does not ensure monotonic increment of $\Psi(\mathbf{x})$ and a line search step must be included to rectify this problem. We first express Equation (53) as a gradient algorithm:

$$x_j^{(k+1/2)} = x_j^{(k)} + s_j^{(k)} \nabla_j \Psi(\mathbf{x}^{(k)}) \quad (56)$$

where $s_j^{(k)} = s_j(\mathbf{x}^{(k)})$ with $s_j(\mathbf{x}) = x_j / (\delta_{j2}(\mathbf{x}) + \epsilon)$. Note that $s_j^{(k)} > 0$ when $x_j^{(k)} > 0$. When $x_j^{(k)} = 0$ we set $s_j^{(k)} = 0$ only if $\nabla_j \Psi(\mathbf{x}^{(k)}) < 0$ (since $x_j^{(k)}$ satisfies the KKT condition in this case); otherwise, we set $s_j^{(k)} = \tilde{\epsilon} / (\delta_{j2}(\mathbf{x}^{(k)}) + \epsilon)$, where $\tilde{\epsilon}$ is another small constant such as 10^{-2} . Equation (56) explains that $x_j^{(k+1/2)}$ emanates from $x_j^{(k)}$ in the gradient direction of Ψ with a non-negative step size $s_j^{(k)}$. For the line search step, the search direction is $\mathbf{d}^{(k)} = \mathbf{x}^{(k+1/2)} - \mathbf{x}^{(k)}$ with $\alpha^{(k)} > 0$ denoting the line search step size. Since $\alpha^{(k)} \leq 1$ guarantees $\mathbf{x}^{(k+1)} \geq 0$, we only search in the fixed range of $0 < \alpha^{(k)} \leq 1$. After including a line search step $\mathbf{x}^{(k+1)}$ is obtained according to

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha^{(k)} \mathbf{d}^{(k)} \quad (57)$$

Due to the fixed search interval, this line search is remarkably simple. One simple and efficient search strategy is provided by the Armijo's rule (e.g., [43]). Armijo line search is a finite terminating algorithm. Briefly, it starts with $\alpha = 1$, and for each α it checks if the following Armijo condition is satisfied:

$$\Psi(\mathbf{x}^{(k)} + \alpha \mathbf{d}^{(k)}) \leq \Psi(\mathbf{x}^{(k)}) - \xi \alpha \nabla \Psi(\mathbf{x}^{(k)})^T \mathbf{d}^{(k)} \quad (58)$$

where $0 < \xi < 1$ is a fixed parameter such as $\xi = 10^{-2}$. If Equation (58) is true then stop; otherwise, reset $\alpha = \rho \alpha$ (such as $\rho = 0.6$) and reevaluate the Armijo condition (58). Note that the repeated evaluations of $\Psi(\mathbf{x}^{(k)} + \alpha \mathbf{d}^{(k)})$ can be made with $A \mathbf{d}^{(k)}$ being computed only once. Therefore, the line search step does not add extra major computations to the MI algorithm.

Convergence properties of the MI algorithm are given in [26,41]. Briefly, under certain regular conditions, MI converges monotonically to a local maxima satisfying the KKT conditions.

For the mean functions given in Equation (4), we have $\nabla_j \mu_i(\mathbf{x}) = a_{ij}$ for emission and $\nabla_j \mu_i(\mathbf{x}) = -b_i e^{-A_i \mathbf{x}} a_{ij}$ for transmission tomography; the corresponding updating formula (53) becomes:

$$x_j^{(k+1/2)} = x_j^{(k)} \frac{\sum_{i=1}^n \nabla l_i(\mu_i^{(k)})^+ a_{ij} - h[\nabla_j J(\mathbf{x}^{(k)})]^- + \epsilon}{-\sum_{i=1}^n \nabla l_i(\mu_i^{(k)})^- a_{ij} + h[\nabla_j J(\mathbf{x}^{(k)})]^+ + \epsilon} \quad (59)$$

for emission tomography, and

$$x_j^{(k+1/2)} = x_j^{(k)} \frac{-\sum_{i=1}^n \nabla l_i(\mu_i^{(k)})^- b_i e^{-A_i \mathbf{x}^{(k)}} a_{ij} - h[\nabla_j J(\mathbf{x}^{(k)})]^- + \epsilon}{\sum_{i=1}^n \nabla l_i(\mu_i^{(k)})^+ b_i e^{-A_i \mathbf{x}^{(k)}} a_{ij} + h[\nabla_j J(\mathbf{x}^{(k)})]^+ + \epsilon} \quad (60)$$

for transmission tomography. The derivative $\nabla l_i(\mu_i)$ in the above formulae depends on the log-density $l_i(\mu_i)$. Some examples are presented below.

Example 5.1 (MI for emission scans with Poisson noise).

For emission tomography with Poisson noise, we have the log-density function for y_i :

$$l_i(\mu_i) = -\mu_i + y_i \log \mu_i \quad (61)$$

where $\mu_i = A_i \mathbf{x} + r_i$. Thus $\nabla l_i(\mu_i) = -1 + y_i/\mu_i$, which gives $\nabla l_i(\mu_i)^+ = y_i/\mu_i$ and $\nabla l_i(\mu_i)^- = -1$. The updating formula (59) becomes, for $j = 1, \dots, p$,

$$x_j^{(k+1/2)} = x_j^{(k)} \frac{\sum_{i=1}^n a_{ij} y_i / \mu_i^{(k)} - h[\nabla_j J(\mathbf{x}^{(k)})]^- + \epsilon}{\sum_{i=1}^n a_{ij} + h[\nabla_j J(\mathbf{x}^{(k)})]^+ + \epsilon} \quad (62)$$

Note that when $h = 0$ (i.e., maximum likelihood reconstruction), $r_i = 0$ and $\epsilon = 0$, this algorithm coincides with the EM algorithm for emission tomography. After line search, the estimate of $\mathbf{x}$ at iteration $k + 1$ is given by Equation (57). In this algorithm, there is only one back-projection (for the numerator of Equation (62)) and one forward-projection in each iteration; its computational burden is the same as EM.

Example 5.2 (MI for randoms-precorrected PET emission scans).

Some PET scans produce measurements that have already been corrected for randoms [44] and their measurements no longer follow Poisson distributions. We consider in this example the model weighted least squares which is also used in [11] but under a different context, i.e., we reconstruct from randoms-precorrected measurements y_i by maximizing the objective Equation (2) where

$$l_i(\mu_i) = -\frac{(y_i - \mu_i)^2}{(\mu_i + 2r_i)} \quad (63)$$

Here μ_i is used to denote $A_i \mathbf{x}$, and for this μ_i formula (59) still applies. Now since

$$\nabla l_i(\mu_i) = \left(\frac{y_i + 2r_i}{\mu_i + 2r_i} \right)^2 - 1 \quad (64)$$

we have $\nabla l_i(\mu_i)^+ = [(y_i + 2r_i)/(\mu_i + 2r_i)]^2$ and $\nabla l_i(\mu_i)^- = -1$. The MI algorithm updates $\mathbf{x}$ first according to

$$x_j^{(k+1/2)} = x_j^{(k)} \frac{\sum_{i=1}^n a_{ij} \left(\frac{y_i + 2r_i}{\mu_i^{(k)} + 2r_i} \right)^2 - h[\nabla_j J(\mathbf{x}^{(k)})]^- + \epsilon}{\sum_{i=1}^n a_{ij} + h[\nabla_j J(\mathbf{x}^{(k)})]^+ + \epsilon} \quad (65)$$

and then, after the line search step, computes $\mathbf{x}^{(k+1)}$ according to Equation (57).

Example 5.3 (MI for polyenergetic transmission scans with Poisson noise).

Application of the MI algorithm to polyenergetic X-ray CT is again extremely easy. Under the assumption of Poisson noise, the log-density for measurement y_i is identical to Equation (61) but now with $\mu_i = \sum_{m=1}^M b_{im} e^{-\sum_j a_{ij} \sum_r u_{mr} z_{rj}} + r_i$; see Equation (17). In Example 5.1 we have already derived $\nabla l_i(\mu_i)^+$ and $\nabla l_i(\mu_i)^-$ for the Poisson noise log-density. On the other hand, the derivative of μ_i with respect to z_{rj} (denoted by $\nabla_{rj} \mu_i$) is

$$\nabla_{rj} \mu_i = -\sum_m b_{im} e^{-\sum_j a_{ij} \sum_r u_{mr} z_{rj}} a_{ij} u_{mr} \quad (66)$$

Thus, the updating formula for polyenergetic transmission is

$$z_{rj}^{(k+1/2)} = z_{rj}^{(k)} \frac{\sum_{i=1}^n a_{ij} \sum_{m=1}^M u_{mr} b_{im} e^{-\sum_j a_{ij} \sum_r u_{mr} z_{rj}^{(k)}} - h[\nabla_j J(\mathbf{z}^{(k)})]^- + \epsilon}{\sum_{i=1}^n a_{ij} (y_i / \mu_i^{(k)}) \sum_{m=1}^M u_{mr} b_{im} e^{-\sum_j a_{ij} \sum_r u_{mr} z_{rj}^{(k)}} + h[\nabla_j J(\mathbf{z}^{(k)})]^+ + \epsilon} \quad (67)$$

for $r = 1, \dots, a$ and $j = 1, \dots, p$. After the line search step specified in Equation (57), $z^{(k+1)}$ is obtained. This iterative formula involves one forward- and two back-projections in each iteration, and therefore it demands similar amount of computations when compared with the alternative minimization algorithm in [34]. When $h = 0$, $r_i = 0$, $\epsilon = 0$ and $m = 1$, this MI algorithm is identical to the algorithm given in [45] for maximum likelihood reconstruction in transmission tomography. Note that unlike the optimization transfer and alternating minimization algorithms, the MI algorithm can be easily derived for other objective functions, such as the weighted least-squares function.

The above examples demonstrate that the MI algorithms are easy to derive and to implement in tomographic imaging. The line search step it requires does not incur significant computational burden.

6. Modified Fisher's Method of Scoring Using Jacobi or Gauss–Seidel Over-Relaxations

In this section we elaborate on another non-negatively constrained method for tomographic imaging, which is a modification to the standard Fisher's method of scoring (FS) algorithm. This method is developed based on the following steps. Firstly, the objective function $\Psi(\mathbf{x})$ is approximated by a quadratic function in each iteration, where the Fisher information matrix (e.g., [46]) is used to define the quadratic term; secondly, an over-relaxation method, either the Jacobi over-relaxation (JOR) or the Gauss–Seidel over-relaxation (also called the successive over-relaxation (SOR)), is employed to solve approximately the linear system derived from zeroing the derivative of this quadratic function. The resulting algorithms are called FS-JOR and FS-SOR and their detailed descriptions can be found in [47,48]. Descriptions of the JOR and SOR methods are available, for example, in [49].

FS is a general optimization algorithm for computing maximum likelihood estimates. Its advantages over the traditional Newton's method have been documented in [50]. Briefly, FS iterations are well defined due to the non-negativeness of the Fisher information matrix, but for the Newton's method, the negative Hessian matrix may not even be non-negative definite, making it unnecessarily proceed in the uphill direction in some applications. Transmission tomography is an example where this problem for the Newton's method indeed occurs; see Example 6.2.

We assume the objective function $\Psi(\mathbf{x})$ in Equation (2) is twice differentiable and let $F(\mathbf{x})$ be the Fisher information matrix, namely $F(\mathbf{x}) = E(-\nabla^2 \Psi(\mathbf{x}))$. At iteration $(k + 1)$ of the Fisher scoring algorithm, $\Psi(\mathbf{x})$ is approximated by the following quadratic function:

$$\Psi(\mathbf{x}) \approx \Psi(\mathbf{x}^{(k)}) + (\mathbf{x} - \mathbf{x}^{(k)})^T \nabla \Psi(\mathbf{x}^{(k)}) - \frac{1}{2} (\mathbf{x} - \mathbf{x}^{(k)})^T F^{(k)} (\mathbf{x} - \mathbf{x}^{(k)}) \equiv \Psi^{(k)}(\mathbf{x}) \quad (68)$$

where $F^{(k)}$ denotes the Fisher information matrix at $\mathbf{x}^{(k)}$. Then the $\mathbf{x}$ estimate is updated by constrained maximization of $\Psi^{(k)}(\mathbf{x})$, namely

$$\mathbf{x}^{(k+1)} = \arg \max_{\mathbf{x} \geq 0} \Psi^{(k)}(\mathbf{x}) \quad (69)$$

The KKT conditions for this optimization are

$$\nabla_j \Psi^{(k)}(\mathbf{x}) = 0 \text{ if } x_j > 0 \text{ and} \quad (70)$$

$$\nabla_j \Psi^{(k)}(\mathbf{x}) \leq 0 \text{ if } x_j = 0 \quad (71)$$

where

$$\nabla_j \Psi^{(k)}(\mathbf{x}) = \nabla_j \Psi(\mathbf{x}^{(k)}) - F_j^{(k)}(\mathbf{x} - \mathbf{x}^{(k)}) \quad (72)$$

Here $F_j^{(k)}$ denotes the j -th row of matrix $F^{(k)}$. The JOR and SOR methods solve, for $j = 1, \dots, p$,

$$\nabla_j \Psi^{(k)}(\mathbf{x}) = 0 \quad (73)$$

in different manners: JOR solves it by fixing all the $\mathbf{x}$ elements, except x_j , at their estimates from the last iteration (*i.e.*, iteration k), but SOR solves it by fixing all the $\mathbf{x}$ elements, except x_j , at their most current estimates.

The above illustrations describe how to incorporate JOR or SOR sub-iterations into the FS algorithm. In fact, in each iteration, JOR or SOR is used to solve approximately the linear system of equations determined by the FS algorithm, and then this approximate solution is used as the starting value for the next FS iteration. These new schemes modify the standard FS method, and are feasible for large estimation problems.

Usually it suffices to run one JOR or SOR sub-iteration. But running more than one sub-iterations is also attractive as it has the potential to reduce the computations for the entire optimization process. Suppose within each Fisher scoring iteration we run m sub-iterations of JOR or SOR. The resulting algorithms are called the m -step FS-JOR and m -step FS-SOR algorithms respectively. Let r be the sub-iteration index for the over-relaxation method and $\mathbf{x}^{(k,r)}$ the estimate of $\mathbf{x}$ at the r -th over-relaxation sub-iteration of the k -th FS iteration. Let $f_{jt}^{(k)}$ be the (j, t) -th element of $F^{(k)}$. Assume $f_{jj}^{(k)} > 0$ for all j . At iteration $k + 1$, first set $\mathbf{x}^{(k,0)} = \mathbf{x}^{(k)}$. If using JOR to solve Equation (73) we have

$$x_j^{(k,r+1)} = x_j^{(k,r)} + \omega \frac{1}{f_{jj}^{(k)}} \left(\nabla_j \Psi(\mathbf{x}^{(k)}) - \sum_{t=1}^p f_{jt}^{(k)} (x_t^{(k,r)} - x_t^{(k)}) \right) \quad (74)$$

and if using SOR to solve we then have

$$x_j^{(k,r+1)} = x_j^{(k,r)} + \omega \frac{1}{f_{jj}^{(k)}} \left(\nabla_j \Psi(\mathbf{x}^{(k)}) - \sum_{t=1}^{j-1} f_{jt}^{(k)} (x_t^{(k,r+1)} - x_t^{(k)}) - \sum_{t=j}^p f_{jt}^{(k)} (x_t^{(k,r)} - x_t^{(k)}) \right) \quad (75)$$

where $r = 0, \dots, m - 1$ and $\omega > 0$ is the relaxation parameter. If any $x_j^{(k,r+1)} < 0$ then it is reset to zero. This resetting is correct since the only possibility for $x_j^{(k,r+1)} < 0$ is that the expressions in the round brackets of Equations (74) and (75) are negative since $x_j^{(k,r)} \geq 0$ and $f_{jj}^{(k)} > 0$. Hence resetting $x_j^{(k,r+1)}$ to zeros assures that the FS-JOR and FS-SOR algorithms converge to, when they converge, the solution satisfying the KKT conditions. At the end of the sub-iterations set $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k,m)}$. Note that when $m = 1$, the last term in the round brackets of either Equation (74) or (75) becomes zero. Thus 1-step FS-JOR is basically a gradient algorithm and we can therefore replace ω by a line search step size $\omega^{(k)}$, where the search range is fixed at $0 < \omega^{(k)} \leq 1$ as this range will keep the estimate non-negative.

The relaxation parameter ω is used to achieve convergence of the FS-JOR and FS-SOR algorithms. Results contained in [47] give convergence properties when $n \rightarrow \infty$ and when the non-negativity constraint is ignored. In fact in this context FS-SOR converges if $0 < \omega < 2$ and FS-JOR converges if $0 < \omega < \xi_{\max}$, where $\xi_{\max}$ is the maximum eigenvalue of $D_F(\hat{\mathbf{x}})^{-1/2} F(\hat{\mathbf{x}}) D_F(\hat{\mathbf{x}})^{-1/2}$. Here $\hat{\mathbf{x}}$ is the MPL solution.

From the updating formulae given in Equations (74) and (75) we can see that both FS-JOR and FS-SOR involve the gradient $\nabla\Psi(\mathbf{x})$ and the Fisher information matrix based operation $F(\mathbf{x})\delta$. The gradient is standard for most reconstruction algorithms, but the computation of $F(\mathbf{x})\delta$ requires more careful consideration. It will become clear in Examples 6.1 and 6.2 that for tomographic reconstructions $F(\mathbf{x})$ usually exhibits as $A^T W(\mathbf{x})A + \nabla^2 J(\mathbf{x})$, where $W(\mathbf{x}) = \text{diag}(w_x(\mathbf{x}), \dots, w_n(\mathbf{x}))$. It is not wise to compute $A^T W(\mathbf{x})A$ first as this involves multiplications of two huge matrices A and A^T . For FS-JOR, a feasible alternative is to use the forward projection to find $A\delta$ first, then to multiply it with the diagonal values of W to get $W(\mathbf{x})A\delta$, and finally to back-project $W(\mathbf{x})A\delta$ to obtain $F(\mathbf{x})\delta$ (ignoring the penalty term). This approach involves only one forward- and one back-projections in every sub-iteration. The situation for FS-SOR is more complicated since δ changes with the pixel index j . The above approach for FS-JOR cannot be used here as otherwise each FS-SOR sub-iteration will demand infeasible p pairs of forward- and back-projections. To confront this problem, let

$$\mathbf{x}_{\prec j}^{(k,r)} = (x_1^{(k,r+1)}, \dots, x_{j-1}^{(k,r+1)}, x_j^{(k,r)}, \dots, x_p^{(k,r)})^T \quad (76)$$

The $F\delta$ part of Equation (75) involves $A(\mathbf{x}_{\prec j}^{(k,r)} - \mathbf{x}^{(k)})$. Note that

$$A_i(\mathbf{x}_{\prec j}^{(k,r)} - \mathbf{x}^{(k)}) = A_i(\mathbf{x}_{\prec j-1}^{(k,r)} - \mathbf{x}^{(k)}) + a_{ij}(x_{j-1}^{(k,r+1)} - x_{j-1}^{(k,r)}) \quad (77)$$

so we can start with $A(\mathbf{x}_{\prec 0}^{(k,r)} - \mathbf{x}^{(k)}) \equiv A(\mathbf{x}_{\prec p+1}^{(k,r-1)} - \mathbf{x}^{(k)})$ and obtain $A(\mathbf{x}_{\prec j}^{(k,r)} - \mathbf{x}^{(k)})$ by applying Equation (77). Although here the number of multiplications for $A\delta$ (where vector δ varies with its index j) becomes the same as $A\mathbf{x}$, it requires column access to the system matrix A , which can be a problem if A is generated on-the-fly.

We next provide examples of applying FS-JOR and FS-SOR to emission and transmission tomography.

Example 6.1 (Emission scans with Poisson noise).

For emission reconstruction with Poisson noise, the log-density of y_i is given by Equation (61). Thus for the corresponding object function $\Psi(\mathbf{x})$ of Equation (2), its gradients are

$$\nabla_j \Psi(\mathbf{x}) = \sum_{i=1}^n a_{ij} \left\{ -1 + \frac{y_i}{\mu_i} \right\} - h \nabla_j J(\mathbf{x}) \quad (78)$$

and its Fisher information matrix elements are

$$f_{jt} = E[-\nabla_{jt}^2 \Psi(\mathbf{x})] = \sum_{i=1}^n \frac{a_{ij} a_{it}}{\mu_i} + h \nabla_{jt}^2 J(\mathbf{x}) \quad (79)$$

where $\mu_i = A_i \mathbf{x} + r_i$, j and $t = 1, \dots, p$. Assuming we run only one sub-iteration for FS-JOR or FS-SOR (*i.e.*, $m = 1$), the FS-JOR iterative formula is

$$\tilde{x}_j^{(k+1)} = x_j^{(k)} + \omega \frac{1}{\sum_i a_{ij}^2 / \mu_i^{(k)} + h \nabla_{jj}^2 J(\mathbf{x}^{(k)})} \left(\sum_{i=1}^n a_{ij} (y_i - \mu_i^{(k)}) / \mu_i^{(k)} - h \nabla_j J(\mathbf{x}^{(k)}) \right) \quad (80)$$

and the FS-SOR formula is

$$\begin{aligned} \tilde{x}_j^{(k+1)} = & x_j^{(k)} + \omega \frac{1}{\sum_i a_{ij}^2 / \mu_i^{(k)} + h \nabla_{jj}^2 J(\mathbf{x}^{(k)})} \left(\sum_{i=1}^n a_{ij} (y_i - \mu_i^{(k)}) / \mu_i^{(k)} - h \nabla_j J(\mathbf{x}^{(k)}) \right) \\ & - \sum_{t=1}^{j-1} \left\{ \sum_{i=1}^n a_{ij} a_{it} / \mu_i^{(k)} + h \nabla_{jt}^2 J(\mathbf{x}^{(k)}) \right\} (x_t^{(k+1)} - x_t^{(k)}) \end{aligned} \quad (81)$$

Then $x_j^{(k+1)} = \max\{0, \tilde{x}_j^{(k+1)}\}$. The formula given in Equation (80) is just a gradient algorithm so ω can be replaced by a line search step size $\omega^{(k)} \in (0, 1]$. Efficient computation of Equation (81) requires column access to matrix A as explicated before. Hudson *et al.* [48] reported simulation results and a real data application for emission reconstruction. They compared FS-JOR and FS-SOR with EM. The computer time required per iteration for the EM and one-step FS-JOR algorithms were similar. By comparison with the EM algorithm, FS-JOR and FS-SOR accelerated convergence when an appropriate value of ω was used. Particularly, FS-SOR had a superior speed of convergence when $\omega = 1$.

Example 6.2 (Transmission scans with Poisson noise).

For transmission reconstructions with Poisson noise, we can easily work out the gradient and Fisher information matrix from its penalized likelihood function. The gradients are

$$\nabla_j \Psi(\mathbf{x}) = \sum_{i=1}^n a_{ij} b_i e^{-\eta_i} \left\{ 1 - \frac{y_i}{\mu_i} \right\} - h \nabla_j J(\mathbf{x}) \quad (82)$$

and the Fisher information matrix elements are

$$f_{jt} = E[-\nabla_{jt}^2 \Psi(\mathbf{x})] = \sum_{i=1}^n \frac{a_{ij} a_{it} (b_i e^{-\eta_i})^2}{\mu_i} + h \nabla_{jt}^2 J(\mathbf{x}) \quad (83)$$

where $\eta_i = A_i \mathbf{x}$, $\mu_i = b_i e^{-\eta_i} + r_i$ and j and $t = 1, \dots, p$. Note that for this example, the Fisher information matrix is non-negative but the negative Hessian matrix may not be non-negative, making the Newton method non-applicable. Corresponding to $m = 1$, the FS-JOR iterative formula is

$$\tilde{x}_j^{(k+1)} = x_j^{(k)} + \omega \frac{1}{\sum_i a_{ij}^2 (b_i e^{-\eta_i^{(k)}})^2 / \mu_i^{(k)} + h \nabla_{jj}^2 J(\mathbf{x}^{(k)})} \left(\sum_{i=1}^n a_{ij} b_i e^{-\eta_i^{(k)}} (-y_i + \mu_i^{(k)}) / \mu_i^{(k)} - h \nabla_j J(\mathbf{x}^{(k)}) \right) \quad (84)$$

and the FS-SOR formula is

$$\tilde{x}_j^{(k+1)} = x_j^{(k)} + \omega \frac{1}{\sum_i a_{ij}^2 (b_i e^{-\eta_i^{(k)}})^2 / \mu_i^{(k)} + h \nabla_{jj}^2 J(\mathbf{x}^{(k)})} \left(\sum_{i=1}^n a_{ij} b_i e^{-\eta_i^{(k)}} (-y_i + \mu_i^{(k)}) / \mu_i^{(k)} - h \nabla_j J(\mathbf{x}^{(k)}) - \sum_{t=1}^{j-1} \left\{ \sum_{i=1}^n a_{ij} a_{it} (b_i e^{-\eta_i^{(k)}})^2 / \mu_i^{(k)} + h \nabla_{jt}^2 J(\mathbf{x}^{(k)}) \right\} (x_t^{(k+1)} - x_t^{(k)}) \right) \quad (85)$$

Then $x_j^{(k+1)} = \max\{0, \tilde{x}_j^{(k+1)}\}$. Again, Equation (84) is a gradient algorithm so that a line search can be used, and efficient implementation of Equation (85) demands unpleasant column access to A .

This section explains the Fisher scoring based image reconstruction algorithms using JOR or SOR sub-iterations. For these algorithms, any negative estimates in each iteration can be corrected by simply resetting to zero, as this way of resetting enforces the KKT conditions. If only one sub-iteration is used, FS-JOR is equivalent to a gradient algorithm. For efficient implementation of FS-SOR, it requires column retrieval of the system matrix A , which can be infeasible for some reconstruction problems.

7. Iterative Coordinate Ascent Algorithms

Another method using SOR is the method of iterative coordinate ascent (ICA) (or iterative coordinate descent (ICD) for minimization problems). ICA was first implemented to tomographic imaging in [51,52]. The basic idea of ICA is to apply SOR directly to the objective function $\Psi(\mathbf{x})$, resulting in a sequence of 1-D functions where each x_j is associated with one of these 1-D functions. Then each function is solved exactly or approximately to update the corresponding x_j . More specifically, using the SOR principle we can define a function for x_j according to

$$\psi_j^{(k+1)}(x_j) = \Psi(x_1^{(k+1)}, \dots, x_{j-1}^{(k+1)}, x_j, x_{j+1}^{(k)}, \dots, x_p^{(k)}) \quad (86)$$

This is a function of x_j only and we can update the x_j estimate by

$$x_j^{(k+1)} = \arg \max_{x_j \geq 0} \psi_j^{(k+1)}(x_j) \quad (87)$$

Since this is a 1-D function, the constraint $x_j \geq 0$ can be easily enforced using, for example, the resetting to zero approach.

One computational issue with ICA when applied to tomographic imaging is that it requires repeated calculations of $\eta_i(\mathbf{x}) = \sum_j a_{ij}x_j$ for all i when updating x_j . This problem can be rectified by the following approach. Let

$$\mathbf{x}_{\setminus j}^k = (x_1^{(k+1)}, \dots, x_{j-1}^{(k+1)}, x_j^{(k)}, \dots, x_p^{(k)})^T \quad (88)$$

Consider the evaluation of $\eta_i(\mathbf{x}_{\setminus j}^k)$. Assuming the update of x_{j-1} is given by $x_{j-1}^{(k+1)} = x_{j-1}^{(k)} + \delta_{j-1}^{(k)}$, then $a_{i,j-1}x_{j-1}^{(k+1)} = a_{i,j-1}x_{j-1}^{(k)} + a_{i,j-1}\delta_{j-1}^{(k)}$, and therefore

$$\eta_i(\mathbf{x}_{\setminus j}^k) = \eta_i(\mathbf{x}_{\setminus j-1}^k) + a_{i,j-1}\delta_{j-1}^{(k)} \quad (89)$$

This relationship explains that $\eta_i(\mathbf{x}_{\setminus j}^k)$ can be cheaply computed using the η_i value before the x_j update plus a correction term. However, similar to FS-SOR, it necessitates column access to A . This can be a potential issue if A is generated on-the-fly.

Next we use again the emission and transmission examples to elaborate the ICA algorithm.

Example 7.1 (Emission scans with Poisson noise).

Firstly, we define

$$\mathbf{x}_{(j)}^{(k)} = (x_1^{(k+1)}, \dots, x_{j-1}^{(k+1)}, x_j, x_{j+1}^{(k)}, \dots, x_p^{(k)})^T \quad (90)$$

From the penalized log-likelihood function of emission measurements y_i (see, for example, Equation (34)), function $\psi_j(x_j)$ is given by

$$\psi_j(x_j) = \sum_{i=1}^n \left\{ -(\eta_i(\mathbf{x}_{(j)}^{(k)}) + r_i) + y_i \log(\eta_i(\mathbf{x}_{(j)}^{(k)}) + r_i) \right\} - hJ(\mathbf{x}_{(j)}^{(k)}) \quad (91)$$

Since this is a non-quadratic function of x_j , exact maximization is infeasible. We can find its approximate optimization by running a single or multi- step of, for example, the Newton or Fisher scoring algorithm. In this example we consider using the Fisher scoring algorithm to optimize $\psi_j(x_j)$ and call the resulting algorithm ICA-FS. After a single step of Fisher scoring we have

$$\tilde{x}_j^{(k+1)} = x_j^{(k)} + \omega_j^{(k)} \frac{1}{\sum_i a_{ij}^2 / \mu_i(\mathbf{x}_{\setminus j}^k) + h \nabla_{jj}^2 J(\mathbf{x}_{\setminus j}^k)} \left(\sum_{i=1}^n a_{ij} (y_i / \mu_i(\mathbf{x}_{\setminus j}^k) - 1) - h \nabla_j J(\mathbf{x}_{\setminus j}^k) \right) \quad (92)$$

where $\mu_i(\mathbf{x}_{\prec j}^k) = \eta_i(\mathbf{x}_{\prec j}^k) + r_i$ and $\omega_j^{(k)}$ is a line search step size enforcing $\psi_j(x_j^{(k+1)}) \geq \psi_j(x_j^{(k)})$, where equality holds only when the algorithm is converged. This monotonic condition eventually leads to $\Psi(\mathbf{x}^{(k+1)}) \geq \Psi(\mathbf{x}^{(k)})$. The update for x_j is then $x_j^{(k+1)} = \max\{0, \tilde{x}_j^{(k+1)}\}$.

Example 7.2 (Transmission scans with Poisson noise).

For this example we have

$$\psi_j(x_j) = \sum_{i=1}^n \left\{ -(b_i e^{-A_i \mathbf{x}_{\prec j}^{(k)}} + r_i) + y_i \log(b_i e^{-A_i \mathbf{x}_{\prec j}^{(k)}} + r_i) \right\} - hJ(\mathbf{x}_{(j)}^{(k)}) \quad (93)$$

where $\mathbf{x}_{(j)}^{(k)}$ is defined in Equation (90). The ICA-FS algorithm gives

$$\begin{aligned} \tilde{x}_j^{(k+1)} = x_j^{(k)} + \omega_j^{(k)} & \frac{1}{\sum_i a_{ij}^2 \left(b_i e^{-A_i \mathbf{x}_{\prec j}^k} \right)^2 / \mu_i(\mathbf{x}_{\prec j}^k) + h \nabla_{jj}^2 J(\mathbf{x}_{\prec j}^k)} \times \\ & \left(\sum_{i=1}^n a_{ij} b_i e^{-A_i \mathbf{x}_{\prec j}^k} (-y_i / \mu_i(\mathbf{x}_{\prec j}^k) + 1) - h \nabla_j J(\mathbf{x}_{\prec j}^k) \right) \end{aligned} \quad (94)$$

where $\mu_i(\mathbf{x}_{\prec j}^k) = b_i e^{-A_i \mathbf{x}_{\prec j}^k} + r_i$, and then $x_j^{(k+1)} = \max\{0, \tilde{x}_j^{(k+1)}\}$.

8. Conclusions

Image reconstruction from projections has wide applications, particularly in medical imaging. Emission and transmission tomography and X-ray CT all fall into this category. Three types of reconstruction methods are available: Fourier methods, algebraic methods and likelihood based reconstruction methods. Our attention in this paper is on the penalized likelihood approaches.

In this paper we present and discuss several important simultaneous MPL reconstruction algorithms, where the non-negativity constraint is enforced. The EM algorithm is limited to maximum likelihood reconstruction problems in emission tomography and is difficult to extend to other imaging modalities and probability models for the likelihood. One variation of EM, called the alternating minimization, is developed for transmission tomography. Another variation of EM, called the OT algorithm, is suitable for any imaging modalities and probability models, but its derivation is often cumbersome as the option for the surrogate function is flexible. The OT algorithm based on the separable parabola surrogate is relatively easy to implement to different tomographic imaging. The MI algorithm, on the other hand, is easy to derive and to implement as its line search step is cheap to compute. Its convergence speed, according to the simulation study, is similar to the separable parabola surrogate algorithm. The FS-JOR and FS-SOR algorithms first apply the Fisher information matrix to obtain a quadratic approximation to the objective function, and then optimize it using JOR or SOR schemes. Implementation of ICA-FS reverses the order of FS and SOR in FS-SOR. For both FS-SOR and ICA-FS, their convergence speeds are usually superior, but their potential problem is that both involves column retrieval of A , which may not be pre-generated and stored.

For some of the algorithms covered in this paper, their corresponding block-iterative algorithms have been developed. Block-iterative algorithms can usually achieve faster convergence than their simultaneous counterpart. However, discussions of the block-iterative algorithms are not included in this paper.

Acknowledgements

I wish to thank the referees for their invaluable comments and suggestions which have greatly enhanced the quality of this paper.

References

1. Phelps, M.E.; Hoffman, E.J.; Mullani, N.A.; Ter-Pogossian, M.M. Application of annihilation coincidence detection to transaxial reconstruction tomography. *J. Nucl. Med.* **1975**, *16*, 210–224.
2. Bailey, D.L.; Townsend, D.W.; Valk, P.E.; Maisey, M.N. *Positron Emission Tomography: Basic Sciences*; Springer-Verlag: Secaucus, NJ, USA, 2005.
3. Parra, L.; Barrett, H.H. List mode likelihood: EM algorithm and image quality estimation demonstrated on 2-D PET. *IEEE Trans. Med. Imaging* **1998**, *17*, 228–235.
4. Barrett, J.F.; Keat, N. Artifacts in CT: Recognition and avoidance. *Radio Graph.* **2004**, *24*, 1679–1691.
5. De Man, B.; Nuyts, J.; Dupont, P.; Marchal, G. Reduction of metal streak artifacts in X-ray computed tomography using a transmission maximum a posteriori algorithm. *IEEE Trans. Nucl. Sci.* **2000**, *47*, 977–981.
6. Fessler, J.A. Penalized weighted least squares image reconstruction for PET. *IEEE Trans. Med. Imaging* **1994**, *13*, 290–300.
7. Titterton, D.M. On the iterative image space reconstruction algorithm for ECT. *IEEE Trans. Med. Imaging* **1987**, *6*, 52–56.
8. Shepp, L.A.; Vardi, Y. Maximum likelihood estimation for emission tomography. *IEEE Trans. Med. Imaging* **1982**, *MI-1*, 113–121.
9. Yavuz, M.; Fessler, J.A. Statistical image reconstruction methods for randoms-precorrected PET scans. *Med. Image Anal.* **1998**, *2*, 369–378.
10. Whiting, B.R. Signal statistics in X-ray computed tomography. *Proc. SPIE 4682, Med. Imaging 2002: Phys. of Medical Imaging* **2002**, 53–60.
11. Anderson, J.M.M.; Mair, B.A.; Rao, M.; Wu, C.H. Weighted least-squares reconstruction methods for positron emission tomography. *IEEE Trans. Med. Imaging* **1997**, *16*, 159–165.
12. Veklerov, E.; Llacer, J. Stopping rule for the MLE algorithm based on statistical hypothesis testing. *IEEE Trans. Med. Imaging* **1987**, *6*, 313–319.
13. Lange, K. Convergence of EM image reconstruction algorithms with Gibbs smoothing. *IEEE Trans. Med. Imaging* **1990**, *MI-9*, 439–446.
14. Lewitt, R.M. Multidimensional digital image representations using generalized Kaiser-bessel window functions. *J. Opt. Soc. Am.* **1990**, *7*, 1834–1846.
15. Silverman, B.W.; Jones, M.C.; Wilson, J.D.; Nychka, D.W. A smoothed EM approach to indirect estimation problems, with particular reference to stereology and emission tomography (with discussion). *J. R. Stat. Soc. B* **1990**, *52*, 271–324.
16. Snyder, D.L.; Miller, M.I.; Thomas, L.J.; Polite, D.G. Noise and edge artifacts in maximum-likelihood reconstructions for emission tomography. *IEEE Trans. Med. Imaging* **1987**, *6*, 228–238.

17. Fessler, J.A. Tomographic Reconstruction Using Information Weighted Smoothing Splines. In *Information Processing in Medical Im.*; Barrett, H.H., Gmitro, A.F., Eds.; Springer-Verlag: Berlin, Germany, 1993; pp. 372–386.
18. La Rivière, P.J.; Pan, X. Nonparametric regression sinogram smoothing using a roughness-penalized Poisson likelihood objective function. *IEEE Trans. Med. Imaging* **2000**, *19*, 773–786.
19. Rudin, L.; Osher, S.; Fatemi, E. Nonlinear total variation based noise removal algorithms. *Physica D* **1992**, *60*, 259–268.
20. Huber, P.J. Robust regression: Asymptotics, conjectures, and Monte Carlo. *Ann. Stat.* **1973**, *1*, 799–821.
21. Yu, D.F.; Fessler, J.A. Edge-preserving tomographic reconstruction with nonlocal regularization. *IEEE Trans. Med. Imaging* **2002**, *21*, 159–173.
22. Evans, J.D.; Polite, D.A.; Whiting, B.R.; O’Sullivan, J.A.; Williamson, J.F. Noise-resolution tradeoffs in X-ray CT imaging: A comparison of penalized alternating minimization and filtered backprojection algorithms. *Med. Phys.* **2011**, *38*, 1444–1458.
23. Ma, J. Total Variation Smoothed Maximum Penalized Likelihood Tomographic Reconstruction with Positivity Constraints. In *Proceedings of the 8th IEEE International Symposium on Biomedical Imaging*, Chicago, USA, April 2011; pp. 1774–1777.
24. Sidky, E.Y.; Duchin, Y.; Pan, X.; Ullberg, C. A constrained, total-variation minimization algorithm for low-intensity X-ray CT. *Med. Phys.* **2011**, *38*, S117–S125.
25. Lauzier, P.T.; Tang, J.; Chen, G.H. Quantitative evaluation method of noise texture for iteratively reconstructed X-ray CT images. *Proc. Med. Imaging 2011: Phys. Med. Imaging, Proc. SPIE* **2011**, 7961, Artical 796135.
26. Ma, J. Positively constrained multiplicative iterative algorithm for maximum penalized likelihood tomographic reconstruction. *IEEE Trans. Nucl. Sci.* **2010**, *57*, 181–192.
27. Dempster, A.; Laird, N.; Rubin, D. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Stat. Soc. B* **1977**, *39*, 1–38.
28. Wei, G.; Tanner, M. A Monte Carlo implementation of the EM algorithm and the Poor Man’s data augmentation algorithm. *J. Am. Stat. Assoc.* **1990**, *85*, 699–704.
29. Lange, K.; Carson, R. EM reconstruction algorithms for emission and transmission tomography. *J. Comput. Assis. Tomogr.* **1984**, *8*, 306–316.
30. Ma, J. On iterative Bayes algorithms for emission tomography. *IEEE Trans. Nucl. Sci.* **2008**, *55*, 953–966.
31. Green, P. Bayesian reconstruction from emission tomography data using a modified EM algorithm. *IEEE Trans. Med. Imaging* **1990**, *9*, 84–93.
32. De Pierro, A.R. A modified expectation maximization algorithm for penalized likelihood estimation in emission tomography. *IEEE Trans. Med. Imaging* **1995**, *14*, 132–137.
33. Csiszár, I.; Tusnády, G. Information geometry and alternating minimization procedures. *Stat. Decis.* **1984**, Supplement Issue, No. 1, 205–237.
34. O’Sullivan, J.; Benac, J. Alternating minimization algorithms for transmission tomography. *IEEE Trans. Med. Imaging* **2007**, *26*, 283–297.

35. Csiszár, I. Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems. *Ann. Stat.* **1991**, *19*, 2032–2066.
36. O’Sullivan, J.A.; Whiting, B.R.; Snyder, D.L. Alternating Minimization Algorithms for Transmission Tomography Using Energy Detectors. In *Proceedings of the 36th Asilomar Conference Signals, Systems and Computers*, St. Louis, USA 2002; Volume 1, pp. 144–147.
37. Lasio, G.M.; Whiting, B.R.; Williamson, J.F. Statistical reconstruction for X-ray computed tomography using energy-integrating detectors. *Phys. Med. Biol.* **2007**, *52*, 2247–2266.
38. Lange, K.; Hunter, D.R.; Yang, I. Optimization transfer using surrogate objective functions. *J. Comput. Graph. Stat.* **2000**, *9*, 1–20.
39. Erdoğan, H.; Fessler, J.A. Monotonic algorithms for transmission tomography. *IEEE Trans. Med. Imaging* **1999**, *18*, 801–814.
40. Böhning, D.; Lindsay, B.G. Monotonicity of quadratic approximation algorithms. *Ann. Inst. Stat. Math.* **1988**, *40*, 641–663.
41. Chan, R.H.; Ma, J. A multiplicative iterative algorithm for box-constrained penalized likelihood image restoration. *IEEE Trans. Image Process.* **2012**, *21*, 3168–3181.
42. Gasso, G.; Rakotomamonjy, A.; Canu, S. Recovering sparse signals with a certain family of non-convex penalties and DC programming. *IEEE Trans. Signal Proc.* **2009**, *57*, 4686–4698.
43. Luenberger, D. *Linear and Nonlinear Programming*, 2nd ed.; J. Wiley: New York, NY, USA, 1984.
44. Ahn, S.; Fessler, J.A. Emission image reconstruction for randoms-precorrected PET allowing negative sinogram values. *IEEE Trans. Med. Imaging* **2004**, *23*, 591–601.
45. Lange, K.; Bahn, M.; Little, R. A theoretical study of some maximum likelihood algorithms for emission and transmission tomography. *IEEE Trans. Med. Imaging* **1987**, *6*, 106–114.
46. Ober, R.J.; Zou, Q.; Zhiping, L. Calculation of the Fisher information matrix for multidimensional data sets. *IEEE Trans. Signal Proc.* **2003**, *51*, 2679–2691.
47. Ma, J.; Hudson, H.M. Modified Fisher scoring algorithms using Jacobi or Gauss-Seidel subiterations. *Comput. Stat.* **1997**, *12*, 467–479.
48. Hudson, H.; Ma, J.; Green, P. Fisher’s method of scoring in statistical image reconstruction: Comparison of Jacobi and Gauss-Seidel iterative schemes. *Stat. Method Med. Res.* **1994**, *3*, 41–61.
49. Ortega, J.M.; Rheinboldt, W.C. *Iterative Solutions of Nonlinear Equations in Several Variables*; Academic Press: New York, NY, USA, 1970.
50. Osborne, M.R. Fisher’s method of scoring. *Int. Stat. Rev.* **1992**, *60*, 99–117.
51. Sauer, K.; Bouman, C. A local update strategy for iterative reconstruction from projections. *IEEE Trans. Signal Proc.* **1993**, *41*, 533–548.
52. Bouman, C.A.; Sauer, K. A unified approach to statistical tomography using coordinate descent optimization. *IEEE Trans. Image Process.* **1996**, *5*, 480–492.