Inverse Problems in Science and Engineering
Vol. 00, No. 00, May 2010, 1–12

RESEARCH ARTICLE

Regularization parameter selection for penalized-maximum likelihood methods in PET

Johnathan M. Bardsley† and John Goldes†
(*Received 00 Month 200x; in final form 00 Month 200x*)

Penalized maximum likelihood methods are commonly used in positron emission tomography (PET). Due to the fact that a Poisson data-noise model is typically assumed, standard regularization parameter choice methods, such as the discrepancy principle or generalized cross validation, can not be directly applied. In recent work of the authors, regularization parameter choice methods for penalized negative-log Poisson likelihood problems are introduced, and the application is image deconvolution. In this paper, we extend those methods to the application of PET, introducing a minor modification that seems to improve the performance of the methods. Moreover, we show how these methods can be used to choose the hyper-parameters in a Bayesian hierarchical regularization approach, also of the authors' previous work.

Keywords: positron emission tomography, inverse problems, regularization parameter selection, Bayesian statistical methods, Poisson noise

AMS Subject Classification: 65J22, 65K10, 65F22.

1. Introduction

A positron emission tomography (PET) scan provides information about the density of a metabolite (such as glucose) marked with a radioactive isotope in a living organism. When the radioactive isotope decays, it emits a positron, which in turn annihilates with an electron, causing a pair of photons to propagate in opposite directions. If these two photons reach two different detectors on the PET machine within a sufficiently short time window they are counted as an event along the corresponding connecting line L, referred to as the Line Of Response (LOR). The collection of all events along all LORs during an experiment constitutes a PET data set. The PET reconstruction problem, and the inverse problem of interest, is to reconstruct, from this data, the density of the radioactive metabolite within the organism being imaged.

In [17, 22], a mathematical model for PET data formation is presented. If there are M LORs and N elements in the uniform $\sqrt{N} \times \sqrt{N}$ tracer density grid, we can write the mathematical/statistical model as follows:

$$\mathbf{b} = \text{Poisson}(\mathbf{A}\mathbf{x} + \boldsymbol{\gamma}), \tag{1}$$

where $\mathbf{b} \in \mathbb{R}^M$ is the vector containing the expected number of events along each of the M LORs; Poisson($\boldsymbol{\lambda}$) denotes a Poisson random vector with mean $\boldsymbol{\lambda}$; $\mathbf{A}$ is the $M \times N$ forward model matrix; $\mathbf{x} \in \mathbb{R}^N$ is the discrete representation of the (unknown) photon emission density function; and $\boldsymbol{\gamma}$ is the vector containing

† Department of Mathematical Sciences, University of Montana, USA. Email: bardsleyj@mso.umt.edu, john.goldes@umontana.edu. This work was supported by the NSF under grant DMS-0915107

ISSN: 1741-5977 print/ISSN 1741-5985 online

DOI: 10.1080/1741597YYxxxxxxxx
http://www.informaworld.com

expected erroneous counts due to accidental coincidences and scattered events [21], which we assume is known.

The attenuation matrix $\mathbf{A}$ in (1) is of the form

$$\mathbf{A} = \mathbf{G}\mathbf{A}^{\text{Radon}}, \tag{2}$$

where $\mathbf{A}^{\text{Radon}}$ is the discrete $M \times N$ Radon transform matrix [16], and $\mathbf{G} = \text{diag}(g_1, g_2, \ldots, g_M)$ is the diagonal matrix with diagonal values

$$g_j = \exp\left(-\int_{L_j} \mu(s)\, ds\right). \tag{3}$$

Here L_j is the jth LOR, μ is the absorption density function of the subject (estimated previous to the PET scan, and assumed known). Note that g_j can be viewed as the probability that an emission event anywhere along line L_j is recorded by the detector [21]. Note also that the ijth element of $\mathbf{A}^{\text{Radon}}$ is the intersection length of the ith LOR with the jth computational grid element [16], and so is sparse.

The matrix $\mathbf{A}$ accounts for attenuation due both to Compton scattering and photo-electric absorption. Our mathematical/statistical model does not take into account detector efficiency or detector deadtime [21], however this has no bearing on our discussion. For a complete model of PET data, see [21].

In order to use model (2), (3), the $\sqrt{N} \times \sqrt{N}$ tracer density vector $\boldsymbol{\mu}$ must be estimated. In practice, this is done by solving the so-called transmission PET problem [17], or via another imaging technique such as computed tomography. We will assume throughout the remainder of the document that we have an estimate of $\boldsymbol{\mu}$ on the tracer density grid, and hence of $\mathbf{A}$, in hand.

Assuming (1), the probability mass function of the data $\mathbf{b}$ conditioned on x is given by

$$p(\mathbf{b} \mid \mathbf{x}) = \prod_{j=1}^{M} \frac{([\mathbf{A}\mathbf{x}]_j + \gamma_j)^{b_j}\, e^{-([\mathbf{A}\mathbf{x}]_j + \gamma_j)}}{b_j!}. \tag{4}$$

The maximum likelihood estimate of the true tracer density $\mathbf{x}_e$ given $\mathbf{b}$ is obtained by maximizing $p(\mathbf{b} \mid \mathbf{x})$ with respect to $\mathbf{x}$, subject to the constraint $\mathbf{x} \geq \mathbf{0}$. Equivalently, we can solve

$$\mathbf{x}_{\text{ML}} = \arg\min_{\mathbf{x} \geq \mathbf{0}} T_0(\mathbf{x}\,; \mathbf{b}), \tag{5}$$

where

$$T_0(\mathbf{x}\,; \mathbf{b}) = \sum_{i=1}^{n} \left\{([\mathbf{A}\mathbf{x}]_i + \gamma_i) - b_i \ln([\mathbf{A}\mathbf{x}]_i + \gamma_i)\right\}. \tag{6}$$

Note that $T_0(\mathbf{x}\,; \mathbf{b})$ is equal, up to an additive constant, to $-\ln p(\mathbf{b} \mid \mathbf{x})$. $\mathbf{x}_{\text{ML}}$ is known as the maximum likelihood estimate of $\mathbf{x}_e$.

However, solving (5) directly yields tracer densities with unrealistic artifacts. Because of this, a regularization term is often added (see, e.g., [1, 10, 11, 13–15, 18, 20, 27]), which can incorporate prior knowledge about the true tracer density.

This results in the following modification of (5): compute

$$\mathbf{x}_\alpha = \arg\min_{\mathbf{x}\geq\mathbf{0}} \left\{ T_\alpha(\mathbf{x}) \stackrel{\text{def}}{=} T_0(\mathbf{x};\mathbf{b}) + \frac{\alpha}{2}\mathbf{x}^T\mathbf{C}\mathbf{x} \right\}. \tag{7}$$

Here α is known as the regularization parameter, and $\mathbf{C}$ as the regularization matrix. To guarantee that T_α has a unique nonnegative minimizer, we make the assumption that the null-spaces of $\mathbf{A}$ and $\mathbf{C}$ intersect only trivially [5]. In the PET literature, (7) is known as a penalized maximum likelihood (PML) problem, and such problems have been studied extensively; see, e.g., [1, 10, 11, 13–15, 18, 20, 27].

In the Bayesian setting, $\mathbf{x}_\alpha$ is the *maximum a posteriori* (MAP) estimator:

$$\mathbf{x}_\alpha = \arg\max_{\mathbf{x}\geq\mathbf{0}} p(\mathbf{b} \mid \mathbf{x}) p_{\text{prior}}(\mathbf{x}), \tag{8}$$

where $p(\mathbf{b} \mid \mathbf{x})$ is defined in (4) and

$$p_{\text{prior}}(\mathbf{x}) = \sqrt{\frac{\alpha|\mathbf{C}|}{(2\pi)^n}} \exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{C}\mathbf{x} \right). \tag{9}$$

The Bayesian formulation of the problem will be important later in the manuscript.

The focus of this paper is to present methods for choosing an appropriate value of α in (7). In the next section, we present three methods, which are extensions of the well-known discrepancy principle, generalized cross validation, and unbiased predictive risk methods for regularization parameter selection in penalized least squares estimation [25]. We then test the methods on several examples in Section 3, and end with conclusions in Section 4

2. Regularization Parameter Choice Methods

Existing methods for estimating the regularization parameter in the least squares case can not be directly applied to (7). However, in [5], a quadratic approximation of T_0, as defined in (6), is computed and used to extend existing regularization parameter selection methods. The quadratic approximation is given by

$$\begin{aligned} T_0(\mathbf{x};\mathbf{b}) = T_0(\mathbf{x}_e;\mathbf{b}) &+ \frac{1}{2}(\mathbf{A}\mathbf{x} - (\mathbf{b}-\boldsymbol{\gamma}))^T \text{diag}\left(\frac{\mathbf{1}}{\mathbf{b}_e}\right)(\mathbf{A}\mathbf{x} - (\mathbf{b}-\boldsymbol{\gamma})) \\ &+ \mathcal{O}(\|\mathbf{h}\|_2^3, \|\mathbf{h}\|_2^2\|\mathbf{k}\|_2, \|\mathbf{h}\|_2\|\mathbf{k}\|_2^2, \|\mathbf{k}\|_2^2), \end{aligned} \tag{10}$$

where $\mathbf{h} = \mathbf{x} - \mathbf{x}_e$, $\mathbf{k} = \mathbf{b} - \mathbf{b}_e$, and $\mathbf{b}_e = \mathbf{A}\mathbf{x}_e + \boldsymbol{\gamma}$. Now we use an application of the mean value theorem to replace $\mathbf{b}_e$ in (10) with $\mathbf{b}_\alpha = \mathbf{A}\mathbf{x}_\alpha + \boldsymbol{\gamma}$, where $\mathbf{x}_\alpha$ is computed from (7). Note that in [5] we proceeded similarly, but replaced $\mathbf{b}_e$ by $\mathbf{b}$ instead. We present this modification here because we have found it to be slightly more effective.

The mean value theorem is used by considering the ith component of $\frac{\mathbf{1}}{\mathbf{b}_\alpha - \mathbf{k}_\alpha}$ as a function of k_{α_i}, where $\mathbf{k}_\alpha = \mathbf{b}_\alpha - \mathbf{b}_e$. Letting $\mathbf{r} = \mathbf{A}\mathbf{x} - (\mathbf{b} - \boldsymbol{\gamma})$, the weighted

sum of squares term in (10) can be written as

$$\frac{1}{2}\mathbf{r}^T\left[\mathbf{r}\odot\left(\frac{\mathbf{1}}{\mathbf{b}_\alpha-\mathbf{k}_\alpha}\right)\right]=\frac{1}{2}\mathbf{r}^T\left[\mathbf{r}\odot\left(\frac{\mathbf{1}}{\mathbf{b}_\alpha}+\mathrm{diag}\left(\frac{\mathbf{1}}{(\mathbf{b}_\alpha-\hat{\mathbf{k}}_\alpha)^2}\right)\mathbf{k}_\alpha\right)\right]$$
$$=\frac{1}{2}\mathbf{r}^T\left(\frac{\mathbf{r}}{\mathbf{b}_\alpha}\right)+\frac{1}{2}\mathbf{r}^T\left(\frac{\mathbf{r}\odot\mathbf{k}_\alpha}{(\mathbf{b}_\alpha-\hat{\mathbf{k}}_\alpha)^2}\right), \tag{11}$$

where $\odot$ indicates component-wise multiplication, the square and quotient are taken component-wise, and $0<|\hat{k}_{\alpha_i}|<|k_{\alpha_i}|$. Noting that $\mathbf{r}=\mathbf{A}\mathbf{h}-\mathbf{k}$ and that $\mathbf{b}_\alpha$ is bounded away from zero, we obtain

$$\mathbf{r}^T\left(\frac{\mathbf{r}\odot\mathbf{k}_\alpha}{(\mathbf{b}_\alpha-\hat{\mathbf{k}}_\alpha)^2}\right)=\mathcal{O}(\|\mathbf{h}\|^2\|\mathbf{k}_\alpha\|,\|\mathbf{h}\|\|\mathbf{k}\|\|\mathbf{k}_\alpha\|,\|\mathbf{k}\|^2\|\mathbf{k}_\alpha\|). \tag{12}$$

Recalling (10), we have the approximation

$$\begin{aligned}T_0(\mathbf{x};\mathbf{b})=&T_0(\mathbf{x}_e;\mathbf{b})+T_0^{\mathrm{wls}}(\mathbf{x};\mathbf{b})\\&+\mathcal{O}(\|\mathbf{h}\|_2^3,\|\mathbf{h}\|_2^2\|\mathbf{k}\|_2,\|\mathbf{h}\|_2\|\mathbf{k}\|_2^2,\|\mathbf{k}\|_2^2)\\&+\mathcal{O}(\|\mathbf{h}\|^2\|\mathbf{k}_\alpha\|,\|\mathbf{h}\|\|\mathbf{k}\|\|\mathbf{k}_\alpha\|,\|\mathbf{k}\|^2\|\mathbf{k}_\alpha\|),\end{aligned} \tag{13}$$

where

$$T_0^{\mathrm{wls}}(\mathbf{x};\mathbf{b})=\frac{1}{2}\|\mathbf{B}_\alpha^{-1/2}(\mathbf{A}\mathbf{x}-(\mathbf{b}-\boldsymbol{\gamma}))\|^2, \tag{14}$$

with $\mathbf{B}_\alpha=\mathrm{diag}(\mathbf{b}_\alpha)$.

This approximation is modified from that of [5] in that the weighting term in (14) is now $\mathbf{B}_\alpha^{-1}$ instead of $\mathbf{B}^{-1}$.

2.0.1. Discrepancy Principle

The idea behind the discrepancy principle, as presented in [5], is to choose α so that $\mathbf{x}_\alpha$ computed from (7) satisfies

$$T_0(\mathbf{x}_\alpha;\mathbf{b})\approx E(T_0(\mathbf{x}_e;\mathbf{b})), \tag{15}$$

where E is the expected value function.

For the right-hand side of (15), from (13), we have

$$E(T_0(\mathbf{x}_e,\mathbf{b}))\approx T_0(\mathbf{x}_e;\mathbf{b}_e)+E\left(T_0^{\mathrm{wls}}(\mathbf{x}_e;\mathbf{b})\right). \tag{16}$$

It can be argued (see [5]) that $\|\mathbf{B}_\alpha^{-1/2}(\mathbf{A}\mathbf{x}_e-(\mathbf{b}-\boldsymbol{\gamma}))\|^2$ is approximately $\chi^2(M)$ distributed, and hence that

$$E\left(T_0^{\mathrm{wls}}(\mathbf{x}_e;\mathbf{b})\right)\approx M/2. \tag{17}$$

For the left-hand side of (15), from (13), we have

$$T_0(\mathbf{x}_\alpha;\mathbf{b}) \approx T_0(\mathbf{x}_e;\mathbf{b}) + T_0^{\text{wls}}(\mathbf{x}_\alpha;\mathbf{b}). \tag{18}$$

Assuming $T_0(\mathbf{x}_e;\mathbf{b}) \approx T_0(\mathbf{x}_e;\mathbf{b}_e)$, (16), (17), and (18) then imply that (15) will hold provided

$$T_0^{\text{wls}}(\mathbf{x}_\alpha;\mathbf{b}) \approx M/2. \tag{19}$$

Thus finally, based on (19), we can define the discrepancy principle choice of the regularization parameter:

$$\alpha_{\text{DP}} = \arg\min_{\alpha\geq 0} \left(T_0^{\text{wls}}(\mathbf{x}_\alpha;\mathbf{b}) - M/2\right)^2, \tag{20}$$

where $\mathbf{x}_\alpha$ is computed from (7).

Recent work of Mead and Renaut [19] successfully implement a different χ^2 criteria in the least squares case. This approach could very likely be extended to the Poisson setting, but we do not pursue that here.

2.0.2. Generalized Cross Validation

In [5], a quadratic approximation of T_0 is also used to extend the method of generalized cross validation (GCV) for regularization parameter choice in the least squares case [23, 25] to the case in which T_0 is the likelihood function. GCV can be viewed as an approximation of leave-one-out cross validation [23] for large-scale problems.

The method is as follows: choose α to be the minimizer of

$$\text{GCV}(\alpha) \stackrel{\text{def}}{=} MT_0^{\text{wls}}(\mathbf{x}_\alpha;\mathbf{b}) \Big/ \text{trace}(\mathbf{I}_n - \mathbf{B}_\alpha^{-1/2}\mathbf{A}\mathbf{A}_\alpha)^2 \tag{21}$$

subject to the constraint $\alpha \geq 0$. Here $\mathbf{x}_\alpha$ is computed from (7), and $\mathbf{A}_\alpha$ is a matrix satisfying $\mathbf{x}_\alpha = \mathbf{A}_\alpha \mathbf{b}_\alpha^{-1/2}(\mathbf{b}-\boldsymbol{\gamma})$.

However, for (7), the data-to-regularized solution (or regularization) operator is nonlinear, and so $\mathbf{A}_\alpha$ must be a linear approximation satisfying $\mathbf{x}_\alpha \approx \mathbf{A}_\alpha \mathbf{B}_\alpha^{-1/2}(\mathbf{b}-\boldsymbol{\gamma})$. In [5], the approximation

$$\mathbf{A}_\alpha = (\mathbf{D}_\alpha(\mathbf{A}^T\mathbf{B}_\alpha^{-1}\mathbf{A} + \alpha\mathbf{C})\mathbf{D}_\alpha)^\dagger \mathbf{D}_\alpha \mathbf{A}^T \mathbf{b}_\alpha^{-1/2} \tag{22}$$

is used, where "†" denotes psuedo-inverse, and $\mathbf{D}_\alpha$ is a diagonal matrix with diagonal entries $[\mathbf{D}_\alpha]_{ii} = 1$ if $[\mathbf{x}_\alpha]_i > 0$ and $[\mathbf{D}_\alpha]_{ii} = 0$ otherwise (see [5] for details).

Even with (22) in hand, however, the computation of the trace in (21) remains impractical, and so randomized trace estimation is used [12, 25]. Here the following fact is exploited: if $\mathbf{v}$ is a discrete white noise vector,

$$\text{trace}(\mathbf{I}_n - \mathbf{B}_\alpha^{-1/2}\mathbf{A}\mathbf{A}_\alpha) \approx \mathbf{v}^T\mathbf{v} - \mathbf{v}^T\mathbf{B}_\alpha^{-1/2}\mathbf{A}\mathbf{A}_\alpha\mathbf{v}. \tag{23}$$

In fact, equality holds in (23) if the right-hand side is replaced by its expected value. As is mentioned in [25], the optimal choice of $\mathbf{v}$ in (23) is to take v_i to be $+1$ or -1 with probability $1/2$.

Finally, because the psuedo-inverse in (22) is impractical to compute, $\mathbf{A}_\alpha\mathbf{v}$ in (23) is approximated using a truncated conjugate gradient iteration applied to the

linear system

$$\mathbf{D}_\alpha(\mathbf{A}^T\mathbf{b}_\alpha^{-1}\mathbf{A} + \alpha\mathbf{C})\mathbf{D}_\alpha\mathbf{x} = \mathbf{D}_\alpha\mathbf{A}^T\mathbf{b}_\alpha^{-1/2}\mathbf{v} \tag{24}$$

with a stopping rule based on the norm of the residual and a choice of maximum number of iterations.

Taking all of the above approximations of GCV(α) into account, and calling the resulting approximate GCV function $\widetilde{\text{GCV}}(\alpha)$, we define the GCV choice of the regularization parameter by

$$\alpha_{\text{GCV}} = \arg\min_{\alpha\geq 0} \widetilde{\text{GCV}}(\alpha). \tag{25}$$

2.0.3. Unbiased Predictive Risk Estimation

In [5], a quadratic approximation of T_0 is also used to extend the method of unbiased predicative risk estimation (UPRE) for regularization parameter choice in the least squares case [25] to the case in which T_0 is the likelihood function.

The motivation behind UPRE is as follows: we seek the value of α that minimizes the *predictive risk* $E(T_0(\mathbf{x}_\alpha; \mathbf{b}_e))$. However since $\mathbf{b}_e$ is unknown, we minimize instead an unbiased estimator of the predictive risk. Following the arguments in [5], such an estimator can be well-approximated by choosing $\mathbf{x}_\alpha$ with α approximately minimizing

$$\text{UPRE}(\alpha) = T_0^{\text{wls}}(\mathbf{x}_\alpha; \mathbf{b}) + \text{trace}(\mathbf{B}_\alpha^{-1/2}\mathbf{A}\mathbf{A}_\alpha) - M/2, \tag{26}$$

where $\mathbf{x}_\alpha$ is computed from (7). The trace is estimated in the same fashion as for GCV, with the exception that in place of (23), we use

$$\text{trace}(\mathbf{B}_\alpha^{-1/2}\mathbf{A}\mathbf{A}_\alpha) \approx \mathbf{v}^T\mathbf{B}_\alpha^{-1/2}\mathbf{A}\mathbf{A}_\alpha\mathbf{v}. \tag{27}$$

This results in an approximate UPRE function $\widetilde{\text{UPRE}}(\alpha)$, and the UPRE choice of the regularization parameter is given by

$$\alpha_{\text{UPRE}} = \arg\min_{\alpha\geq 0} \widetilde{\text{UPRE}}(\alpha). \tag{28}$$

3. Numerical Experiments

We test our approach on synthetically generated data. The true emission density $\mathbf{x}_e$ is shown in Figure 1. The noisy sinogram data $\mathbf{b}$, generated using statistical model (1) and MATLAB's `poissrnd` function, is shown on the right in Figure 1. We assumed that $\boldsymbol{\gamma}$ is a constant vector of 1s at all pixels, and that the density vector $\boldsymbol{\mu}$ was a vector of zeros – typical in the PET literature – so that $\mathbf{A}$ is the discrete Radon transform matrix. Our computational grid is defined by 128 detectors and angles, as well as a 128×128 uniform computational grid for the unknown emission density. Thus $M = N = 128^2$.

In order to test our approach on multiple data sets we vary the signal-to-noise ratio. The signal-to-noise ratio for data with statistical model (1) is defined as

$$SNR = \sqrt{\frac{\|\mathbf{A}\mathbf{x} + e + \boldsymbol{\gamma}\|^2}{E(\|\mathbf{b} - (\mathbf{A}\mathbf{x}_e + \boldsymbol{\gamma})\|^2)}}, \tag{29}$$

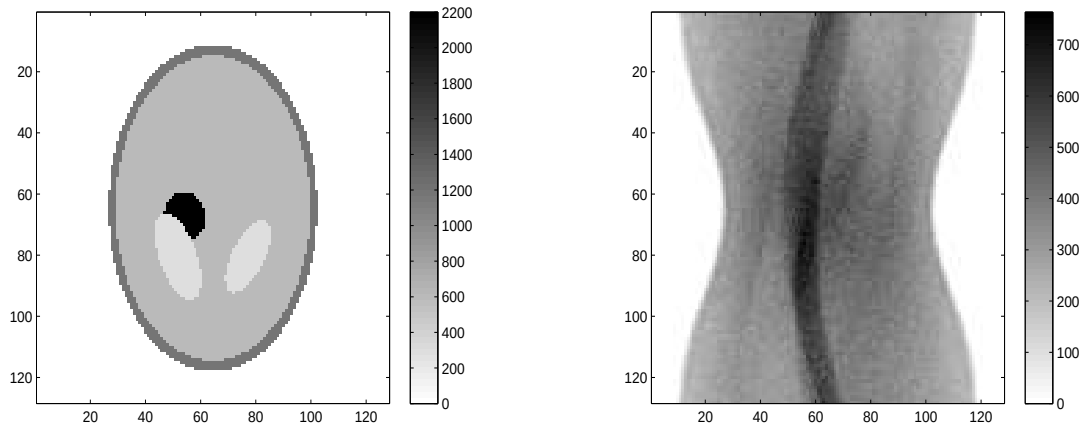

Figure 1. $\mathbf{x}_e$ *is plotted on the left and* $\mathbf{b}$ *is plotted on the right. The signal-to-noise ratio of* $\mathbf{b}$ *is 20.*

with

$$E(\|\mathbf{b} - (\mathbf{A}\mathbf{x}_e + \boldsymbol{\gamma})\|^2) = \sum_{i=1}^{M}([\mathbf{A}\mathbf{x}_e]_i + \gamma_i). \tag{30}$$

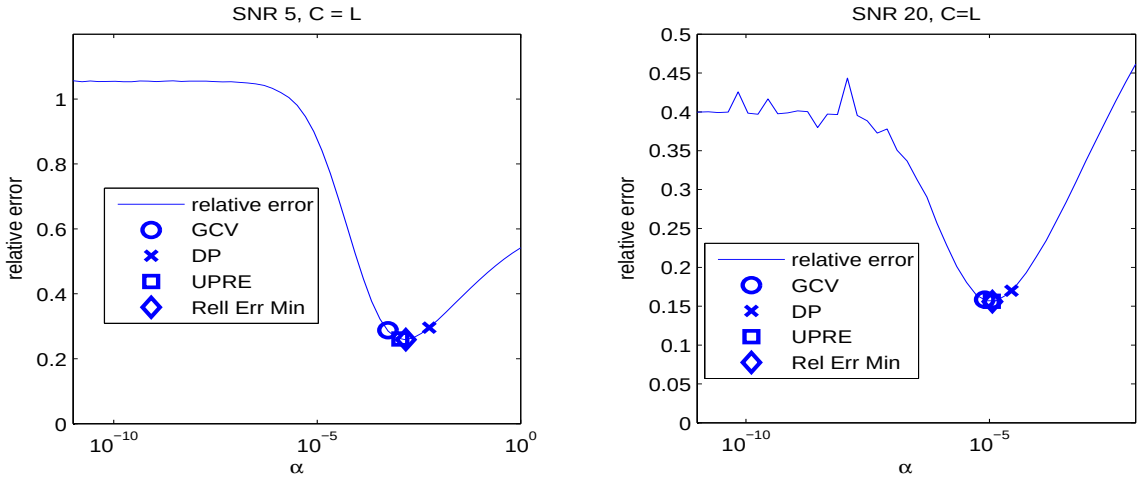

Figure 2. *Plots of* α *versus relative error are shown. The plot on the left is from data with a SNR of 5 and the plot on the right is from data with a SNR of 20.*

MATLAB's `fminbnd` function was used for computing approximate solutions to (20), (25), and (28). A more efficient method, exploiting the unique structure of our problem, is likely possible, however we do not pursue that here.

To test the effectiveness of the regularization parameter selection methods, we plot the relative error

$$\frac{\|\mathbf{x}_\alpha - \mathbf{x}_e\|}{\|\mathbf{x}_e\|} \tag{31}$$

for a range of α values, together with the values of α chosen by the three methods. This can be seen in Figure 2. The regularized solution $\mathbf{x}_\alpha$ is calculated from (7) with $\mathbf{C} = \mathbf{L}$, where $\mathbf{L}$ is a discretization of the negative Laplacian (smoothing)

operator. Note that when α is close to zero the relative error is large due to the presence of unrealistic artifacts, whereas when α is too large the relative error is large because the penalty term dominates the reconstruction.

In both cases, the GCV and UPRE methods yielded similar recommendations for α, while the DP (discrepancy principle) yielded a recommendation that was slightly worse in terms of relative error. Figure 3 contains the reconstructions that were obtained from the two data sets using the DP and UPRE recommendations. The GCV recommendation yielded a reconstruction that was visibly very similar to that obtained from the UPRE recommendation.

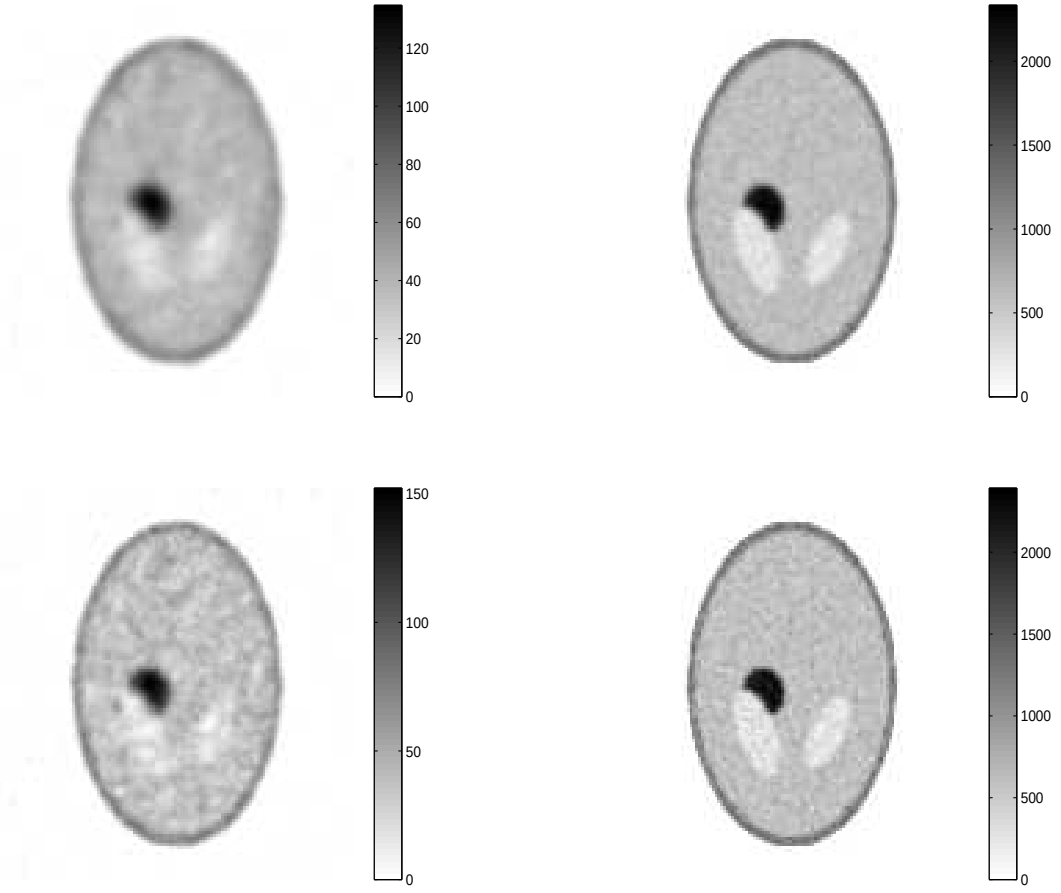

Figure 3. *Plots of the reconstructions obtained from the two data sets, with the reconstructions obtained from data with SNR=5 on the left and SNR=20 on the right. The top row contains the reconstructions that were computed with the DP recommendation and the bottom row contains reconstructions that were computed with the UPRE recommendation.*

3.1. *Anisotropic Diffusion via Hierarchical Regularization*

Most PET images are piece-wise smooth with sharp jumps in intensity, corresponding, e.g., to tissue boundaries. In pixels corresponding to sharp intensity jumps, a smoothing regularization, or prior, is not desirable. Thus we would like a means of adapting the regularization given information about the location of edges in the image.

A very effective approach along these line, incorporating Bayesian hierarchical models, was introduced in [7] for regularized least squares problems. This method-

ology was extended to the PET case in [3]. Here we show how to use the regularization parameter choice methods presented above to automate the choice of parameters in the approach of [3].

First, similar to (8)-(9), using Bayes' Law, we define a maximum a posteriori (MAP) estimation problem of the form

$$\arg\max_{\mathbf{x}\geq\mathbf{0},\boldsymbol{\theta}} p(\mathbf{x},\boldsymbol{\theta}\mid\mathbf{b}), \tag{32}$$

where

$$p(\mathbf{x},\boldsymbol{\theta}\mid\mathbf{b}) = p(\mathbf{b}\mid\mathbf{x})p_{\text{prior}}(\mathbf{x}\mid\boldsymbol{\theta})p_{\text{hyper}}(\boldsymbol{\theta}) \tag{33}$$

with p given by (4).

In our definition of the prior, we allow for a variable penalty on the horizontal and vertical partial derivatives, which we denote, in the discrete setting, by $\mathbf{D}_1$ and $\mathbf{D}_2$. Our assumption regarding the distribution of the image $\mathbf{x}$ is as follows:

$$\mathbf{D}_1\mathbf{x}, \mathbf{D}_2\mathbf{x} \sim \text{Normal}(\mathbf{0}, \mathbf{D}_\theta), \quad \mathbf{D}_\theta = \text{diag}(\theta_1,\ldots,\theta_N). \tag{34}$$

Assuming that these two random vectors are independent, the prior p_{prior} is then defined

$$\begin{aligned} p_{\text{prior}}(\mathbf{x}\mid\boldsymbol{\theta}) &= p(\mathbf{D}_1\mathbf{x}\mid\boldsymbol{\theta})p(\mathbf{D}_1\mathbf{x}\mid\boldsymbol{\theta}) \\ &\propto \det(\mathbf{D}_\theta^{-1})\exp\left(-\frac{1}{2}\mathbf{x}^T(\mathbf{D}_1^T\mathbf{D}_\theta^{-1}\mathbf{D}_1 + \mathbf{D}_2^T\mathbf{D}_\theta^{-1}\mathbf{D}_2)\mathbf{x}\right) \\ &= \exp\left(-\frac{1}{2}\mathbf{x}^T(\mathbf{D}_1^T\mathbf{D}_\theta^{-1}\mathbf{D}_1 + \mathbf{D}_2^T\mathbf{D}_\theta^{-1}\mathbf{D}_2)\mathbf{x} - \sum_{i=1}^{N}\log\theta_i\right). \end{aligned} \tag{35}$$

See [3] for details. Note that if $\theta = \alpha^{-1}\mathbf{1}$ is deterministic, this is equivalent to using $\mathbf{C} = \mathbf{D}_1^T\mathbf{D}_1 + \mathbf{D}_2^T\mathbf{D}_2$ in (7).

It remains to define $\pi_{\text{hyper}}(\boldsymbol{\theta})$. As in [3], we assume

$$\theta_i \sim \text{Gamma}(\alpha_0,\theta_0), \quad i = 1,\ldots,n,$$

so that

$$p_{\text{hyper}}(\boldsymbol{\theta}) \propto \Pi_{i=1}^{N}\theta_i^{\alpha_0-1}\exp\left(\frac{\theta_i}{\theta_0}\right). \tag{36}$$

The MAP estimate is computed by minimizing $-\ln p(\mathbf{x},\boldsymbol{\theta}\mid\mathbf{b})$, defined in (33), with respect to $\mathbf{x}$ and $\boldsymbol{\theta}$. A simple cyclic iteration has been found to be effective for this problem [3, 7]. The outline of the algorithm is as follows:

Step 0: Initialize $\boldsymbol{\theta}_0 = \alpha_0\theta_0\mathbf{1}$, $k = 1$.
Step 1: Update the estimate $\mathbf{x}^k$:
$\mathbf{x}_k = \text{argmin}_{\mathbf{x}\geq 0} - \ln p(\mathbf{x},\boldsymbol{\theta}_{k-1}\mid\mathbf{b})$.
Step 2: Update the estimate of $\boldsymbol{\theta}$:
$\boldsymbol{\theta}_k = \text{argmin}_{\boldsymbol{\theta}\geq 0} - \ln p(\mathbf{x}_k,\boldsymbol{\theta}\mid\mathbf{b})$.
Step 3: Increase k by 1 and return to **Step 1**. Repeat until convergence.

Note that optimization problem in **Step 1** has the form of (7), with

$$\mathbf{C} = \mathbf{D}_1^T \mathbf{D}_{\theta_{k-1}}^{-1} \mathbf{D}_1 + \mathbf{D}_2^T \mathbf{D}_{\theta_{k-1}}^{-1} \mathbf{D}_2,$$

and the update in **Step 2** has the analytic form

$$\boldsymbol{\theta}_{k,j} = \theta_0 \left(\frac{\alpha_0 - 2}{2} + \sqrt{\frac{[\mathbf{D}_1 \mathbf{x}_k]_j^2 + [\mathbf{D}_2 \mathbf{x}_k]_j^2}{2\theta_0} + \frac{(\alpha_0 - 2)^2}{4}} \right). \tag{37}$$

Left unspecified is the choice of the parameters α_0 and θ_0 in the gamma hyperprior. For $0 < \alpha_0 - 2 << 1$, using (37) yields a regularization function closely related to total variation; note that when $\alpha_0 = 2$, total variation results. Since in our application we would like to encourage piece-wise smooth (total variation-like) reconstructions, in our experiments we chose $\alpha_0 = 2.01$.

For the choice of θ_0, we note that the mean of Gamma(α_0, θ_0) is $\alpha_0\theta_0$, which should be approximately equal to α^{-1}, where α is obtained by applying one of the above regularization parameter selection methods to (7) with $\mathbf{C} = \mathbf{D}_1^T\mathbf{D}_1 + \mathbf{D}_2^T\mathbf{D}_2$. Thus we advocate choosing $\theta_0 = 1/(\alpha\alpha_0)$.

Figure 4 shows the reconstructions $\mathbf{x}_k$ and $\boldsymbol{\theta}_k$ that were obtained after 6 iterations of the algorithm, with α (and hence $\theta_0 = 1/(2.01\alpha)$) chosen using the GCV method.

4. Conclusions

We have presented regularization parameter choice methods for penalized negative-log Poisson likelihood problems arising in positron emission tomography. These methods are extensions of the discrepancy principle, generalized cross validation, and unbiased predictive risk estimation for regularization parameter choice in the least squares setting. The approach set forth here corresponds to a minor modification of that presented in [5].

The numerical results show that the methods yield good estimates of the regularization parameter and, moreover, that they can be used for choosing values of the hyper-parameters in the hierarchical regularization approach of [3].

References

[1] Sangtae Ahn and Jeffrey Fessler, *Globally Convergent Image Reconstruction for Emission Tomography Using Relaxed Ordered Subsets Algorithms*, IEEE Transactions on Medical Imaging, 22(5), pp. 613-626, 2003.

[2] Johnathan M. Bardsley, *An Efficient Computational Method for Total Variation-Penalized Poisson Likelihood Estimation*, Inverse Problems and Imaging, vol. 2, no. 2, 2008, pp. 167 - 185.

[3] Johnathan M. Bardsley, Daniela Calvetti, and Erkki Somersalo, *Hierarchical regularization for edge-preserving reconstruction of PET images*, submitted.

[4] Johathan M. Bardsley and John Goldes, *An Iterative Method for Edge-Preserving MAP Estimation when Data-Noise is Poisson*, accepted in the SIAM Journal on Scientific Computing.

[5] Johathan M. Bardsley and John Goldes, *Regularization Parameter Selection Methods for Ill-Posed Poisson Maximum Likelihood Estimation*, accepted in Inverse Problems.

[6] J. M. Bardsley and C. R. Vogel, *A Nonnnegatively Constrained Convex Programming Method for Image Reconstruction*, SIAM Journal on Scientific Computing, 25(4), 2004, pp. 1326-1343.

[7] D. Calvetti and E. Somersalo, *Hypermodels in the Bayesian Imaging Framework*, Inverse Problems, **24**, 2008, 034013 (20pp) doi: 10.1088/0266-5611/24/3/034013.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin, *Maximum Likelihood From Incomplete Data via the EM algorithm*, Journal of the Royal Statistical Society, B, 39, pp. 1-38, 1977.

[9] Jeffrey Fessler, *Penalized Weighted Least Squares Image Reconstruction for Positron Emission Tomography*, IEEE Transactions on Medical Imaging, 13(2), pp. 290-300, June 1994.

[10] Jeffrey Fessler and Alfred O. Hero, *Penalized Maximum-Likelihood Image Reconstruction Using Space*

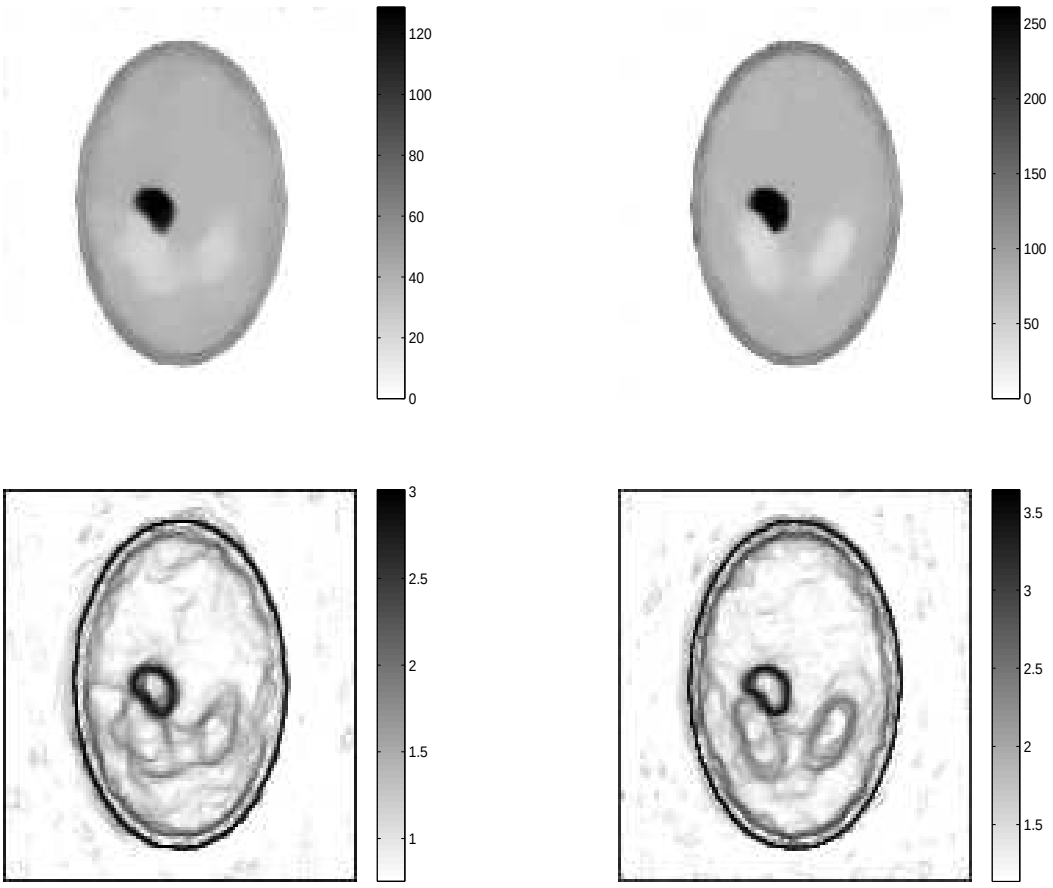

Figure 4. *On the top row are plots of the reconstructions that were obtained using hierarchical regularization. The reconstruction and $\boldsymbol{\theta}$ estimate from the data with SNR 5 is on the left and the reconstruction and $\boldsymbol{\theta}$ estimate from the data with SNR 20 is on the right.*

Alternation Generalized EM Algorithms, IEEE Transactions on Image Processing, 4(10), pp. 1417-29, Oct. 1995.

[11] J. A. Fessler and A. O. Hero, *Space-alternating generalized EM algorithms*, IEEE Transactions on Signal Processing, 42(10), pp. 2664-2677, Oct. 1994.

[12] G. Golub and U. von Matt, *Generalized Cross-Validation for Large-Scale Problems*, Journal of Computational and Graphical Statistics, Vol. 6, No. 1, 1997, pp. 1-34.

[13] Peter Green, *Bayesian reconstructions from emission tomography data using a modified EM algorithm*, IEEE Transactions on Medical Imaging, 9(1), pp. 84-93, March 1990.

[14] T. Herbert and R. Leahy, *A generalized EM algorithm for 3-D Bayesian reconstruction from Poisson data using Gibbs priors*, IEEE Transactions on Medical Imaging, 8(2), pp. 194-202, June 1989.

[15] T. Hsiao, A. Rangarajan, and G. Gindi, *Bayesian image reconstruction for transmission tomography using deterministic annealing*, Journal of Electronic Imaging, Vol. 12, 7 (2003), doi:10.1117/1.1526103.

[16] Jari Kaipio and Erkki Somersalo, *Satistical and Computational Inverse Problems*, Springer 2005.

[17] K. Lange and R. Carson, *EM reconstruction algorithms for emission and transmission tomography*, Journal of Computer Assisted Tomography, 8, pp. 306-316, 1984.

[18] S.-J. Lee, A. Rangarajan, and G. Gindi, *Bayesian image reconstruction in SPECT using higher order mechanical models as priors*, IEEE Tansactions on Medical Imaging, 14(4), pp. 669-680, 1995.

[19] Jodi Mead and Rosie Renaut, *A Newton root-finding algorithm for estimating the regularization parameter for solving ill-conditioned least squares problems*, Inverse Problems 25, 2, 2009.

[20] E. U. Mumcuoglu, R. Leahy, S. R. Cherry, Z. Zhou, *Fast Gradient-based methods for Bayesian reconstruction of transmission and Emission PET images*, IEEE Transactions on Medical Imaging, 13, pp. 687-701, 1994.

[21] John M. Ollinger and Jeffrey A. Fessler, *Positron-Emission Tomography*, IEEE Signal Processing Magazine, January 1997.

[22] L. A. Shepp and Y. Vardi, *Maximum likelihood reconstruction in positron emission tomography*, IEEE Transactions on Medical Imaging, vol. MI-1, pp. 113-122, 1982.

[23] G. Wahba, *Practical Approximate Solutions to Linear Operator Equations When the Data are Noisy*, SIAM Journal on Numerical Analysis, vol. 14, 1977, pp. 651-667.

[24] Y. Vardi, L. A. Shepp, and L. Kaufman, *A stastical model for positron emission tomography*, Journal

of the American Statistical Association, 80, pp. 8-37, 1985.

[25] Curtis R. Vogel, *Computational Methods for Inverse Problems*, SIAM, Philidelphia, 2002.

[26] M. Yavuz and J. A. Fessler, *New statistical models for randoms-precorrected PET scans*, Information Processing in Medical Im., J Duncan and G Gindi, editor. Springer-Verlag, Berlin, pp. 190-203, 1997.

[27] D. F. Yu and J. A. Fessler, *Edge-Preserving Tomographic Reconstruction with Nonlocal Regularization*, IEEE Trans. on Medical Imaging, 21(2), 2002, pp. 159-173.