An Efficient MCMC Method for Uncertainty Quantification in Inverse Problems

Johnathan M. Bardsley

Department of Mathematical Sciences
University of Montana
Missoula, Montana 59812
USA

E-mail: bardsleyj@mso.umt.edu

MSC numbers: 15A29, 62F15, 65F22, 94A08

Keywords: inverse problems, regularization, image reconstruction, Bayesian inference, Markov chain Monte Carlo, uncertainty quantification.

Abstract. The connection between Bayesian statistics and the technique of regularization for inverse problems has been given significant attention in recent years. For example, Bayes' law is frequently used as motivation for variational regularization methods of Tikhonov type. In this setting, the regularization function corresponds to the negative-log of the prior probability density; the fit-to-data function corresponds to the negative-log of the likelihood; and the regularized solution corresponds to the maximizer of the posterior density, known as the maximum a posteriori (MAP) estimator. While a great deal of attention has been focused on the development of techniques for efficient computation of MAP estimators (regularized solutions), less explored is the problem of uncertainty quantification, which corresponds to the problem of determining the shape, at least to some degree, of the posterior density in high probability regions. One way to do this is to sample from the posterior density using a Markov chain Monte Carlo (MCMC) method. In this paper, we present an MCMC method for use on linear inverse problems with independent and identically distributed Gaussian noise and Gaussian priors (quadratic regularization functions). From the MCMC samples, an estimator (regularized solution), and measures of variability in the estimator, are computed. Additionally, samples of the noise and prior precision parameters are computed, making regularization parameter selection unnecessary.

1. Introduction

In this paper, we consider linear models of the form

$$\mathbf{b} = \mathbf{A}\mathbf{x} + \boldsymbol{\eta}, \tag{1}$$

where $\mathbf{b} \in \mathbb{R}^m$ corresponds to observed data; $\mathbf{x}$ is the $n \times 1$ vector of unknowns; $\mathbf{A}$ is the $m \times n$ forward model matrix obtained via a numerical discretization of the forward map; and $\boldsymbol{\eta}$ is an $m \times 1$ independent and identically distributed (iid) Gaussian random vector with variance σ^2 across all pixels. We note that if $\boldsymbol{\eta} \sim N(\mathbf{0}, \boldsymbol{\Lambda})$, where the covariance matrix $\boldsymbol{\Lambda}$ is not constant diagonal, multiplying (1) by $\boldsymbol{\Lambda}^{-1/2}$ yields an iid Gaussian noise model with variance 1 across all pixels. Non-Gaussian noise models are outside the scope of this paper.

For inverse problems arising in imaging applications, the values of m and n are typically large. This is natural if pixel representations of the observed and unknown images are used. However, it is also necessary when the forward map is compact with infinite dimensional domain and range [12] and no other information about the unknown is given. Moreover, compactness in the forward map results in an ill-conditioned matrix $\mathbf{A}$, with eigenvalues clustered near zero corresponding to high frequency modes in both the unknown and observations.

Instabilities in solutions of inverse problems occur due to the fact that the subspace in which the noise $\boldsymbol{\eta}$ lives overlaps nontrivially with the subspace spanned by the singular vectors of $\mathbf{A}$ with extremely small, positive singular values. To illustrate, if we assume that $\mathbf{A}$ is invertible, multiplying equation (1) by $\mathbf{A}^{-1}$ yields $\mathbf{A}^{-1}\mathbf{b} = \mathbf{x} + \mathbf{A}^{-1}\boldsymbol{\eta}$, which will be dominated by the noise term $\mathbf{A}^{-1}\boldsymbol{\eta}$ since some of the singular vectors of $\mathbf{A}^{-1}$ spanning the noise subspace have extremely large singular values. For a more rigorous discussion see the texts [5, 12].

Regularization is the standard technique for handling such instabilities. In this paper we consider Tikhonov regularization, which we take to have the general form

$$T(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 + \alpha\, \mathbf{x}^T\mathbf{C}\mathbf{x}, \tag{2}$$

where α is the regularization parameter and $\mathbf{C}$ the regularization matrix. Various choices of $\mathbf{C}$ are used in practice, and a number of methods exist for choosing an appropriate value for α. For general discussions of regularization, see one of the many excellent tests on the subject, e.g., [4, 5, 12].

What is missing from standard treatments of inverse problems and Tikhonov regularization is a statistical formulation of the problem. Such motivation follows from Bayes' Theorem and has been recently exploited by many authors, most significantly in the texts [1, 9]. Bayes' Theorem states that if $p(\mathbf{b}|\mathbf{x})$ is the probability for the data $\mathbf{b}$ given the unknown $\mathbf{x}$ (i.e. the likelihood function), and $p(\mathbf{x})$ is the assumed probability distribution for the unknown $\mathbf{x}$ (i.e. the prior), then the probability of $\mathbf{x}$ given $\mathbf{b}$ satisfies

$$p(\mathbf{x}|\mathbf{b}) \propto p(\mathbf{b}|\mathbf{x})p(\mathbf{x}). \tag{3}$$

Here $p(\mathbf{x}|\mathbf{b})$ is known as the posterior probability density. The standard approach taken in the inverse problems literature is to compute the maximizer of $p(\mathbf{x}|\mathbf{b})$, which is appropriately named the maximum a posterior (MAP) estimator. Equivalently, one can minimize '$-\ln p(\mathbf{x}|\mathbf{b})$', which gets us close to (2) as we will see in a moment.

Given our assumptions above regarding (1), the likelihood function has the form

$$p(\mathbf{b}|\mathbf{x}) \propto \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2\right).$$

Assuming, furthermore, that $\mathbf{x} \sim N(\mathbf{0}, \delta^{-1}\mathbf{C}^{-1})$, our prior has the form

$$p(\mathbf{x}) \propto \exp\left(-\frac{\delta}{2}\mathbf{x}^T\mathbf{C}\mathbf{x}\right),$$

where the inverse covariance $\delta\mathbf{C}$ is called the precision matrix, and hence the MAP estimator is the minimizer of

$$-\ln p(\mathbf{x}|\mathbf{b}) \propto \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 + (\sigma^2\delta)\, \mathbf{x}^T\mathbf{C}\mathbf{x}. \tag{4}$$

Thus we can equate the regularization parameter α in (2) with the product of σ^2 and δ. In cases where the noise variance σ^2 is known a priori, (2) and (4) are equivalent, and it remains to estimate α or δ, depending on which approach is taken.

The above discussion illustrates the current paradigm in computational inverse problems, which is to work toward writing down a variational problem, or corresponding Euler-Lagrange equation, that is solved using an efficient computational method. The resulting point estimator is then presented as a regularized solution to the problem. As a result, much recent research in inverse problems (this author's included) sits squarely in its intersection with computational math [4, 5, 9, 11, 12], whether a Bayesian formulation such as (4), or the more standard (2), are used.

However, the MAP estimator insufficiently characterizes the posterior density $p(\mathbf{x}|\mathbf{b})$, since it only provides the location of it's peak. To quantify uncertainty in the

MAP estimator, knowledge of the shape of the posterior in a neighbor of the estimator is needed. Another approach is to compute samples from the posterior density using a Markov chain Monte Carlo (MCMC) method, and then to use those samples both to compute an estimator, e.g., the sample mean, as well as to quantify uncertainty, e.g., by computing credibility intervals (Bayesian confidence intervals) for each sampled parameter. This is the general approach taken here. Similar approaches have been studied by researches in the field of inverse problems; see, e.g. [1, 8, 9, 10, 13].

In order to define the posterior density function, in addition to (1), we assume a Gaussian prior on the unknown $\mathbf{x}$, and Gamma hyper-prior probability densities on the noise precision $\lambda = 1/\sigma^2$ and the prior precision parameter δ. The use such hierarchical models in inverse problems has been studied by other researchers, see e.g., [1, 2, 7], though in these papers the focus is mainly on computing MAP estimators. In this paper, following [6], our prior and hyper-priors are chosen because they are conjugate to the normal distribution, making efficient Markov chain Monte Carlo (MCMC) sampling possible.

The paper is organized as follows. In the next section, we present the posterior probability density function. Then, in Section 3, we present the MCMC method used to sample from the posterior density. Finally, in Section 4, we test the methodology on some standard examples from image processing, and finish with conclusions in Section 5.

2. Definition of the posterior probability density function

As was stated above, we assume that the unknowns include $\mathbf{x}$, $\lambda = 1/\sigma^2$ and δ, so that the posterior density function has the form $p(\mathbf{x}, \lambda, \delta|\mathbf{b})$. Defining probability distributions for λ and δ defines a hierarchical model; note that prior for $\mathbf{x}$ in turn depends upon the parameter δ. Bayes' Theorem in this setting can be expressed

$$p(\mathbf{x}, \lambda, \delta|\mathbf{b}) \propto p(\mathbf{b}|\mathbf{x}, \lambda)p(\lambda)\, p(\mathbf{x}|\delta)p(\delta), \tag{5}$$

where, similar to above, the likelihood is given by

$$p(\mathbf{b}|\mathbf{x}, \lambda) \propto \lambda^{n/2} \exp\left(-\frac{\lambda}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2\right), \tag{6}$$

and the prior by

$$p(\mathbf{x}|\delta) \propto \delta^{n/2} \exp\left(-\frac{\delta}{2}\mathbf{x}^T\mathbf{C}\mathbf{x}\right). \tag{7}$$

The probability densities for λ and δ are taken to be Gamma distributions

$$p(\lambda) \propto \lambda^{\alpha_\lambda - 1} \exp(-\beta_\lambda \lambda), \tag{8}$$

$$p(\delta) \propto \delta^{\alpha_\delta - 1} \exp(-\beta_\delta \delta). \tag{9}$$

Note that the Gamma distribution $\Gamma(\alpha, \beta)$ has a probability density function satisfying $g(t|\alpha, \beta) \propto t^{\alpha-1} \exp(-\beta\, t)$. Thus (5) can be written

$$p(\mathbf{x}, \lambda, \delta|\mathbf{b}) \propto \lambda^{n/2+\alpha_\lambda - 1} \delta^{n/2+\alpha_\delta - 1} \exp\left(-\frac{\lambda}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 - \frac{\delta}{2}\mathbf{x}^T\mathbf{C}\mathbf{x} - \beta_\lambda\lambda - \beta_\delta\delta\right). \tag{10}$$

In order to completely define the hyper-priors, we must define the parameter pairs $(\alpha_\lambda, \beta_\lambda)$ and $(\alpha_\delta, \beta_\delta)$. For this, following [6], we take $\alpha_\lambda = \alpha_\delta = 1$ and $\beta_\lambda = \beta_\delta = 10^{-4}$. Then the hyper-priors can be deemed to be "uniformative", since the mean and variance of the corresponding Gamma distributions is $\alpha/\beta = 10^4$ and $\alpha/\beta^2 = 10^8$, respectively. Note that no other parameters remain to be defined.

In [7], the problem of computing a minimizer of (5) is considered. In particular, existence of minimizers is proved and an efficient method for computing a MAP estimator is presented. Similar hierarchical models are also considered in [2]. However, here we make use of the conjugacy relationship between the Gamma and Normal densities for MCMC sampling.

Note that the full conditional densities have the form

$$p(\mathbf{x}|\lambda, \delta, \mathbf{b}) \propto \exp\left(-\frac{\lambda}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 - \frac{\delta}{2}\mathbf{x}^T\mathbf{C}\mathbf{x}\right), \tag{11}$$

$$p(\lambda|\mathbf{x}, \delta, \mathbf{b}) \propto \lambda^{n/2+\alpha_\lambda-1} \exp\left(\left[-\frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 - \beta_\lambda\right]\lambda\right), \tag{12}$$

$$p(\delta|\mathbf{x}, \lambda, \mathbf{b}) \propto \delta^{n/2+\alpha_\delta-1} \exp\left(\left[-\frac{1}{2}\mathbf{x}^T\mathbf{C}\mathbf{x} - \beta_\delta\right]\delta\right). \tag{13}$$

Conjugacy guarantees that each of (11)-(13) corresponds to a distribution in the same family as the corresponding prior/hyper-prior. In particular, note that

$$\mathbf{x}|\lambda, \delta, \mathbf{b} \sim N\left((\lambda\mathbf{A}^T\mathbf{A} + \delta\mathbf{C})^{-1}\lambda\mathbf{A}^T\mathbf{b}, (\lambda\mathbf{A}^T\mathbf{A} + \delta\mathbf{C})^{-1}\right), \tag{14}$$

$$\lambda|\mathbf{x}, \delta, \mathbf{b} \sim \Gamma\left(n/2 + \alpha_\lambda, \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 + \beta_\lambda\right), \tag{15}$$

$$\delta|\mathbf{x}, \lambda, \mathbf{b} \sim \Gamma\left(n/2 + \alpha_\delta, \frac{1}{2}\mathbf{x}^T\mathbf{C}\mathbf{x} + \beta_\delta\right). \tag{16}$$

The power in (14)-(16) lies in the fact that samples from these three distributions can be easily computed using standard statistical software, though numerical linear algebra techniques may be needed for (14). A Gibbsian approach then yields an efficient Markov chain Monte Carlo (MCMC) sampling scheme for the posterior density function, as we will see in Section 3.

However, it remains to define the regularization matrix $\mathbf{C}$.

2.1. Defining the regularization matrix via Gaussian Markov random fields

Staying with the statistical flavor of the presentation, we use Gaussian Markov random fields (MRF's), as in [6], to define $\mathbf{C}$.

A MRF on the pixel grid is defined by specifying a neighborhood system at each pixel, as well as a set of n conditional densities $\{p(x_i|\mathbf{x}_{\partial_i}),\ i = 1, \ldots, n\}$, where $\mathbf{x}_{\partial_i} = \{x_j | j \in \partial_i\}$, with $j \in \partial_i$ if $i \neq j$ and pixels i and j are neighbors. We note that MRF's have the property that for all i, the full conditional density $p(x_i|\mathbf{x}_{-i}) = p(x_i|\mathbf{x}_{\partial_i})$, where $\mathbf{x}_{-i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$, i.e. it depends only upon it's neighbors.

For a given neighborhood system, we define the conditional densities to be Gaussian and of the form

$$x_i|\mathbf{x}_{\partial_i} \sim N(\bar{x}_{\partial_i}, 1/n_i),$$

where n_i is the number of neighbors belonging to x_i and $\bar{x}_{\partial_i}$ is the average of the x_j's neighboring x_i. The resulting joint Gaussian density for $\mathbf{x}$ is then given by [6]

$$p(\mathbf{x}) \propto |\mathbf{C}|^{1/2} \exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{C}\mathbf{x}\right),$$

where

$$[\mathbf{C}]_{ij} = \begin{cases} n_i & i = j, \\ -1 & j \in \partial_i, \\ 0 & \text{otherwise.} \end{cases}$$

Note that if zero boundary pixels are assumed and a standard first-order neighborhood is used – which in 1D includes the pixels to the left and right, and in 2D those to the left, right, above, and below – the standard discrete negative-Laplacian results. Also, recall that in practice, we include the scaling parameter δ (see (7)), which is necessary in any practical application [6].

3. MCMC sampling of the posterior distribution

As was stated above, the full conditional distributions (14)-(16) can be used to define an efficient Gibbs sampler for sampling from the posterior distribution $p(\mathbf{x}, \delta, \lambda|\mathbf{b})$. It is given as follows.

An MCMC Method for Sampling from $p(\mathbf{x}, \delta, \lambda|\mathbf{b})$.

0. Initialize δ_0, and λ_0, and set $k = 0$;
1. Compute $\mathbf{x}^k \sim N\left((\lambda_k\mathbf{A}^T\mathbf{A} + \delta_k\mathbf{C})^{-1}\lambda_k\mathbf{A}^T\mathbf{b}, (\lambda_k\mathbf{A}^T\mathbf{A} + \delta_k\mathbf{C})^{-1}\right)$;
2. Compute $\lambda_{k+1} \sim \Gamma\left(n/2 + \alpha_\lambda, \frac{1}{2}\|\mathbf{A}\mathbf{x}^{k+1} - \mathbf{b}\|^2 + \beta_\lambda\right)$;
3. Compute $\delta_{k+1} \sim \Gamma\left(n/2 + \alpha_\delta, \frac{1}{2}(\mathbf{x}^{k+1})^T\mathbf{C}\mathbf{x}^{k+1} + \beta_\delta\right)$;
4. Set $k = k + 1$ and return to Step 1.

Since the parameters λ and δ are scalar, the scalar random draws required in Steps 2 and 3 are very efficient and easy to compute given the appropriate software.

For inverse problems, the computational bottleneck occurs in Step 1. However, if an easily invertible square-root of the precision matrix is known, i.e.

$$\lambda_k\mathbf{A}^T\mathbf{A} + \delta_k\mathbf{C} = \mathbf{L}_k\mathbf{L}_k^T,$$

then the sample can be computed as follows:

$$\mathbf{x}^k = \mathbf{L}_k^{-T}(\mathbf{L}_k^{-1}\lambda_k\mathbf{A}^T\mathbf{b} + \mathbf{v}), \quad \text{where} \quad \mathbf{v} \sim N(\mathbf{0}, \mathbf{I}). \tag{17}$$

Note that if $\mathbf{v} \sim N(\mathbf{0}, \mathbf{I})$, then

$$\begin{aligned}\mathbf{L}_k^{-T}\mathbf{v} &\sim N(\mathbf{0}, \mathbf{L}_k^{-T}\mathbf{L}_k^{-1}) \\ &= N(\mathbf{0}, (\lambda_k \mathbf{A}^T\mathbf{A} + \delta_k \mathbf{C})^{-1}).\end{aligned}$$

A key requirement for this approach to be efficient is that multiplication by $\mathbf{L}_k^{-T}$ and $\mathbf{L}_k^{-1}$ is efficient. This will be the case if, for example, $\mathbf{L}_k$ is obtained from the Cholesky factorization, or from an eigenvalue decomposition, of $\lambda_k \mathbf{A}^T\mathbf{A} + \delta_k \mathbf{C}$.

3.1. Assessing MCMC chain convergence

A number of techniques can be employed to assess convergence of an MCMC chain. In [1] and elsewhere, the use of time series of graphs of individual sampled parameters is discussed. However, in [3], it is mentioned that this is a "notoriously unreliable method of assessing convergence", and several supporting references are given. Moreover, this approach is infeasible for the large scale problems of interest to us.

The recommended approach presented in [3] requires the computation of multiple MCMC chains, with radonomly chosen starting points, based on observations that the variance within an individual chain will often converge before the variance between chains converges. With multiple chains in hand, a statistic for each sampled parameter is then computed, whose value provides a measure of convergence.

This statistic is defined as follows. Suppose we compute n_r parallel chains, each of length n_s (after discarding the first half of the simulations), and that $\{\psi_{ij}\}$, for $i = 1, \ldots, n_s$ and $j = 1, \ldots, n_r$, is the collection of samples of a single parameter. Then we define

$$B = \frac{n_s}{n_r - 1}\sum_{j=1}^{n_r}(\overline{\psi}_{\cdot j} - \overline{\psi}_{\cdot\cdot})^2, \quad \text{where} \quad \overline{\psi}_{\cdot j} = \frac{1}{n_s}\sum_{i=1}^{n_s}\psi_{ij}, \quad \text{and} \quad \overline{\psi}_{\cdot\cdot} = \frac{1}{n_r}\sum_{j=1}^{n_r}\overline{\psi}_{\cdot j}\,;$$

and

$$W = \frac{1}{n_r}\sum_{j=1}^{n_r} s_j^2, \quad \text{where} \quad s_j^2 = \frac{1}{n_s - 1}\sum_{i=1}^{n_s}(\psi_{ij} - \overline{\psi}_{\cdot j})^2.$$

Note that $\overline{\psi}_{\cdot j}$ and $\overline{\psi}_{\cdot\cdot}$ are the individual chain mean and overall sample mean, respectively. Thus B provides a measure of the variance between the n_r chains, while W provides a measure of the variance within individual chains.

The marginal posterior variance $\mathrm{var}(\psi|\mathbf{b})$ can then be estimated by

$$\widehat{\mathrm{var}}^{+}(\psi|\mathbf{b}) = \frac{n_s - 1}{n_s}W + \frac{1}{n_s}B,$$

which is an unbiased estimate under stationarity [3]. The statistic of interest to us, however, is

$$\widehat{R} = \sqrt{\frac{\widehat{\mathrm{var}}^{+}(\psi|\mathbf{b})}{W}}, \tag{18}$$

which declines to 1 as $n_s \to \infty$.

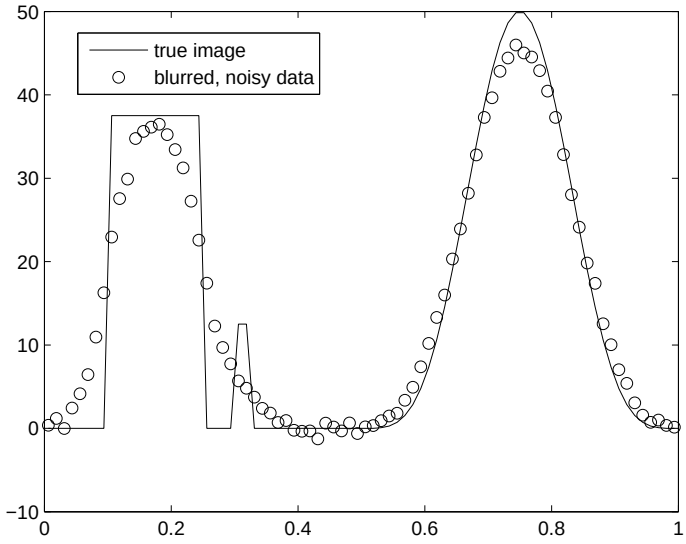

Figure 1. The one-dimensional true image and blurred noisy data.

Once $\widehat{R}$ is 'near' 1 for all sampled parameters, the $n_s n_r$ samples from the last half of all of the sequences together can be treated as samples from the target distribution [3]. We will see that in our examples, $\widehat{R}$ will be much nearer to 1 than 1.1, which is deemed acceptable in [3].

4. Numerical Examples

Suppose that for a given example, we compute n_r parallel MCMC chains, each of length n_s, using the algorithm of the previous section. Suppose, furthermore, that the first half of all sequences has already been discarded, and that the value of $\widehat{R}$ computed from the remaining samples, is sufficiently 'near' 1 for all sampled parameters. We now show how this output can be used to obtain parameter estimates, as well as for the quantification of uncertainty in those estimates.

4.1. A one-dimensional example

We begin with the one-dimensional image deblurring example of [12]. The data model is of the form (1), with $\mathbf{b} = \mathbf{Ax}$ obtained via mid-point quadrature applied to the convolution equation

$$b(s) = \int_0^1 A(s - s')x(s')ds',$$

with a Gaussian convolution kernel

$$A(s) = \frac{1}{\sqrt{\pi\gamma^2}} \exp(-s^2/2\gamma^2), \quad \gamma > 0.$$

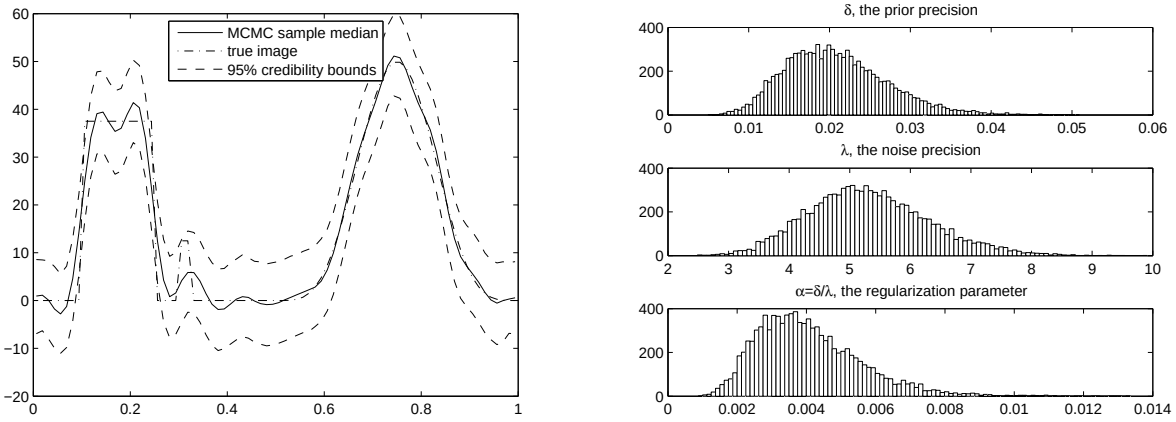

Figure 2. One dimensional example. On the left are plots of the median, and the 0.025 and 0.975 quantiles of the image samples. On the right are histograms of the samples of the precision parameters δ and λ, as well as of the regularization parameter $\alpha = \delta/\lambda$.

Then $\mathbf{A}$ has the form

$$[\mathbf{A}]_{ij} = \frac{h}{\pi\gamma^2} \exp\left(-\frac{((i-j)h)^2}{2\gamma^2}\right), \quad 1 \le i, j \le n, \tag{19}$$

where $h = 1/n$ with n the number of grid points in $[0, 1]$. We use $n = 80$, and the resulting $\mathbf{A}$ has full column rank with condition number on the order of 10^{16}, resulting in a severely ill-conditioned problem. The true image is given by the solid line in Figure 1, and the data $\mathbf{b}$ is generated using (1), with the noise variance σ^2 chosen so that the noise strength is 2% that of the signal strength.

We sample from the posterior density $p(\mathbf{x}, \lambda, \delta)$ by computing 40 MCMC chains each of length 500, which took approximately 30 seconds. The maximum $\widehat{R}$ value at the end of the run was 1.011. The initial values δ_0 and λ_0 in Step 0 were chosen randomly from the uniform distributions $U(0, 10)$ and $U(0, 1)$, respectively. And finally, direct image sampling (17) was used in Step 1 of the algorithm, where $\mathbf{L}_k\mathbf{L}_k^T$ is the Cholesky factorization of $\lambda_k\mathbf{A}^T\mathbf{A} + \delta_k\mathbf{C}$.

From the image samples, on the left in Figure 1, we plot the median (or 0.5 quantile) as our reconstruction, and 95% credibility images given by the 0.025 and 0.975 quantiles of the samples at each pixel; these three quantiles were computed using MATLAB's `quantile` function. From the samples for λ and δ, on the right in Figure 2, we plot histograms for λ, δ, and the regularization parameter $\alpha = \delta/\lambda$, which has a 95% credibility interval [0.0016,0.0067]. Finally, as a verification of our statistical model, we note that the true noise precision $1/\sigma^2 \approx 5.35$ is contained within the 95% credibility interval for λ, $[3.56, 7.46]$, computed using MATLAB's `quantile` function.

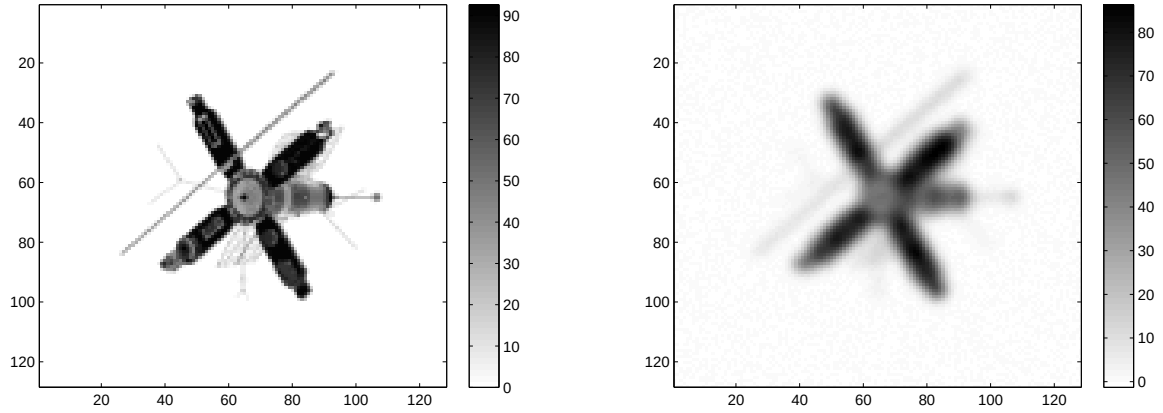

Figure 3. On the left is the two-dimensional true image, and on the right is the blurred noisy data.

4.2. The Two-Dimensional Case

In many inverse problems applications, e.g. imaging, the spatial domain is two-dimensional. Thus we must show that our method is also effective on two-dimensional problems. Two-dimensional convolution has the form

$$b(s,t) = \int_0^1 \int_0^1 A(s-s', t-t')x(s',t')ds'dt'.$$

As above, we choose a Gaussian convolution kernel A, and discretize using mid-point quadrature on an 128×128 uniform computational grid over [0,1]×[0,1]. Periodic boundary conditions for the image are assumed, so that $\mathbf{A}$ is an $n^2 \times n^2$ block circulant with circulant blocks matrix, and hence is diagonalizable by the two-dimensional discrete Fourier transform (DFT) [5, 12]. Finally, the data $\mathbf{b}$ is generated using (1) with the noise variance σ^2 chosen so that the noise strength is 2% that of the signal strength. The true image and data are shown in Figure 3.

We sample from the posterior density $p(\mathbf{x}, \lambda, \delta)$ by computing 25 MCMC chains each of length 500. We choose a smaller number of chains than in the one-dimensional example due to memory constraints. The max $\widehat{R}$ value at the end of the run was 1.023. We use direct image sampling (17) in Step 1 of the MCMC algorithm, with $\mathbf{L}_k\mathbf{L}_k^T$ computed efficiently using the DFT-based diagonalization of $\mathbf{A}$. The computation of the chain took only 68 seconds due to the efficiency of the fast Fourier transform; indeed memory issues were more restrictive than computation time.

We plot the median of the sampled images as the reconstruction on the left in Figure 4. From the samples for λ and δ, on the right in Figure 2, we plot histograms for λ, δ, and the regularization parameter $\alpha = \delta/\lambda$, which has a 95% credibility interval [0.00036,0.00041]. The overall approach is validated by the fact that the true noise precision $1/\sigma^2 = 8.09$ is contained within the sample 95% credibility interval for λ, [7.81, 8.20].

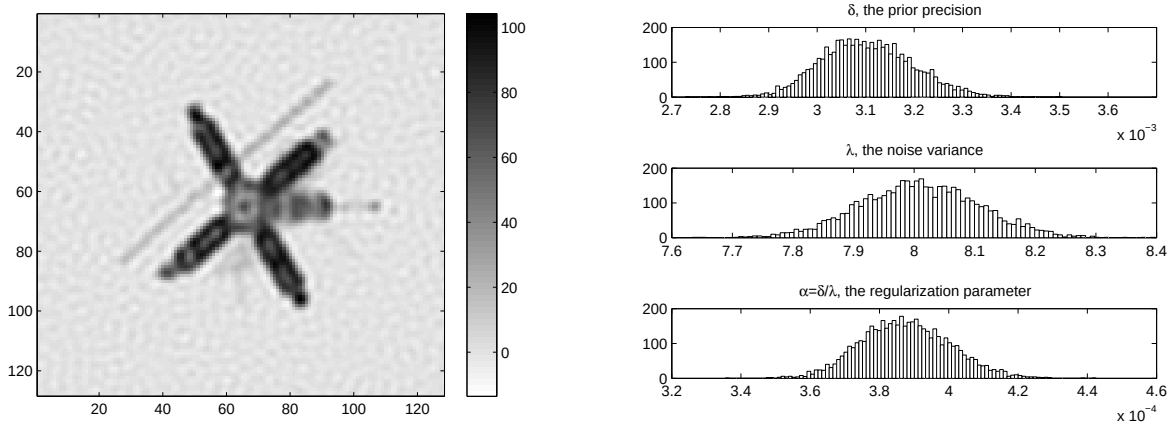

Figure 4. Two-dimensional example. On the left is the median image. On the right are histograms of the samples of the precision parameters δ and λ, as well as of the regularization parameter $\alpha = \delta/\lambda$.

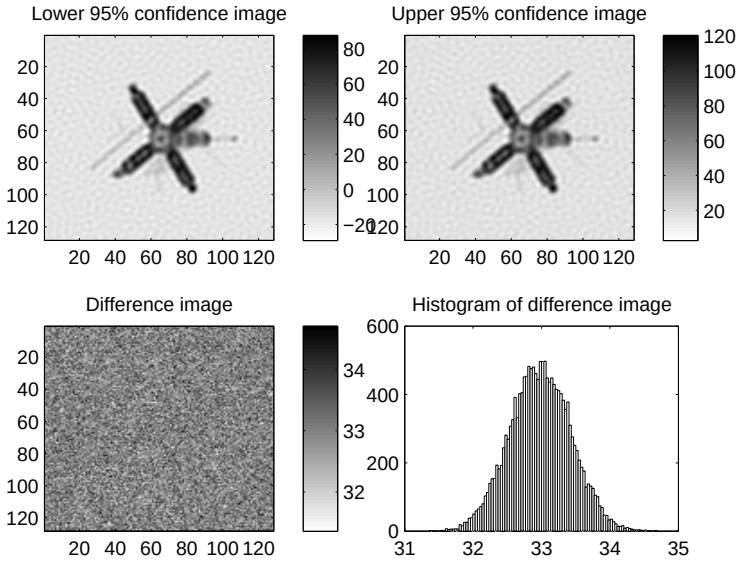

Figure 5. Two-dimensional example. In the upper-left and right are the lower and upper 95% credibility images, respectively. In the lower-right is the differnce of these two images, and on the lower-left is a histogram of this difference image.

In order to quantify uncertainty in the image samples, we plot the upper and lower 95% credibility images, obtained using MATLAB's `quantile` function, in the upper-left and -right, respectively, in Figure 5; note the difference in the intensity values for these images compared to the median image given in Figure 4. The difference between the upper and lower 95% credibility images is given in the lower-left of Figure 5, and a histogram of this difference image is shown on the lower-right, which has mean 33 and variance 0.2.

5. Conclusions

In this paper, we present an MCMC sampling approach for solving large-scale, linear inverse problems with independent and identically distributed Gaussian noise. We assume that the variance of the noise is unknown, but that it's inverse (the precision) λ arises from an uninformative Gamma random variable. The regularization function is determined by the prior probability density function, which we define using a Gaussian Markov random field. A scaling factor for the prior δ, akin to a regularization parameter, is also assumed to arise from an uninformative Gamma random variable. From this hierarchical model, a posterior probability density function is defined.

The standard approach at this point would be to maximize the posterior density function, yielding the maximum a posteriori (MAP) estimator. Instead we sample from the posterior using a Markov chain Monte Carlo (MCMC) method. Our MCMC method takes advantage of conjugacy relationships between likelihood, prior, and hyper-prior densities, making it very efficient (see CPU times above).

Using our MCMC method, we compute samples from several parallel chains and monitor convergence both within and between chains using the statistic $\widehat{R}$ [3], for each parameter sampled. Once $\widehat{R}$ is deemed to be sufficiently 'near' 1 for all parameters, the last half of all chains are treated as samples from the posterior.

The median of the samples is taken to be the reconstructed image, and 95% credibility intervals (and images) are also computed using MATLAB's `quantile` function. Moreover, histograms are created from the samples of δ and λ, which are in turn used to create a histogram for the regularization parameter $\alpha = \delta/\lambda$, illustrating that in this approach, regularization parameter selection is unnecessary. In all cases, the 95% credibility interval for λ includes the true noise precision $1/\sigma^2$, which serves to validate the approach.

We conclude that the Bayesian MCMC approach that we've presented in this paper is both efficient and effective, given that it yields high quality reconstructions, does not require the use of a regularization parameter choice method, and provides a means of uncertainty quantification for all estimated parameters, including the reconstructed image.

Acknowledgements

This work was supported by the National Science Foundation under grant DMS-0915107. The author would like to thank the University of Montana and the Department of Physics at the University of Otago, New Zealand, and Dr. Colin Fox in particular, for their support during his 2010-11 sabbatical year.

References

[1] D. Calvetti and E. Somersalo, *Introduction to Bayesian Scientific Computing*, Springer 2007.

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

[3] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis, Second Edition*, Chapman & Hall/CRC, Texts in Statistical Science, 2004.

[4] P. C. Hansen, *Rank-Deficient and Discrete Ill-Posed Problems*, SIAM, Philadelphia, 1997.

[5] P. C. Hansen, J. Nagy, and D. O'Leary, *Deblurring Images: Matrices, Spectra, and Filtering*, SIAM, Philadelphi, 2006.

[6] Dave Higdon, *A primer on space-time modelling from a Bayesian perspective*, Los Alamos Nation Laboratory, Statistical Sciences Group, Technical Report, LA-UR-05-3097.

[7] B. Jin and J. Zou, *Augmented Tikhonov Regularization*, Inverse Problems, **25(2)**, 2009, doi: 10.1088/0266-5611/24/3/034013.

[8] J. P. Kaipio, V. Kolehmainen, E. Somersalo, and M. Vauhkonen, *Statistical inversion and Monte Carlo sampling methods in electrical impedance tomography*, Inverse Problems, **16(5)**, 2000, pp. 14871522.

[9] J. Kaipio and E. Somersalo, *Statistical and Computational Inverse Problems*, Springer 2005.

[10] G. Nicholls and C. Fox, *Prior modelling and posterior sampling in impedance imaging*, Bayesian Inference for Inverse Problems, Proc. SPIE 3459, 1998, pp. 116-127.

[11] A. N. Tikhonov, A. V. Goncharsky, V. V. Stepanov and A. G. Yagola, *Numerical Methods for the Solution of Ill-Posed Problems*, Kluwer Academic Publishers, 1990.

[12] C. R. Vogel, *Computational Methods for Inverse Problems*, SIAM, Philadelphia, 2002.

[13] D. Watzenig and C. Fox, *A review of statistical modeling and inference for electrical capacitance tomography*, Measurement Science and Technology, **20(5)**, 2009, doi:10.1088/0957-0233/20/5/052002.