Inverse Problems, Design and Optimization Symposium
Miami, Florida, USA, April 16-18, 2007

RECENT ADVANCES IN INFERENTIAL SOLUTIONS TO INVERSE PROBLEMS

Colin Fox
Department of Mathematics
The University of Auckland
Auckland, New Zealand
fox@math.auckland.ac.nz

ABSTRACT

Inferential solutions to inverse problems provide substantial advantages over over deterministic methods, such as: quantitative estimates with posterior (data-dependent) error estimates, predictive densities, model comparison, and direct support for optimal decisions. The ability to include arbitrary forward maps, and hence use high-level representations of the unknowns, allows structure-preserving model reduction and also allows 'classification' to be performed within the 'imaging' step. Since inferential methods make (provably) optimal use of data, the ability to reduce data to a minimal set gives cost savings in applications where collecting data is expensive.

The price of these advantages is presently the relatively high computational cost of sampling algorithms for computing estimates. Hence the most significant advances are in computational methods for sample-based inference in inverse problems. In this paper we review the inferential formulation of inverse problems, some reasons why it is necessary to take on the extra machinery of inferential solutions, the 'basement level' methods for computing inferential solutions, and summarize some recent advances in computational methods for inferential solutions to inverse problems.

INTRODUCTION

Inverse problems occur when observed data d depend on unknowns x via a measurement process, and we want to recover x from d, or at least answer quantitative questions about x given d. Simulation of the measurement process for given x defines the *forward map* $A : x \mapsto d$ giving data in the absence of errors.

Typical examples are the various modalities of imaging from wave scattering used in non-invasive medical diagnostics, geophysical prospecting, and industrial process monitoring. In these problems the forward map can be modelled using ideas from theoretical Physics, and simulation usually requires solution of a partial differential equation. Inverse problems also occur in a myriad of other settings such as inverse spectral problems (determining internal structure or shape from resonance frequencies), interferometric imaging, and mapping of flows subject to physical laws, to name just a few.

The classical, or deterministic, inverse problem is to invert the the function A to obtain unknowns x in terms of data d. Practical inverse problems are usually *ill-posed* by failing to have solutions or unique solutions, while idealized inverse problems in which all possible measurements are made are usually unstable, i.e., small changes in data d cause large or unbounded changes in recovered value(s) x. This latter property is routinely displayed by least-squares or maximum likelihood solutions to inverse problems, even when the number of data exceed the number of unknowns. For many inverse problems this behavior can be understood mathematically when the forward map is compact, implying that the inverse is discontinuous. Regularized inversion consists of applying a regular approximation to the inverse of A to give a *single* estimate of unknowns x – presenting a 'take it or leave it' solution. In contrast, inferential methods, as we will see, can make estimates and predictions by summarizing *all* feasible solutions.

Inferential Formulation of Inverse Problems

The ubiquitous presence of measurement errors, or noise, means that a practical measurement process is probabilistic, and the inverse problem is naturally a problem in statistical inference. To fix ideas, consider additive noise n with probability density function $\pi_N(n)$. The measurement process can then be written

$$d = A(x) + n \tag{1}$$

and we see that the data is now a random variable that is dependent on x. The (conditional) probability density for measuring d given that x is the true set of parameters is

$$\pi(d|x) = \pi_N(d - A(x)), \tag{2}$$

since the Jacobian determinant for the change of variables from n to d is 1. In this formulation, making a set of measurements corresponds to drawing a sample d from $\pi(d|x)$ which is a probability distribution parameterized by the unknowns x via the forward map A. Given a set of measurements d, the job in inverse problems is to work out what we can say about the parameter x. Statisticians have developed a beautiful and principled set of tools for exactly this problem[1] making the field of 'statistical inference'. That toolbox is applied here, so solution methods presented are structurally just sample-based inference as formulated in Bayesian statistics. However, inferential methods applied to inverse problems have a particular structure arising from the ill-posed nature of the inverse problem, and because the forward map can be modelled using Physics and simulated by intensive computation.

As a function of d, $\pi(d|x)$ is a probability density function with all the usual properties: it integrates to 1 and transforms as a distribution. As a function of x, $\pi(d|x)$ is a function but not a probability density: its integral equals 1 only by coincidence and it does not transform as a distribution. Hence we write $l(d|x)$ for $\pi(d|x)$ and refer to the *likelihood function.*

Most commonly the measurement error has an exponential family or Gibbs distribution [Kaipio and Somersalo, 2005], so the likelihood function takes the form

$$l(d|x) \propto \exp\{-\chi(d-A(x))\}$$

where $\chi(\cdot)$ is an 'energy' function. For example $\chi(y) = y^{\mathrm{T}}B^{-1}y/2$ when the noise comes from a Gaussian process with known covariance matrix B.

In a Bayesian formulation, inference about x is based on the posterior density

$$\pi(x|d) = \frac{l(d|x)\pi(x)}{\pi(d)} \tag{3}$$

where $\pi(x)$ denotes the prior density modelling beliefs about the unknown x independent of the data d. For our purposes it is sufficient to take $\pi(d)$ to be a finite constant[2], thus ensuring that the posterior density is normalizable.

Exploratory analyses typically employ a low-level (e.g. pixel or voxel) representation of the unknowns with a Gibbs or Markov random field (MRF) prior distribution [Geman and Geman, 1984]. This is the typical case in regularized inversion, which may be viewed as a special case of inferential methods. This follows since regularized inverses are the same as maximum a posteriori (MAP) estimates with regularization functionals that almost always correspond to a proper (or improper) Gibbs prior distribution written as the exponential of minus a norm (or semi-norm) of the unknown x. Then the posterior density has the form

$$\pi(x|d) \propto \exp\{-\chi(d-A(x))-\rho(x)\} \tag{4}$$

where χ and ρ are relatively simple functions.

Despite that similarity of mathematical form of many regularization functionals to the log GMRF, the formulation of a statistically-sensible prior distribution is a major practical difference between regularization and inferential methods. As remarked above, the prior density $\pi(x)$ models beliefs about the unknowns in the absence of data, and hence typical states in the support of the prior distribution should look like reasonable values of the unknown x. In a Bayesian analysis it is typical to test modelling assumptions by drawing several samples from the prior distribution and ensuring that they look reasonable. In contrast, typical regularization functionals would fail this test.

Also, since the prior distribution must be normalizable, the space X of allowable reconstruction must have finite (actually unit) volume. This is not only a philosophical issue, but also a practical issue that enables sampling algorithms to be provably convergent. In contrast, typical regularization functionals correspond to improper prior distributions, i.e. that are not normalizable. Attempts to naively apply sampling algorithms to (the exponential of minus) a regularization functional lead to ill-defined results depending on numerical happenstance.

A strong practical advantage of inferential methods is that any bias introduced by the prior may be calculated and corrected. This is easily done by drawing samples from the prior distribution and applying the solution/estimation procedure to obtain the density of estimates independent of measured data. Bias in estimates may be removed by suitably adjusting the prior distribution, usually by scaling the metric in some direction. A nice example of this procedure is given by Watzenig [2006] when estimating the area of an inclusion using electrical capacitance tomography.

For this reason, there is no single 'Bayesian analysis' since there is no single prior distribution, rather the prior distribution that should be used depends on the purpose of performing the inverse problem. For example, the prior distribution for estimating a 'best' reconstruction would be different to that for estimating some feature of the reconstruction, or from making a decision based on the reconstruction, etc.

As in the field of image analysis, geometric information about the unknowns may be included using a prior distribution based on an intermediate-level representation of the unknowns, such as Nicholls [1998] continuum triangulation of the plane (see e.g. Andersen et al. [2003]), or using a high-level representation of the type introduced by Grenander and Miller [1994]. These represent substantial difficulties for regularization methods since state spaces are typically not Euclidean, usually not equipped with a norm, and hence regularization functionals are not applicable. The ability to use these more informative representations is a substantial advantage of inferential (or model fitting) methods since there is the potential of dimensionality reduction.

Before looking into the details of *how* we may perform statistical inference for inverse problems, we examine a simple problem that demonstrates some of the reasons *why*

[1] Parameter estimation problems in astronomy led Laplace to a formulation of probability, that is seen as one of the starting points of Statistics.

[2] This number plays a central role in problems where *model selection* is at issue.

we should use inferential methods.

Mode and Mean in a Binary Imaging Problem

The first 'recent result' we visit is actually a 'recent understanding' that has been made possible by new algorithms. For some decades computational limitations have restricted the summary statistics that can be calculated to the *mode* of the posterior distribution — by using efficient optimization algorithms. However, in any inverse problem where the forward map is non-linear, or when measurement error is not Gaussian, or when using general prior distributions, or where image space is discrete, (so just about all inverse problems) there is no reason that the mode should be representative of the bulk of feasible reconstructions. We find, instead, that summary statistics based on the bulk of probability in the posterior distribution are required.

We consider a problem of recovering a binary (black and white) image after pixel-wise addition of zero-mean Gaussian noise. Recovery of binary images occurs in the practical setting of 'image segmentation' problems (see e.g. Kumar and Hebert [2006]). As we will see, this problem has a posterior distribution for which the mode is very different to the mean. We follow the analysis of Fox and Nicholls [2000] who calculated the mode exactly and estimated the mean using a provably convergent sampling algorithm.

Figure 1 shows the true image (left) and the pixel-wise degraded version (right).

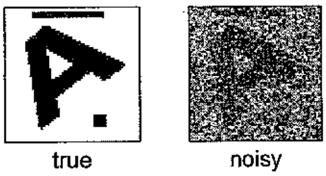

Figure 1. True binary image (left) and gray-scale image showing the image after pixel-wise addition of Gaussian noise (right).

The $N \times N$ binary image is $x = (x_1, x_2 \dots x_{N^2})$ where each $x_i \in \{0,1\}$. The pixel-wise corrupted version, $g = (g_1, g_2 \dots g_{N^2})$, is given by

$$g_i = x_i + \varepsilon_i$$

with each $\varepsilon_i \sim \text{Normal}(0, \sigma^2)$ with known variance σ^2.

In terms of the model for inverse problems in equation 1, it appears that the forward map is the identity (at least to the untrained eye). A common procedure in conventional imaging of binary images of is to treat the unknowns as gray-scale pixels and attempt inversion of the forward map. In this case it is certainly easy to invert the forward map, but it makes no improvement at all!

The likelihood function for image x given measurements g, may be written

$$l(g|x) \propto \prod_{i=1}^{N^2} \exp\left(-\frac{(x_i - g_i)^2}{2\sigma^2}\right).$$

In this form, it looks reasonable to perform inversion by requiring that each recovered pixels be binary valued, i.e. black or white, and finding the best-fit image in the least-squares sense. Since the noise is iid Gaussian, this equals the maximum likelihood solution

$$x_{\text{MLE}} = \arg\max_{x \in X} l(g|x).$$

That solution is shown in figure 2. As can be seen, x_{MLE} does not make a good estimate of the true image. Failure of the least-squares solution should present no surprise to practitioners in inverse problems where it is commonplace. In this case the failure is perhaps more expected when considering the equivalent form of the likelihood as a function of x, i.e. once data g is measured and hence fixed,

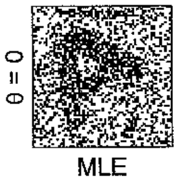

Figure 2. Least-squares or maximum likelihood solution.

$$l(g|x) \propto \exp\left(\sum_{i=1}^{N^2} \lambda_i x_i\right)$$

where

$$\lambda_i = \frac{2g_i - 1}{2\sigma^2}.$$

It is prudent to remember that the likelihood as a function of the unknowns x is an object of quite a different nature to a probability distribution, hence the tempting interpretation that 'most likely' is related to quality of reconstruction is actually misleading.

We specify a prior distribution by modelling x on the pixel lattice as an Ising Markov random field, with distribution

$$\pi(x) \propto \exp\left(\theta \sum_{i=1}^{N^2} \sum_{j \sim i} \delta_{x_i, x_j}\right)$$

where the sum over $j \sim i$ is a sum over pixel neighbors, θ is a smoothing parameter, and $\delta_{a,b}$ is the indicator function for the event $a = b$. For the binary images defined here, the prior is explicitly

$$\pi(x) \propto \exp\left(-\theta \sum_{i=1}^{N^2} \sum_{j \sim i} (2x_i - 1)(2x_j - 1)\right)$$

which is a Gaussian distribution on the binary variables.

The joint posterior distribution for an image x given measurements g is given by Bayes' rule as

$$\begin{aligned} \pi(x|g) &\propto l(x|g)\,\pi(x) \\ &\propto \exp\left(\sum_{i=1}^{N^2} \lambda_i x_i - \theta \sum_{i=1}^{N^2} \sum_{j \sim i} (2x_i - 1)(2x_j - 1)\right) \quad (5) \end{aligned}$$

which is also Gaussian. It is this distribution that we need to explore to learn about the unknown image.

Figure 3 shows four statistics that can be used to summarize the posterior distribution, for a range of smoothing parameters θ. The first column shows the MAP state, which maximizes the posterior distribution

$$x_{\text{MAP}} = \arg\max_{x \in X} \pi(x|g).$$

Inspection of equation 5 shows that MAP state is also the solution found by total-variation regularized inversion, or since the image is binary also by gradient regularization. Note that as the smoothing parameter θ increases, x_{MAP} becomes smoother, providing possible reconstructions at $\theta \approx 0.25$, first loosing the center of the A, then the "legs" and finally, for all θ greater than some critical value in the range $(0.5, 0.675)$, $x_{\text{MAP}} = \mathbf{1}$, i.e. the all-white state.

The second column shows the (posterior) mean image

$$\bar{x} = \mathrm{E}_\pi[x] = \sum_{x \in X} x\pi(x|g).$$

While the mean image is gray-scale and not a feasible reconstruction for binary images, it does give a good indication of the position of the bulk of posterior probability mass. As can be seen, the posterior clusters increasingly around better reconstructions as θ increases, with good reconstructions for $\theta \gtrsim 0.5$.

The third column shows a sample drawn from the posterior, and confirms that the the mean is representative of 'typical' states in the posterior distribution.

The fourth column is the marginal posterior mode (MPM), which shows each pixel as the mode of the marginal distribution of that pixel, and hence takes the value that the pixel most frequently took in the samples drawn from the posterior. For the case of binary images, the MPM is just the thresholded mean, $x_{\text{MPM}} = [\bar{x}]$.

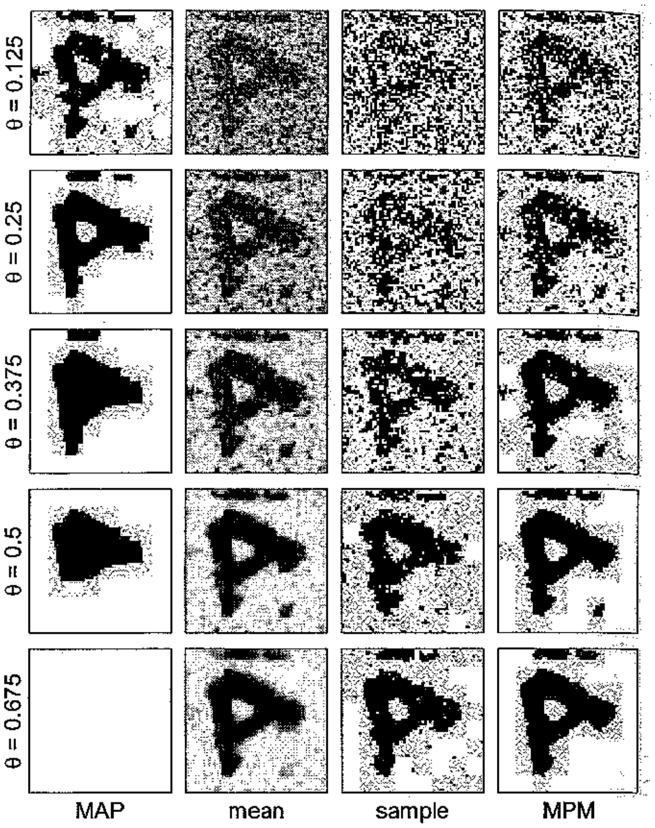

Figure 3. Tableau of maximum *a posteriori* (MAP) state, mean, a single sample from the posterior, and the marginal posterior mode (MPM) for a range of smoothing parameters θ.

This simple, yet instructive, example shows that the MAP state is not a robust estimate of the true image, and does a poor job of summarizing the posterior distribution at *all* values of θ. At larger smoothing parameters, $\theta \gtrsim 0.5$ for this example, when the prior distribution is doing an excellent job of shaping the posterior so that the bulk of posterior probability mass contains smooth images that fit the data well and themselves make good reconstructions, the MAP state is a hopeless reconstruction precisely because it is entirely unrepresentative of typical samples. For $\theta \gtrsim 0.675$ the MAP state is an extreme outlier while the posterior is dominated by states from which a good recovered image could be formed.

In this example, even a single sample drawn from the posterior distribution would be a better estimate of the true image than given by the MAP state (or regularized inversion). This is a typical situation in many inverse problems, and shows that improvement over current methods does not require extensive sampling. Single samples can provide good reconstructions. A few samples, say 2-4, can establish a scale and nature of ambiguity in the reconstructions (see e.g.McKeague et al. [2005]), while extensive sampling allows accurate estimates of posterior variability in applications where that is needed.

Our attention now moves to *how* inferential solutions are computed, and some recent advances in sampling algorithms for inverse problems.

COMPUTATIONAL INFERENCE: MCMC

Answers to questions about the true image may be calculated as posterior expectations of some function f

$$\mathsf{E}[f(\cdot)] = \sum_{x \in \mathsf{X}} \pi(x|d) f(x). \qquad (6)$$

For example, if f is an indicator function for an image showing that the patient has cancer, then equation 6 gives the probability of cancer based on the measurements and the prior information.

The expectation in equation 6 may be calculated using Monte Carlo integration as follows. If $\{X(t), t = 1, 2, \dots, N\}$ are distributed according to the posterior distribution, $\pi(\cdot|d)$, then

$$\mathsf{E}[f(\cdot)] \approx \frac{1}{N} \sum_{t=1}^{N} f(X(t)). \qquad (7)$$

The task then is to draw samples from the given posterior $\pi(\cdot|d)$. The Markov chain Monte Carlo (MCMC) algorithm achieves sampling by generating $\{X(t)\}_{t=0}^{\infty}$ as a Markov chain of random variables $X(t) \in \mathsf{X}$, with a t-step distribution $\mathsf{P}(X(t) = x | X_0 = x^{(0)})$ that tends to $\pi(x|d)$, as $t \to \infty$. Thus the algorithm produces a random walk through the space of feasible images with the long-term probability that the walk will visit a particular image x tending to the desired posterior distribution.

The Metropolis-Hastings Algorithm

In the following we write the abbreviated $\pi(x)$ for $\pi(x|d)$. We construct a Markov chain that evolves on X that has $\pi(\cdot)$ as its limiting distribution by repeatedly applying the transition law P defined by

$$\mathsf{P}\{X(t+1) = x | X(t) = y\} = P(x,y), \quad \forall x, y \in \mathsf{X}.$$

To ensure that π is the equilibrium distribution it is necessary to satisfy the linear *invariance equation:*

$$\pi P = \pi,$$

where π is seen as a row vector, P as a matrix. Hence

$$\sum_{x \in \mathsf{X}} \pi(x) P(x,y) = \pi(y), \quad \forall y \in \mathsf{X}.$$

In practice it is easier to assert the stronger *detailed balance condition:*

$$\pi(x) P(x,y) = \pi(y) P(y,x) \quad \forall x, y \in \mathsf{X},$$

which implies the invariance equation as can be seen by summing over x, and using $\sum_{x \in \mathsf{X}} P(x,y) = 1, \forall y \in \mathsf{X}$.

Detailed balance can be achieved by *simulating* the transition law using the Metropolis-Hastings algorithm, given in algorithm 1. Note that this algorithm depends only on the *ratio* of densities, and hence the normalization constant is not required (but it is important that it exists). The stochastic update is then: if the chain is in state x at step t we set $X(t+1) = \mathrm{MH}(x)$.

```
$y = \mathrm{MH}(x)$
Draw $x'$ from proposal distribution $T(x,x')$
Draw $u$ from Uniform$[0,1]$
$\alpha(x,x') \leftarrow \min\left(1, \frac{\pi(x')T(x',x)}{\pi(x)T(x,x')}\right)$
if $u \leq \alpha(x,x')$
  $y = x'$
else
  $y = x$
```

Algorithm 1: Metropolis-Hastings algorithm simulating operation by the transition kernel $P(x,y)$. The proposal distribution $T(x,x')$ may be *any* distribution that guarantees the chain is aperiodic and irreducible.

A beautiful feature of the MH (or MH-type) algorithm is that it is easy to implement, and the conditions on the proposal distribution are straightforward to achieve. It also turns out to be robust to numerical roundoff error.

Efficiency of this algorithm is strongly dependent on the proposal distribution. The simplest proposal distributions are given by random walks centered at the current state x. However, these lead to very inefficient algorithms in high-dimensional inverse problems. While improved generic proposal distributions are available, such as the Langevin proposal and the hybrid Monte Carlo schemes, the most efficient algorithms result from proposals designed with specific knowledge of the structure of the forward map. These are typically designed using multiple 'moves', where each move explores a known uncertainty in solutions. An example of such a proposal distribution, tailored to a specific inverse problem, is given by Nicholls and Fox [1998].

Computing Inferential Solutions to Inverse Problems

In principle, the posterior density in equation 3 or 4 can be evaluated and hence sampled using Metropolis-Hastings dynamics allowing summary statistics to be evaluated, effectively solving the inverse problem. However, the need to calculate the posterior density at each step in a standard Metropolis–Hastings algorithm, with typically many thousands or millions of steps required to give sufficiently small variance in estimates, appears to be computationally prohibitive for realistic inverse problems. In this aspect, computational MCMC for inverse problems shares many of the goals and problems of numerical optimization.

Nevertheless, there are a growing number of demonstrations of comprehensive posterior sampling, conditioned on measured data, for inverse problems implementing a physically-based forward simulator. Recent examples include the work by McKeague et al. [2005]) mapping ocean circulation, Haario et al. [2004] recovering atmospheric gas density, Cornford et al. [2004] who retrieve fields of wind vectors, and Cui [2005] calibrating numerical models of geothermal fields.

The massive scale of computation in each of these examples indicates that considerable improvement in efficiency of MCMC algorithms for inverse problems is required if the method is to be widely applied. Indeed, each of the works cited employs an enhanced MCMC to improve computational efficiency. For example Haario et al. [2004] used a novel adaptive Metropolis algorithm in which the covariance matrix in a d-dimensional Gaussian proposal distribution is calculated from the history of the output chain. Another interesting development is the Metropolis coupled MCMC of Higdon et al. [2003] that simultaneously runs chains with the spatial parameters coarsened to various degrees. Information from the faster running, though approximate, coarse formulations speeds up mixing in the finest scale chain, from which samples are taken.

Simulated Tempering

Consider the case where the Metropolis-Hastings algorithm is used to sample from posterior distribution $\pi(\cdot)$ using proposal distribution $T(x,y)$ and it is found that the resulting chain is evolving slowly, or worse still is getting stuck. This can happen because of multi-modality of $\pi(\cdot)$, or because of strong correlations as is typical in inverse problems where the support of $\pi(\cdot)$ can be effectively a low-dimensional subspace of X. Simulated tempering (with the name and idea adapted from simulated annealing for optimization) is a general method that can overcome some of these difficulties, while using the existing proposal distribution $T(x,y)$.

The methods augments the state space to $\mathrm{X} \times \{0,1,\ldots,N\}$ and defines a set of distributions $\{\pi_k(\cdot)\}_{k=0}^{N}$ where $\pi_0 = \pi$ and $\pi_1(\cdot), \pi_2(\cdot), \ldots, \pi_N(\cdot)$ are a sequence of distributions that are increasingly easy to sample from. The distribution over the augmented space is taken as

$$\pi(x,k) \propto \pi_k(x)$$

with transitions for a fixed k being derived from the proposal $T(x,y)$ and are interspersed with proposals that change k (perhaps by a random walk in k) with both accepted/rejected by a standard Metropolis-Hastings algorithm. The random walk then occurs in (x,k) space. Samples that have $k=0$, i.e. from the conditional density $\pi(x,k|k=0)$, are samples from the desired distribution.

A simple example of such a sequence is the scheme due to Marinari and Parisi [1992] who introduced simulated tempering. Define the positive numbers (inverse temperatures) $1 = \beta_0 < \beta_1 < \cdots < \beta_N$ and pseudo prior constants $\lambda_0, \lambda_1, \cdots, \lambda_N$ with $\sum_{k=1}^{N} \lambda_k = 1$. The sequence of distributions is then given by

$$\pi_k(x) = \lambda_k \pi^{\beta_k}(x)$$

which are increasingly unimodal.

Other schemes for generating sequences of distributions have found greater success in inverse problem applications. The simple idea of increasing the variance used to calculate the likelihood function was used by Palm [1999] to overcome sampling difficulties arising in electrical conductance imaging where high-accuracy data leads to posterior distributions that are too narrow to easily sample. Sampling difficulties due to a very narrow posterior distributions also arises when the data consists of a long time series, such as is measured in ultrasound imaging or from observations of a dynamical system. Then an effective tempering scheme can be to simply reduce the length of data considered. So if

$$\{d\} = \{d_0\} \supset \{d_1\} \supset \cdots \supset \{d_N\}$$

is a sequence of decreasing data sets, we define

$$\pi_k(x) = \pi(x|d_k).$$

Parallel tempering is similar to simulated tempering except that the N chains (one for each value of k) are maintained simultaneously. An example is the Metropolis coupled MCMC of Higdon et al. [2003], mentioned above, that simultaneously runs chains with the spatial parameters coarsened to various degrees. The recent 'evolutionary Monte Carlo' algorithms introduced by Liang and Wong [2001] are examples of parallel tempering with moves that are inspired by the genetic optimization algorithms, amongst other ideas.

Parallel Rejection Algorithm

We now consider an algorithm to decrease the CPU time per MCMC update by utilizing multiple compute processors using a straightforward parallelizing of the serial Metropolis-Hastings algorithm[3].

Consider the serial Markov chain $\{X(t)\}_{t=0}^{\infty}$ that is at state $X(t) = x(t)$ for some t. Suppose that n processors are available, that we take to be independent for computing purposes. We run on each processor an *independent* instance of the Metropolis-Hastings algorithm initialized at state $x(t)$ to give the n independent Markov chains $\{Y(r,k)\}_{k=0}^{\infty}$ for $r = 1,2,\ldots,n$ with $Y(r,0) = x(t)$. We enumerate the resulting states by $s(r,k) = r + n(k-1)$ for $r = 1,2,\ldots,n$ and

[3] This idea is due to Geoff Nicholls.

$k = 1,2,\ldots$ giving the total ordering $s = 1,2,\ldots$. Note that s is an invertible function so we may refer to states on the parallel computer by the total ordering s, i.e. as $\{Y(s)\}_{s=0}^{\infty}$. We run the n parallel chains until the first non-trivial acceptance (in the order s) occurring at $s_{\min}$, i.e. the minimum s for which $Y(s) \neq x(t)$. Then set $X(j) = x(t)$ for $j = t+1, t+2, \ldots, t+s_{\min}-1$ and $X(t+s_{\min}) = Y(s_{\min})$, and reinitialize the n parallel chains.

The block stochastic update given by parallel rejection is shown in algorithm 2. The update returns a variable number of states that are appended to the existing chain.

```
$\underline{y(1), y(2), \ldots, y(s_{\min}) = \mathrm{PR}(x)}$
Initiate $n$ parallel chains $\{Y(r,k)\}_{k=0}^{\infty}$ for $r = 1,2,\ldots,n$
  with $Y_{r,0} = x$
Run until there is some $s_{\min}$ with $Y(s_{\min}) \neq x$ AND
  $Y(s) = x \; \forall s < s_{\min}$ in the ordering $s(r,k) = r + n(k-1)$
Set $y(1), y(2), \ldots, y(s_{\min}-1) = x$
Set $y(s_{\min}) = Y(s_{\min})$
```

Algorithm 2: Parallel rejection algorithm for updating a variable-length block. In practice each of the parallel chains can be halted after it has an acceptance, or when it is performing an iteration with index s greater than an acceptance in another chain.

In the simplest case where time per MCMC step is constant, the speedup is achieved because the serial chain is advanced m steps in time proportional to $\lfloor m/n \rfloor + 1$ rather than time proportional to m. A more accurate calculation taking into account the acceptance rate α and time for transactions relative to the forward map β shows that the expected speedup is

$$\frac{(1-(1-\alpha)^n)}{\alpha}\frac{1}{1+n\beta}.$$

In practice the transaction time limits the speedup achieved, and is therefore critical in implementations. Parallel rejection was first used to speed up calibration of numerical models of geothermal fields by Cui [2005].

Using Approximations

The simulated- and parallel-tempering algorithms mentioned above may be thought of as using approximations to the posterior distribution as a means of improving sampling efficiency. Two quite different algorithms have been recently introduced that explicitly use approximations to the forward map, with both achieving substantial reductions in time required to compute inferential solutions to inverse problems.

As a means of model reduction (and counteracting inverse crimes) Kaipio and Somersalo [2005] introduced the enhanced error model to correct for model errors introduced by coarse numerical approximations. For the case of Gaussian prior and noise distributions, they considered the accurate linear model

$$d = Ax + n$$

and the coarse approximation

$$d = \tilde{A}\tilde{x} + \tilde{n}$$

where $\tilde{x} = Px$ is a coarse approximation to the unknowns x resulting from a projection by P, and $\tilde{A}$ is the (cheap) approximation to A on coarse variables. Then

$$\tilde{n} = (A - \tilde{A}P)x + n$$

defines the enhanced error model by assuming the two terms on the right hand side are uncorrelated. Use of the coarse approximation necessarily increases the uncertainty of recovered values, since model error has been introduced. However, Kaipio and Somersalo [2005] give examples in which a tolerably small increase in posterior uncertainty is traded for a huge reduction in compute time without introducing bias in estimates, and demonstrate that accurate real-time inversion is possible.

A second use of approximations was introduced by Christen and Fox [2005] who considered the state-dependent approximation $\pi_x^*(\cdot)$ to the posterior distribution calculated using a cheap approximation to the forward map, to give a modified Metropolis-Hastings MCMC. Once a proposal is generated from the proposal distribution $T(x,y)$, to avoid calculating $\pi(y)$ for proposals that are rejected, they first evaluate the proposal using the approximation $\pi_x^*(y)$ to create a second proposal distribution $T^*(x,y)$ that is then used in a standard Metropolis-Hastings algorithm. The full definition is given in algorithm 3. An appealing feature of algorithm 3 is that computer implementation can make use of the same problem-specific functions required for a gradient-based optimization, when the approximation is based on a local linearization of the forward map. Christen and Fox [2005] present an example using a local linearization, and demonstrate an order of magnitude speedup in a stylized version of electrical impedance imaging.

CONCLUSIONS

Computational inference for inverse problems is currently a rapidly developing area – as can be seen by the recent advances occurring in the past few years. These advances have allowed the computation of inferential solutions to substantial inverse problems, including examples with many thousands of unknowns recovered from measured data and using complex physical simulations of the forward map.

```
$y = \mathrm{DAMH}(x)$
Draw $x'$ from proposal distribution $T(x,x')$
Draw $u_1$ from Uniform$[0,1]$
$\alpha(x,x') \leftarrow \min\left(1, \frac{\pi_x^*(x')T(x',x)}{\pi_x^*(x)T(x,x')}\right)$
if $u_1 \leq \alpha(x,x')$
  Promote $x'$ and go on
else
  $y = x$, i.e. reject $x'$ and exit
Define $T^*(x,y) = \alpha(x,y)T(x,y)$
Draw $u_2$ from Uniform$[0,1]$
$\beta(x,x') \leftarrow \min\left(1, \frac{\pi(x')T^*(x',x)}{\pi(x)T^*(x,x')}\right)$
if $u_2 \leq \beta(x,x')$
  $y = x'$
else
  $y = x$
```

Algorithm 3: Modified Metropolis-Hastings algorithm that makes use of an approximation.

It is notable that all the advances reported here are enhancements of the basic Metropolis-Hastings algorithm. Since that algorithm implements a random-walk (or a diffusion when averaged) over state space, it is necessarily slow to explore all feasible solutions. It is likely that the future will see new algorithms that circumvent the need for detailed balance, and hence need much reduced sampling times. Such algorithms could potentially reduce the computational cost of sampling to the current cost of optimization, and thereby make inferential solutions the method of choice across the field of inverse problems.

ACKNOWLEDGMENT

The author is indebted to Tiangang Cui (Auckland), Andrés Christen (CIMAT), Jari Kaipio (Kuopio), and Geoff Nicholls (Oxford), for sharing their insights during the course of this research.

REFERENCES

Andersen, K., S. Brooks, and M. Hansen (2003). Bayesian inversion of geoelectrical resistivity data. *Journal of the Royal Statistical Society Ser. B 65*, 619–642.

Christen, J. A. and C. Fox (2005). MCMC using an approximation. *Journal of Computational and Graphical Statistics 14*, 795–810.

Cornford, D., L. Csato, D. Evans, and M. Opper (2004). Bayesian analysis of the scatterometer wind retrieval inverse problem: Some new approaches. *Journal of the Royal Statistical Society Ser. B 66*, 1–17.

Cui, T. (2005). Bayesian inference for geothermal model calibration. Master's thesis.

Fox, C. and G. Nicholls (2000). Exact map states and expectations from perfect sampling: Greig, Porteous and Seheult revisited. In A.-M. Djafari (Ed.), *Twentieth Int. Workshop on Bayesian Inference and Maximum Entropy Methods in Sci. and Eng.* AIP.

Geman, S. and D. Geman (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Analysis and Machine Intelligence 6*, 721–741.

Grenander, U. and M. Miller (1994). Representations of knowledge in complex systems. *Journal of the Royal Statistical Society Ser. B 56*, 549–603.

Haario, H., M. Laine, M. Lehtinen, E. Saksman, and J. Tamminen (2004). Markov chain Monte Carlo methods for high dimensional inversion in remote sensing. *Journal of the Royal Statistical Society Ser. B 66*(3), 591–608.

Higdon, D., H. Lee, and C. Holloman (2003). Markov chain Monte Carlo-based approaches for inference in computationally intensive inverse problems. In J. M. Bernardo, M. J. Bayarri, J. O. Berger, D. H. A. P. Dawid, A. F. M. Smith, and M. West (Eds.), *Bayesian Statistics 7*. Oxford University Press.

Kaipio, J. and E. Somersalo (2005). *Statistical and Computational Inverse Problems*. Number 160 in Applied Mathematics. Springer.

Kumar, S. and M. Hebert (2006). Discriminative random fields. *International Journal of Computer Vision 68*(2), 179–201.

Liang, F. and W. Wong (2001). Real-parameter evolutionary Monte Carlo with applications to Bayesian mixture models. *Journal of the American Statistical Association 96*(454), 653–666.

Marinari, E. and G. Parisi (1992). Simulated tempering: A new Monte Carlo scheme. *Europhysics Letters 19*, 451–458.

McKeague, I. W., G. Nicholls, K. Speer, and R. Herbei (2005). Statistical inversion of south atlantic circulation in an abyssal neutral density layer. *Journal of Marine Research 63*, 683–704.

Nicholls, G. K. (1998). Bayesian image analysis with Markov chain Monte Carlo and coloured continuum triangulation mosaics. *Journal of the Royal Statistical Society Ser. B 60*, 643–659.

Nicholls, G. K. and C. Fox (1998). Prior modelling and posterior sampling in impedance imaging. In A. Mohammad-Djafari (Ed.), *Bayesian Inference for Inverse Problems*, Volume 3459, pp. 116–127. SPIE.

Palm, M. (1999). Monte carlo methods in electrical conductance imaging. Master's thesis.

Watzenig, D. (2006, June). *Bayesian inference for process tomography from measured electrical capacitance data.* Ph. D. thesis.