Monte Carlo (stochastic) inversions - similarities and differences to gradient descent methods.

Ross C. Brodie

Introduction

Geophysical inversion is the process of estimating the parameters of Earth model(s) that are consistent with a set of geophysical observations, taking into account the noise on the data. There are two main viewpoints from which this can be approached, sometimes referred to as deterministic and probabilistic. There are also two common mechanisms for obtaining the solution which are known as gradient descent and direct search methods. This abstract aims to demonstrate some similarities and differences between these approaches using simple examples from the 1D inversion of airborne electromagnetic data.

Deterministic and probabilistic

Inversion problems usually require the definition of an objective (cost or error) function that expresses a combination of degree of fit (closeness or agreement) between the theoretical forward response of an Earth model and the observed data (e.g., least squares data misfit) and degree of agreement between the Earth model and prior information (e.g., closeness to a reference value). The objective function may be exactly the same in both deterministic and probabilistic approaches.

In the deterministic approach a single Earth model, which according to the objective function is optimum, is found and judged to be the solution to the inverse problem. The deterministic approach will usually be concerned with finding the global extrema of the objective function, a process known as global optimisation. This approach is thus often described as an optimisation approach or problem.

In contrast, the probabilistic approach aims, not to settle upon a single model, but to gather statistics about all models that are feasible after consideration of both data fit and the prior information. This will typically be an ensemble (collection) of those models in the vicinity of the global extrema and possibly several local extrema of the objective function. This approach is often called a probabilistic, ensemble or Bayesian inference approach.

In probabilistic inference all information is expressed in terms of probability density functions (PDFs). Figure 1 shows the concept behind Bayes' rule which, more or less, says that the posterior PDF (what is considered probable after the data is collected) is proportional to the product of the prior PDF (what was considered probable before the data was collected) times the likelihood function (the likelihood of the data being observed given the model and noise).

Probabilistic inference methods are concerned with making good estimates of the Posterior PDF histogram shown in red on the right of Figure 1. In contrast, deterministic methods are only concerned with locating one point on the horizontal axis, perhaps for example the position of the peak of the posterior PDF if prior information is being used. Or alternatively, if prior information is not available (constant prior PDF), the position of the peak of the likelihood functions.

Figure 1. Demonstration of the concepts behind Bayes' rule. Reproduced with permission from Sambridge (2013).

The terms Monte Carlo and stochastic are sometimes rather loosely used to mean probabilistic. This comes about because probabilistic inversions are more often than not implemented via Monte Carlo methods. Monte Carlo is only prefix indicating that the method or approach makes use of repeated trials, or sampling, generated with the use of random numbers, and is named after the famous French city associated with casinos (Sambridge, and Mosegaard, 2002). Monte Carlo methods may, in fact, be used in either deterministic or probabilistic approaches.

Gradient descent and direct search

Gradient descent (or downhill, or gradient based) methods make use of the gradient of the objective function surface to traverse parameter space in a downhill direction from a starting model to toward a local or global extrema. They are always used for deterministic inversion. These methods are widely used for deterministic linear (e.g., Li and Oldenburg, 1996) and non-linear (e.g., Constable et al., 1987) geophysical inverse problems.

Monte Carlo methods fall into the broader category of direct search methods which do not require calculation of derivatives of the objective function. Uniform and grid searches are examples of non-Monte-Carlo direct search methods. Simulated Annealing (Rothman, 1985) and Genetic Algorithms (Sen and Stoffa, 1992) are the most well-known direct search deterministic Monte Carlo methods used for global optimisation problems.

In the probabilistic approach the emphasis is less on optimisation and more on sampling the most probable regions of parameter space as determined by the posterior PDF (Sambridge, and Mosegaard, 2002). The idea is that the sampling density is proportional to the height of the posterior PDF and is known as importance sampling. The Metropolis-Hastings algorithm and the Gibbs Sampler are two so-called Markov chain Monte Carlo (McMC) algorithms (Hastings, 1970) that are designed to generate samples of an initially unknown PDF and are the favoured tools for importance sampling in probabilistic geophysical inference. The Neighbourhood Algorithm (Sambridge, 1999a and 1999b) is another form of Monte Carlo probabilistic inference.

A relatively new class of Monte Carlo methods are the transdimensional or reversible-jump (rj-McMC) algorithms (e.g., Green, 1995; Malinverno, 2002; Bodin and Sambridge, 2009). These allow the number of unknowns (dimensionality) to be varied as the sampling progresses. Hence the generated ensemble consists of models with different numbers of parameters. This allows great flexibility in the models and the possibility making inferences about the number of parameters (e.g., layers) required to explain the data.

Contrasting the methods

Gradient descent methods require that the derivative of the objective function be either analytically or numerically differentiable, with respect to the model parameters. Monte Carlo (and other direct search methods) do not require derivatives and are thus more flexible and can be applied to problems having complex model-data relationships and sophisticated objective functions.

Gradient descent methods are efficient at locating local minima of the objective function near the starting position, but because they always traverse downhill they are prone to getting trapped in local minima in the vicinity of the starting model and may never find the global minimum at all. They are thus better suited to problems with simple linear or weakly non-linear (Figure 2a) model-data relationships, rather than highly non-linear problems with multiple minima (Figure 2b). On the other hand, having the element of randomness, Monte Carlo methods have the (unguaranteed)

ability to climb out of local minima and have a greater chance of locating the global minimum. They are thus more suitable for strongly non-linear global optimisation problems.

Figure 2. Examples of (a) weakly non-linear and (b) strongly non-linear objective functions. Reproduced with permission from Sambridge (2013).

Gradient descent methods tend to be orders of magnitude more computationally efficient than direct search global optimisation methods because they do not rely on trial and error but are informed by the extra information of the gradients. The probabilistic Monte-Carlo methods are even more computationally expensive than deterministic Monte Carlo since they must sample an ensemble of models rather than seek a single minimum.

Gradient descent methods are feasible for as many as millions of unknown model parameters and are thus suitable for large scale inversions, such as voxel based potential field inversion (Li and Oldenburg, 1996). In contrast Monte Carlo methods are usually only feasible for a few, tens to perhaps hundreds of parameters. This is due to the so called curse of dimensionality because the number of models that need to be tested goes up with the power of the number of parameters.

Example

In this section dual moment SkyTem-508 airborne electromagnetic (AEM) data are inverted using a gradient descent deterministic method and a probabilistic rj-McMC method. The data are from a costal groundwater investigation survey and were supplied courtesy of the Western Australian Department of Water. In both cases every AEM sounding was inverted independently to a 1D conductivity-depth model, and were then stitched together to make conductivity sections.

The deterministic gradient descent inversion was developed at Geoscience Australia and is based on the Occam's inversion (Constable et al., 1987). The parameterization was a 30 layer model with fixed layer thicknesses. The objective function,

$$\Phi = \Phi_d + \lambda \Phi_c = \Phi_d + \lambda (\Phi_c + \alpha \Phi_s)$$

consists of the noise normalised least squares data misfit (Φ_d) and a model regularization term (Φ_m) weighted by a regularization parameter (λ). The model regularization term consists of the L_2 distance (Φ_c) between the inversion model and a conductivity reference model, and a second finite difference model roughness term (Φ_s) weighted by the smoothness parameter (α).

The non-linear inversion iteratively reduces Φ_d until and acceptable level of data misfit value (Φ_d = 1) is found. Within each iteration a line search must be performed on λ to select a suitable value that will result in a slow reduction of the data misfit. The selection of the smoothness parameter (α) is a subjective user supplied choice. Figure 3 shows the inversion of one flight line of the data using reference model conductivity of 0.001 S/m and with three different values of the smoothness parameter of 1 (panel b), 1,000 (panel c) and 1000,000 (panel d). Clearly the section with the low smoothness value is not plausible. It has put too much emphasis on the reference model at the expense of the model smoothness. It has also prevented the convergence to an acceptable misfit (see panel a), probably because of being trapped in local minima. In contrast both higher smoothness values have allowed the inversion to always converge and have resulted in geologically plausible models.

Figure 3. Comparison of the use of three different smoothness parameters for the deterministic inversion.

Figure 4 shows the inversion of the same flight line using a smoothness parameter of 1,000 and a conductivity reference model of 0.001 S/m (panel b) and 0.050 S/m (panel c). In both cases the data are always fitted and the top portions of the sections are distinctly similar. The difference lies in the conductivity toward the bottom of the sections. In both cases the data is too insensitive to the conductivity at depth and is unable to influence the conductivity. Accordingly the inversion model conductivity at depth (especially beneath the saline lake at distance 3,000 m) stay more or less clamped at the respective reference model values. This is sometime used as a method of estimating the depth of investigation.

Figure 4. Comparison of the use of two different conductivity reference models for the deterministic inversion.

The rj-McMC inversion program was developed at Geoscience Australia. It is adapted from the work of Bodin and Sambridge (2009). The algorithm used for this particular inversion was described by Brodie and Sambridge (2012). Minsley (2011) has also described a similar algorithm for inversion of frequency-domain AEM data. For this survey the model domain was allowed to have up to 10 layers with uniform prior probability. Layer interfaces were allowed to lie between the surface and 100 m depth with uniform probability. The conductivity of any particular layer had log-uniform prior probability in the range 0.001 S/m to 10 S/m.

The rj-McMC sampling begins by initializing the Markov chain with a single-layer model chosen uniform randomly from the prior. In the sampling loop new model are proposed for possible addition to the ensemble. One of four types of changes to the current model may be proposed: (i) change a layer's conductivity, (ii) move an interface, (iii) add an interface, or (iv) delete an interface. The proposed model is either accepted or rejected. If accepted it is appended to the Markov chain, otherwise, a copy of the current model is appended. A total of 300,000 models were generated with a 10,000 sample burn-in period to allow time to converge to a suitably low data misfit. Once the ensemble is generated various information are extracted.

Figure 5 shows a comparison between the rj-McMC and the deterministic inversion conductivity sections. Figures 6, 7 and 8 show details of the results at the individual AEM soundings at the positions marked A, B and C on Figure 5.

A 2D discrete frequency histogram representing the posterior PDF is generated from the ensemble. These are depicted by the grey shading on Figures 6b, 7b and 8c, where darker areas represent more probable regions of the posterior PDF. From the 2D histogram various other metrics are also extracted including the mean, mode, 10^{th}, 50^{th} and 90^{th} percentile models. This is done by calculating the statistics from each row (depth bin) of the 2D PDF histogram. Also the 'best' individual model with the largest posterior PDF in the ensemble is stored.

The 10^{th} and 90^{th} percentile models are superimposed on Figures 6b, 7b and 8c. The deterministic inversion models, for the same AEM sounding, but two different conductivity reference models are also shown for comparison.

Figure 5a shows the best and mean data-misfit of the models in the rj-McMC ensemble. They are generally <1 indicating the algorithm is predominantly sampling models in the data acceptable region. Figure 5b shows the same deterministic conductivity section as Figure 3d. Figure 5c and d

show the conductivity section representing the best and mean models extracted from the rj-McMC ensemble as discussed above.

All three sections show the same broad conductivity structure. Without additional information there is no objective way of determining that one is any better than any other. The main differences lie in the sharpness of the apparent boundaries between layers. This is no doubt partly due to the large vertical smoothing parameter of 1,000,000 that was applied to the deterministic inversion shown in Figure 5b to prevent too much oscillation in the solution.

Another piece of information that can be extracted is the changepoint histogram. This is a 1D histogram of the position of the interface depths discretised into 1 m bins (see Figure 6c, 7c and 8c). The peak position of the changpoint histogram is shown as black dots on Figure 5c and d. It suggests the possibility of being able to be used as an automated interface picker if multiple peaks could be conveniently extracted,

The difference or spread between the 10^{th} and 90^{th} percentile models can also be considered as a measure of the uncertainty at that particular depth. To convey this uncertainty to the interpreter we use progressively increase the transparency of the conductivity section as the spread increases above a certain threshold value. This is clearly evident in the vicinity of the saline lake at the position Marked A on the sections. Figure 6b shows that at this point the spread is large at this point.

Figures 6d, 7d and 8d show the histograms of the number of models in the ensemble. It is not understood why the results for position A (Figure 6d) are showing a high probability of 10 layers given the limited depth of investigation. However at position B (Figure 7d) and C (Figure 8d) are clearly showing a high probability of 3 layer models.

To invert the 11,228 soundings in the complete dataset, the deterministic gradient descent inversion took 15.63 CPU hours (5 s/sounding). However the rj-McMC inversion was 345 times more expensive, costing 5,399 CPU hours (1,730 s/sounding).

Figure 5 A comparisons of conductivity sections for the deterministic gradient descent inversion and the best and mean models from the rj-McMC inversion.

Figure 6 Summary of the rj-McMC inversion results for the sounding marked A on Figure 5. It shows (a) the data misfit for each sample, (b) the 2D histogram of the posterior PDF (grey shading), (c) the 1D changepoint histogram and (d) the histogram of the number of layers.

Figure 7 Same as Figure 6 but for the AEM sounding at position marked B on Figure 5.

Figure 8 Same as Figure 6 but for the AEM sounding at position marked C on Figure 5.

Acknowledgements

I would like to thank the Western Australia Department of Water for permission to use the Lake Thetis SkyTEM dataset for the purpose of testing and demonstrating the application of the rj-McMC inversion.

I also appreciate the assistance from Malcolm Sambridge, Research School of Earth Science, Australian National University for guidance in development of the rj-McMC algorithm and for permission to use course diagrams used herein.

The computational work was carried out on the National Computational Infrastructure (NCI, http://www.nci.org.au). The NCI is supported by the Australian Government through the Department of Innovation, Industry, Science and Research and by a number of major co-investing partner organisations including the Australian National University, CSIRO, the Australian Bureau of Meteorology and Geoscience Australia.

References

Bodin, T., and M. Sambridge, 2009, Seismic tomography with the reversible jump algorithm. Geophysical Journal International 178, 1411-36.

Brodie, R., and M. Sambridge, 2012, Transdimensional Monte Carlo Inversion of AEM Data: 22nd International Geophysical Conference and Exhibition, Australian Society of Exploration Geophysicists, Extended Abstracts.

Constable, S. C., R. L. Parker, and C. G. Constable, 1987, Occam's inversion; a practical algorithm for generating smooth models from electromagnetic sounding data. Geophysics 52, 289-300.

Green, P. J., 1995, Reversible jump MCMC computation and Bayesian model selection. Biometrika 82, 711–32.

Hastings, W. K., 1970, Monte Carlo sampling methods using Markov Chain and their applications, Biometrika, 57, 97–109.

Li, Y. and D. W. Oldenburg, 1996, 3-D inversion of magnetic data. Geophysics 61, 394-408.

Malinverno, A., 2002, Parsimonious Bayesian Markov chain Monte Carlo inversion in a nonlinear geophysical problem. Geophysical Journal International 151, 675-88.

Minsley, B. J., 2011, A trans-dimensional Bayesian Markov chain Monte Carlo algorithm for model assessment using frequency-domain electromagnetic data. Geophysical Journal International 187, 252-72.

Rothman, D. H., 1985, Nonlinear inversion statistical mechanics, and residual statics corrections, Geophysics, 50, 2784–2796.

Sambridge, M., 1999a, Geophysical inversion with a neighborhood algorithm; I, Searching a parameter space. Geophysical Journal International 138, 479-494.

Sambridge, M., 1999b, Geophysical inversion with a neighbourhood algorithm--II. Appraising the ensemble. Geophysical Journal International 138, 727-746.

Sambridge, M., and K. Mosegaard, 2002, Monte Carlo methods in Geophysical inverse problems. Reviews in Geophysics 40, 1-29.

Sambridge, M., 2013, An introduction to Inverse Problems: EMSC8012 Course Notes, Research School of Earth Sciences Australian National University, available from http://rses.anu.edu.au/~malcolm/Teaching/EMSC8012/lectures/EMSC8012.pdf.

Sen, M. K., and P. L. Stoffa, 1992, Rapid sampling of model space using genetic algorithms, Geophysical Journal International, 108, 281–292.