Guidelines for probabilistic inversion

Klaus Mosegaard
Thomas Mejer Hansen
Andrea Zunino

Niels Bohr Institute
University of Copenhagen

Overview

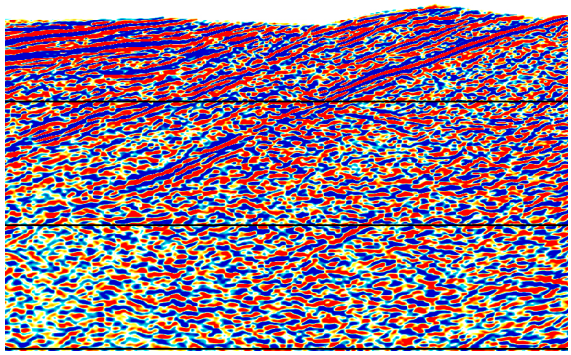
Introduction

The following guidelines are based on two methods:

1. **Careful and conscious preparation of input information for the inversion procedure**. Input parameters and their uncertainties must be specified with great care, using all available information, and with particular caution when such uncertainties are not readily found.
2. **External quality control of the relation between input and output information**. There are three basic ways to test the correctness of an algorithm:
   - Mathematical proof
   - Internal tests
   - **External tests**

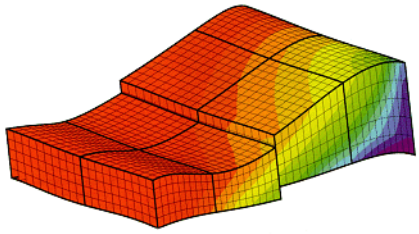
Parameterization of the model

Example: Seismic Inversion



Selection of basic model parameters

1. Choose parameters for which you have prior information.
   - Parameterization in terms of P- or S-wave velocity and density, or reflectivity, is often used, but such parameters are hard to define meaningful prior information for.
   - Parameters such as porosity, sand content or layer thickness are examples of parameters with a more direct geological meaning.
2. If possible, use a uniform grid (in space, and in time) to represent subsurface properties.
   - If two subsurface parameters represent different volumes in space (or intervals of time), your a priori expectations about them may be very different.



Vertical and horizontal gridding

Choosing the Number of Model Parameters

1. Do not underparameterize the model. If you choose too few parameters to represent the target zone you have lost your possibilities for resolving structure that is not allowed by the parameterization.
2. Too few parameters may result in distorted (and even meaningless) probability distributions.
3. Parameterize densely enough to allow even finer structure to be represented, and then carry out a resolution analysis from the final inversion results.

Hard constraints on the model

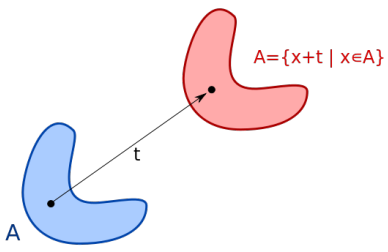
Hard constraints on the model fall in two related categories:

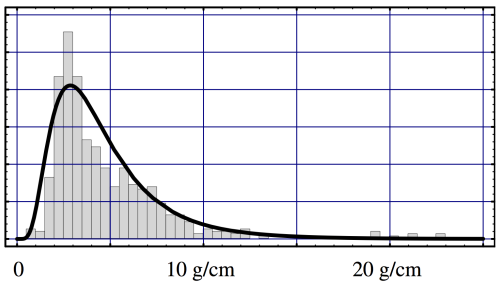
1. **Hard parameter limits**
   - Monte Carlo sampling requires that (possibly unnormalized) probability densities are *integrable*. It is therefore common to work within a limited domain of the model space.
   - Choose the parameter limits in such a way that all physical/geological relevant solutions are allowed. If the true solution is outside the limits, there is usually no way to compute the error from the inversion output.
2. **Fixed parameters**
   - Some parameters are physical constants that are known with great precision, but others may be Earth parameters that *assumed* accurately known. If the actual uncertainty of those parameters are non-negligible, using them as fixed parameters may result in too optimistic (too small) uncertainties on the inversion result.

Probabilistic prior information on the model parameters

There are two reliable sources of prior information in practical inversion:

1. **Prior information from symmetries** Example: We have good reasons to believe that the reflector is located between T_A and T_B, but within that interval we have no reason to favor any value of T over others.
2. **Prior information from external data** Example: Older geophysical or geological data analyzed through probabilistic methods may have resulted in output (posterior) probabilities that can be used as prior probabilities in our analysis of the new data.





Defining the noise model

Besides the definition of prior information, the other main source of input to probabilistic inversion is the noise (uncertainty) in the data:

1. **Measurement noise** An assessment of measurement noise requires information from instrument manufacturers, on-site tests of seismic sources, etc., and all this should, in principle, be correctly combined into a noise model for the data.
2. **Parameterization noise** Even when we carefully parameterize the target zone (as described above), there is bound to be structure in the Earth that is not represented in the model, and this structure will give rise to data that we cannot account for in our modeling. This data will therefore appear as noise.
3. **Modelization noise** The third, and potentially very important, kind of noise is modelization noise which stems from the fact that out forward algorithms (e.g., seismic data simulators) are inaccurate or simplified.

Building the noise model

The above-mentioned sources of noise are very difficult to describe individually, and they are likely to differ from situation to situation. For this reason, we recommend the following approach:

1. Compute synthetic data using from wavelets and the elastic parameters at each well site.
2. Subtract synthetic data from the seismic data to be inverted, to obtain *data residuals* $\mathbf{r}$.
3. Compute the empirical covariance matrix $\mathbf{C} = \{C_{ij}\}$ as:

$$C_{ij} = \frac{1}{N-1} \sum_{n=1}^{N} (r_n - \bar{r})(r_{n+(i-j)} - \bar{r}) \tag{1}$$

   where

$$\bar{r} = \frac{1}{N} \sum_{n=1}^{N} r_n \, .$$

4. Use $\mathbf{C}$ as the noise covariance matrix $\mathbf{C}_D$.

Quality Control

Quality Control of the Inversion Process

Testing Strategy

The underlying strategy of any external, algorithmic testing scheme is to use test functions (with outputs "Pass" or "Fail") satisfying that

1. They can be computed with little additional cost.
2. The result "Pass" is a necessary condition for the algorithm to work correctly (but unfortunately not a satisfactory condition - that would require that we checked the entire algorithm in detail).
3. The result "Pass" means that there is a high probability that the algorithm works correctly.

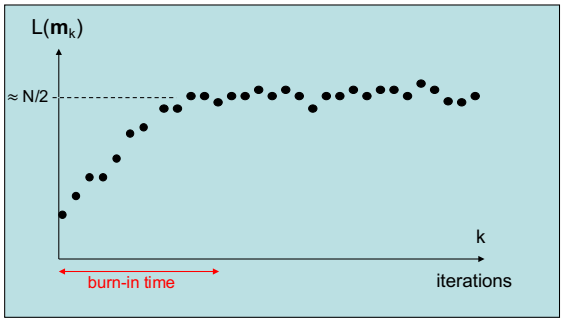
Practical testing of the performance of a Monte Carlo algoritm

Two properties of a Monte Carlo sampler should be tested to ensure proper performance of the algorithm:

1. How many statistically near-independent samples is it producing?
2. Is this number sufficient to characterize the posterior probability distribution for our purposes?

Testing the statistical independence of the samples

Quality control of a Monte Carlo sampling:



- Test that the log-likelihoods

$$\log(L(\mathbf{m}_k)) = -\frac{1}{2}(\mathbf{d}_{\text{obs}} - \mathbf{g}(\mathbf{m}_k)^T \mathbf{C}_D^{-1}(\mathbf{d}_{\text{obs}} - \mathbf{g}(\mathbf{m}_k)$$

  for consecutive models $\mathbf{m}_k, \mathbf{m}_{k+1}, \ldots$ have a mean value of the order of $N/2$, where N is the number of data values.

($-2\log(L(\mathbf{m}_n))$ is χ^2-distributed with mean N and variance $2N$, and hence the distribution of $-\log(L(\mathbf{m}_n))$ will have mean value $N/2$ and standard deviation $\sqrt{N/2}$.)

Test of Independence

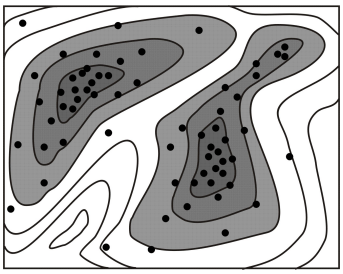
The number of iterations I_w between (near-)independent models can be computed as follows:

1. If f_n is the values of the same model parameter taken from a string of consecutive models $\mathbf{m}_k, \mathbf{m}_{k+1}, \ldots$, the waiting time I_w between independent samples can be estimated from the autocorrelation

$$a(k) = \sum_n f_n f_{n+k}$$

2. If the total number of iterations is I, we only have about I/I_w independent samples.

Ensemble Test



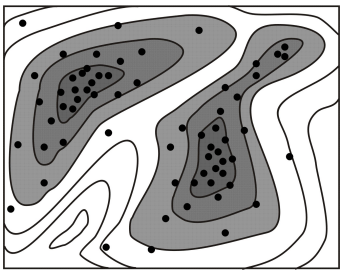
Correct sampling of the posterior can further be assured by an *ensemble test*:

1. Perform a sequence of K runs by the Monte-Carlo algorithm
2. Use, for instance, the *Average Linkage* method to measure the similarity between two such sets

$$\{\mathbf{m}_k^{(1)}, \mathbf{m}_{k+1}^{(1)}, \ldots\} \quad \text{and} \quad \{\mathbf{m}_k^{(2)}, \mathbf{m}_{k+1}^{(2)}, \ldots\} \; .$$

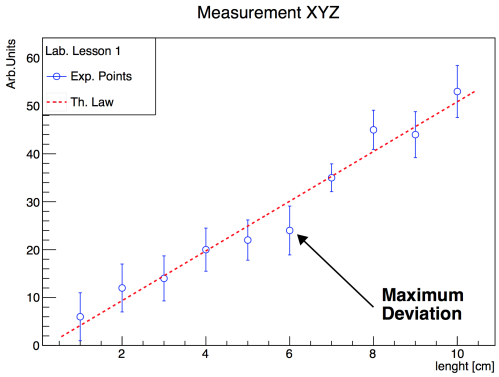
The minimum number of runs K required to obtain similar sample sets is a rough estimate of the number of 'solution islands' in the model space.

Ensemble Test



3. Each solution island must be represented by at least $M + 1$ independent sample models, where M is the dimension of the model space.
4. An estimate of the absolute minimum number of iterations required to represent the posterior is now calculated as $KI_w(M + 1)$.

Quality control of the data fit



The Most important Principle of Inversion

Solutions must fit the data (within the uncertainty)!

In a probabilistic formulation, if data and prior are compatible ('non-contradicting'), a solution not fitting the data within 2-3 standard deviations has probability almost zero (assuming Gaussian noise).

Testing the data misfit

1. Compute synthetic data and subtract it from the inverted data to obtain data residuals $\mathbf{r}_n$.
2. Compute the variance of the data residual:

$$\sigma^2 = \frac{1}{N-1}\sum_{n=1}^{N}(r_n - \bar{r})^2 \ . \tag{2}$$

3. Compare σ^2 with the diagonal elements of $\mathbf{C}_D$. They should be of the same order of magnitude.
   - If σ^2 is large compared to $C_D(i,i)$ there are two possibilities:
     - Incomplete convergence of the inversion algorithm.
     - The prior information is in conflict with the data.
   - If σ^2 is small compared to $C_D(i,i)$, the data are overfitted

Quality control in the model domain

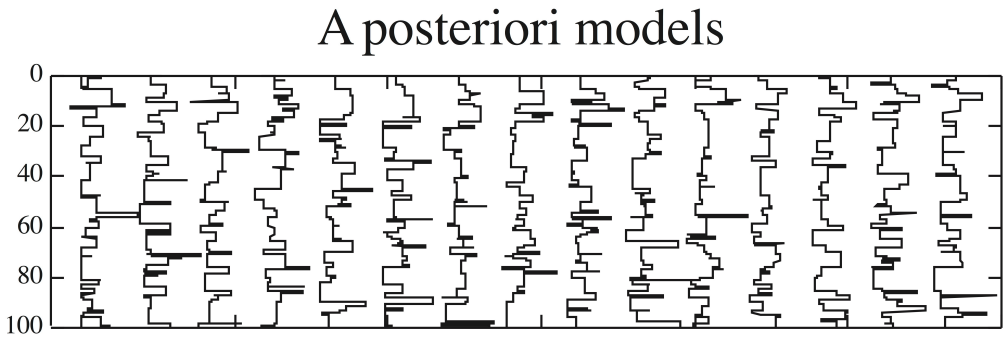
Checking the prior The prior information used in the inversion can be tested in the following way:

- Run the algorithm with $L(\mathbf{m}) = 1$ for all models $\mathbf{m}$. In this mode, the algorithm will sample the prior.
- Expect to see models that are geologically reasonable. In particular, pay attention to extreme parameter values: Are they outside reasonable bounds, or are their values too restricted?
- If the prior carries little information you should not expect realistic geological structure to appear in the prior sample models.

Checking the posterior

- Compute samples of posterior data, that is, data that are computed from the posterior models.
- Subtract a number of different posterior data sets. The noise standard deviation of such differences must be (of the order of) 2 times the estimated standard deviation of the data itself.

Interpreting the results



Samples from the posterior probability distribution is the full solution to the inverse problem, but in practice more work i required to extract useful information from these models:

1. **Presentation of output model parameters**
   - Display all output models side by side (or on top of each other, or as a movie on the computer screen)
   - The spread of the models in such a display will immediately show the uncertainty of the models.
   - The display will furthermore give an impression of the parameter correlations (showing highly probable *simultaneous* parameter variations).

Interpreting the results

2. **The distribution of model parameters** is derived from posterior output models $\mathbf{m}^{(n)}, \ldots, \mathbf{m}^{(N)}$. The posterior uncertainty of model parameter m_k is now inferred from histogram of the numbers

$$m_k^{(n)}, \ldots, m_k^{(N)}, \tag{3}$$

   approximating the posterior probability distribution of m_k.

3. **The posterior uncertainty of the combined model parameters** (m_k, m_l), can be inferred from a 2D histogram of the numbers

$$(m_k, m_l)^{(n)}, \ldots, (m_k, m_l)^{(N)}, \tag{4}$$

   giving an approximation to the joint posterior probability distribution of (m_k, m_l) showing how m_k depends on m_l.