Monte Carlo Solution of Inverse Problems

M. Griffa

Dept. of Physics, Polytechnic of Torino
Corso Duca degli Abruzzi 24, 10129, Torino (Italy)
and
Bioinformatics and High Performance Computing Labs, Bioindustry Park of Canavese
via Ribes 5, 10010, Colleretto Giacosa (Italy)
E-mail: *michele.griffa@polito.it*

Abstract

With this document, a brief (so incomplete) review of the use of Monte Carlo methods in the solution of Inverse Problems is presented along with some examples of applications to Elastodynamics. The main examples are taken from the literature in Geophysics, e.g. in Explorative Seismology, but not only.
A clear distinction is made between Inverse Problems treated as optimization problems and a Bayesian approach to their theory. In the first case, Monte Carlo methods are used for random direct search in the model space of an Inverse Problem, while, in the second case, Monte Carlo sampling of probability density functions arises as the natural method for solving teh Inverse Problem, treated from a probabilistic point of view. This second approach to such Problems is relatively new and has been developed in recent years (mainly) within the Geophysics community.

Chapter 1

Introduction

When seeking to model a physical system or phenomenon, one wishes to determine mathematical models describing the relationship between the excitation introduced in the system (input) and its related response (output). A **Forward Problem** (herefore denoted as FP) consists in the prediction of the response of the system, once the excitation and/or the internal properties are known, whereas an **Inverse Problem** (herefore denoted as IP) aims at reconstructing the excitation and/or internal structure, starting from the response.

IPs are universally omni-present in each field of research of Natural Sciences, they are at the core of the modern scientific method, because they consist in inferring theoretical information about a system or phenomenon starting from experimental data (measurements), e.g. estimating parameters of a theoretical (mainly mathematical, but not only) model, which can next be used for making predictions on other aspects of the system/phenomenon.

The classical mathematical formulation and theory of IPs is well established: solving an IP correspond, for example, to solving a problem of Linear Algebra or an integral equation, according to the type of spaces involved in its mathematical formulation. Then, Linear Algebra and Functional Analysis (respectively) are the natural contexts for studying and solving such problems.
Numerical methods for approximated solutions have been developed and used with IPs since the early stages of the development of the field, particularly for tackling the intrinsic difficulties related to the ill-posedness of most of IPs (see chapter 1 for a brief definition), related to noise in experimental data and uncertainties in model formulation.

Monte Carlo methods have been applied to the solution of IPs since the late 1960s, mainly within the Geophysics communities. A famous paper by Backus and Gilbert [1] established the foundations of the geophysical inverse theory, stressing the non-uniqueness of the problems as a fundamental recurrent property.
IPs are omni-present in the study of physical systems involving wave propagation phenomena, independently of the physical nature of the waves (electromagnetic, elastic, gravitational, or Schrodinger's). This is beacuse the way of propagation and the features of the patterns of wave interactions are directly connected with the media of propagation or the source of the perturbation: basically, the waves have a high content of information about the physical system they have travelled through or about the source they have been emitted from; measuring features of wave propagation phenomena can lead to indirect measurements (or calculations) of properties of the source or of the medium or of the systems they have interacted with.

IPs arising from wave propagation experiments and calculations are more relevant in Geophysics than in other fields of Natural Sciences due to scale issues: for studying space-extended systems such as the Earth crust in a meaningful way it is necessary to use seismic waves and sometimes this is the only manner to obtain information about its structure, composition, the physical phenomena occuring within it. Geophysical systems usually are too large and complicated for being studied by samples in laboratory, so field experiments are necessary.
For these reasons, the Geophysics communities have had a main role in the development of the theories of IPs solution, both proposing new types of problems and contributing to their mathematical formulations.

The first use of Monte Carlo methods in IPs solution is also attributed to geophysicists [2] who used them for dealing with the non-uniqueness problem. At that time (geophysical) **Monte Carlo inversion** (**MCI**) meant generating discrete Earth models (i.e. sets of parameters related to physical observables of Earth's systems) in a uniform random fashion between pairs of upper and lower bounds chosen *a priori*; each generated Earth model was tested for its fit to the available data and then accepted or rejected. Subsequent applications were to inversion of seismic body-wave travel times (compressional and shear) and to the 97 eigenfrequencies of Earth free vibrations for estimation of the variations of Earth global compressional/shear wave velocities and density as a function of depth [3, 4, 5].

At that time, the main appeal of **MCI** consisted in the fact that it avoids all assumptions of linearity or non-linearity between the observables and the unknowns representing the system (Earth in those cases) model, upon which the classical techniques rely. However, the main problem with uniform sampling in a model space (see chapter 1 for its mathematical definition) was that it is never known wether a sufficient number of models had been tested.

During the 1970s, **MCI** lost attention within the Geophysicists communities because uniform random searching of parameter spaces was thought to be inefficient and too inaccurate for problems involving a large number of unknowns, e.g. more than 50. The successes obtained in that period by the theory of linear/linearized IPs and regularization methods were mainly due to the previous development of high performance numerical methods for solving Linear Algebra and Functional Analysis problems (e.g. analytical and numerical methods for solving Fredholm linear integral equations).

As long as many IPs were recognized to be highly non-linear, their formulation was transformed into that of an optimization problem in a high-dimensional parameter space. Usually, some objective function is devised that measures the discrepancy between observables and theoretical predictions of the model and between theoretical predictions and a priori information/constraints (regularization terms). These optimization problems have been treated with gradient-based optimization numerical techniques (in the case of linear or linearizable problems) or with stochastic global optimization methods like **Simulated Annealing** (**SA**) [6, 7, 8] or **Genetic Algorithms** (**GAs**)[9, 10]. SA can be ascribed to the class of Monte Carlo methods, particularly in the class of Metropolis-Hastings algorithms, while GAs are (improperly) considered as special cases of Monte Carlo methods in many articles of geophysical IPs. Other global optimization techniques have been used in the last three decades for studying and solving ill-posed geophysical IPs: evolutionary programming methods [11], Tabu search [12, 13], neighbourhood algorithms [14, 15].

However, Monte Carlo techniques have been more directly applied to the solution

of IPs formulated in a complete probabilistic way, using a Bayesian approach developed since the mid of the 1980s by some geophysicists [16, 17, 18, 19]. This innovative and different Bayesian inference approach is presented in chapter 3 and is based on the notion of "state of information as a **probability density function**"(**PDF**) in model/data spaces. The Bayesian inference is more general than the solution based on the classical formulations of IPs, it is completely uncorrelate with the linear or non-linear nature of the forward problem: it combines the prior information known on the model with the observed data and produces the **posterior PDF** on the model parameters, which is taken as the complete solution to the given IP. With this statistical approach, uncertainties in measured data and model parameters are completely and directly considered and analyzed in the formulation and solution of the IP.

Chapter 2

Classical formulation of Inverse Problems

The general mathematical formulation of an IP makes use of the two-spaces schematisation.
Let $\wp$ be a physical system under investigation. The scientific procedure for studying it can be divided in:

- parameterization of the system, i.e. discovery of a minimal set of **model parameters** whose values completely characterize the system;
- forward modeling, i.e. discovery of the physical laws allowing us, given values of the model parameters, to make predictions on the results of measurements on some **observable parameters**; such measurements are called **experimental data**;
- inverse modeling, i.e. use of actual results of some measurements of the observable parameters to infer the actual values of the model parameters.

The distinction between model parameters and observable parameters can be formalized in the subsequent way.
Let's introduce a manifold (in the sense of Differential Geometry and Topology) $\mathcal{M}$ called the **model parameters space** and let $\mathbf{m} = \{m_1, m_2, m_3, ...\}$ be a set of local coordinates on it. Each point in the model parameters space is labeled with a set of values $\mathbf{m}$ and represents a specific model (in the common sense) of the physical system under study.
Let's now introduce another manifold $\mathcal{D}$ called the **observable parameters space**, with a set of local coordinates $\mathbf{d} = \{d_1, d_2, d_3, ...\}$ identificating a general point of it. $\mathcal{D}$ is also defined as the **data space** and can be interpreted as the space of all conceivable responses from a measurement apparatus.

In many cases, $\mathcal{M}$ and $\mathcal{D}$ might be linear spaces, M and D respectively, but they could be also infinite dimensional manifold, as function spaces. The use of manifolds instead of linear spaces is a useful generalization for considering more general situations in which there are particular constraints between different model parameters or observable ones.

According to the classical definition cited above, solving an IP corresponds to find a point in the model space, given a point in the data space. The discrete or continous nature of the model and data spaces determines the type of IP and particularly of the forward operator, which maps a point of the model space into a point of the data space:

$$F(\mathbf{m}) \;=\; \mathbf{d} \tag{2.1}$$

where F is called the forward operator and is a matrix if both spaces are linear spaces and $\mathbf{m}$, $\mathbf{d}$ are Cartesian coordinates of the respective points, a differential or integral or integro-differential operator if the spaces are function spaces, and so on.
In some cases the forward operator isn't a mathematical object but a computational procedure, e.g. a computational model derived through elaborations from theories or from algorithms for the numerical solution of mathematical problems like ordinary or partial differential equations, intregal equations, linear algebra problems, variational problems, etc... .

In many important applications, however, the model $\mathbf{m} = \{m_1, m_2, m_3, ...\}$ can include parameters with the meaning of input for the system or of its internal features. In the following, we will adopt a general definition of IP, denoting as "model"any kind of information one wishes to determine for characterizing the system, including input determining the measured data and/or parameters characterizing the system, i.e. the forward operator.

Given the FP, solving the IP involves finding the model $\mathbf{m}$ for a given data set $\mathbf{d}$. Independently of whether $\mathbf{m}$ and $\mathbf{d}$ are continuous or discrete in nature, the IP is termed well-posed [20] if it satisfies the following conditions:

- existence, i.e. a solution exists for any given data set $\mathbf{d}$;
- uniqueness, i.e., given the data point $\mathbf{d}$, the solution is unique in the model space $\mathcal{M}$;
- continuity, i.e. the inverse mapping $\mathbf{d} \longmapsto \mathbf{m}$ is continous.

The first two requirements simply state that the operator F should have a well-defined inverse F^{-1}, with co-domain equal to the entire data space, whereas the third one is a necessary, yet not sufficient, condition for the stability of the solution.
A solution can be considered stable if a small deviation $\Delta\mathbf{d}$ in the data point results in a small deviation $\Delta\mathbf{m}$ of the corresponding model point. An important quantity for characterizing the stability of an IP is the condition number, $cond(F)$, which can be defined as

$$cond(F) \;= \parallel F \parallel \cdot \parallel F^{-1} \parallel \tag{2.2}$$

where $\parallel F \parallel$ is the norm of the operator F and F^{-1} indicates the inverse (or pseudo-inverse in the case the inverse does not exist) of the same operator. It can be shown that

$$\frac{\parallel \Delta\mathbf{m} \parallel}{\parallel \mathbf{m} \parallel} \leq cond(F) \cdot \frac{\parallel \Delta\mathbf{d} \parallel}{\parallel \mathbf{d} \parallel} \tag{2.3}$$

where $\Delta\mathbf{d}$ is the variation of $\mathbf{d}$ and $\Delta\mathbf{m}$ the corresponding variation of $\mathbf{m}$. Eq. 2.3 entails that the condition number controls relative error propagation from the data to the solution, so that the IP admits stable solutions only if it is also **well-conditioned**, i.e. the condition number is not too large. It is clear that the definition of ill-conditioned problems is rather vague, compared to that of ill-posed ones. However, it should be noted that ill-conditioned problems can show properties very similar to those of ill-posed ones, in terms of sensitivity to noise and high-frequency perturbations.

Chapter 3

Inverse Problems solution as an optimization problem: classical and Monte Carlo methods

In practical applications, data are collected through measurements, and thus are affected by noise. Usually, measured data can be represented as the superposition of the "true"data vector $\mathbf{d}$, which can be obtained through the forward process as formulated in Eq. 2.1, and a set $\mathbf{n}$ of stochastic parameters representing the noise process; the IP thus can be formulated as:

$$F(\mathbf{m}) \;=\; \mathbf{d} \;=\; \widetilde{\mathbf{d}} + \mathbf{n}. \tag{3.1}$$

where $\widetilde{\mathbf{d}}$ is the measured data set.

However, after adding random noise, this equation may no longer admit a solution: the IP must therefore be reformulated as an optimization problem, where the quantity to be minimized is the misfit $C(\mathbf{m})$ between the measured data $\mathbf{d}_{obs}$ and the data calculated from a given model $\mathbf{d}$. Thus, an approximated solution can be found by minimizing the following function:

$$C(\mathbf{m}) = \| \; \mathbf{d}_{obs} \; - \; F(\mathbf{m}) \; \| \tag{3.2}$$

In the presence of noise and ill-conditioned problems, the invertibility of the operator F turns out to be an issue of relatively little interest: even if the problem can be exactly solved from a mathematical point of view, the effects of noise amplification can be disruptive to the point that the solution is actually determined by the noise itself, rather than by relevant measurement information. Due to the uncertainty introduced by noise, the global minimum of $C(\mathbf{m})$ could be not the optimal solution, while better results can be obtained by considering a **feasible set of solutions** (specifically those satisfying the condition $C(\mathbf{m}) \leq C_0$, with C_0 depending on the level of noise) which can be considered consistent with the observed data.

Even if experimental data were noise-free, the IP could still admit multiple feasible solutions because of its indetermination, either due to the lack of available experimental data or because the forward operator, failing conditions 1-2, is not exactly invertible. When multiple potential solutions are available, and minimizing the misfit function may lead to instability, a compromise between stability and accuracy of the solution can be reached by including a priori information.

In fact, one usually has an idea of how a good solution should "look like", i.e. of which properties it should reasonably possess. In IPs, techniques employed to take into account such a priori information are known as **regularization techniques**. Common methods include Tikhonovs, Levenbergs and Levenberg-Marquardts regularization techniques[21]. Although an exhaustive discussion is beyond the scope of this review, a brief explanation of at least the most common and widely known technique, Tikhonov's regularization, is deemed essential.

Let $\mathbf{m_0}$ be the default solution for a given IP and let F be a linear operator. For instance, $\mathbf{m_0}$ could be determined according to a priori information, when available, or can be simply set equal to the null solution . The Tikhonov's regularization scheme consists in minimizing, instead of the quantity given in Eq. 3.2, the following function

$$\Lambda(\mathbf{m}) \;=\; \lambda^2\Omega(\mathbf{m}) + C(\mathbf{m}) \;=\; \lambda^2 \parallel L(\mathbf{m} \;-\; \mathbf{m_0}) \parallel^2 + \parallel \widetilde{\mathbf{d}} \;-\; F(\mathbf{m}) \parallel^2 . \qquad (3.3)$$

Two competing terms are thus jointly minimized: the former is the misfit function, while the latter penalizes solutions distant from the default solution, according to the operator F. In the simplest case, i.e. with $\mathbf{m} = \mathbf{0}$ and L equal to the identity operator, this term simply reduces to the norm of the solution. The weight parameter λ controls the amount of regularization of the solution: by adjusting its value, one can regulate the sensitivity of the solution to measured data, and therefore to the noise therein, in order to counterbalance the effects of perturbations. It is thus clear that the optimal value for λ is noise dependent.

Many techniques have been proposed for solving IPs, which are often domain-specific and exploit the peculiarities of a given problem. However, even in this variegated scenario some common characteristics can be identified.
In some problems, the unknown model and the data can be related by an invertible, although generally nonlinear, operator, so that Eq. 2.1 can be exactly solved. This approach is suitable only for restricted classes of IPs, since these methods are unable to deal with ill-posedness, ill-conditioning, data uncertainty and underdetermination.

As previously mentioned, IPs are usually re-formulated as optimization problems. Supposing that an analytical representation of the forward operator is available, a formal solution can often be readily found. In the simplest case, F is a linear operator, and the problem is reduced to zeroing the derivatives of the misfit function $C(\mathbf{m})$ with respect to $\mathbf{m}$ and solving an (often large) system of coupled linear equations. In the most general case, however, that is when the operator is non linear, an analytical solution of this system may not be available. Unless the problem is somehow simplified (e.g. by linearization), we have to resort to iterative methods for multidimensional, nonlinear optimization, such as the steepest descent algorithm (belonging to the set of conjugate gradient methods) or the Gauss-Newton method. Such methods are based on the exploration of the search space, starting from an initial guess for the solution and then moving towards a local minimum based on the values of the derivates in the current point. These optimization techniques represent a valid method in the solution of IPs.
However, they may suffer from various drawbacks. In particular, they tend to be computationally intensive and liable to the presence of local minima, issue which may be particularly critical when the error landscape tends to present many local optima. Furthermore, these techniques require an analytical formulation of the objective function to be minimized, e.g. $C(\mathbf{m})$, which is generally not possible.

In many problems (most of the cases in geophysical inversion), the data-model relationship is very complicated or not expressed in a closed analytical form; linearization is not always possible or practical, e.g. when the obsevables are not differentiable functions of the unknowns (model parameters).
In these cases, stochastic direct search in the model space techniques, based on Monte Carlo methods or on Genetic Algorithms are very useful and also very reliable in appraising the solution, estimating uncertainty by means of model covariance and resolution matrices.
Monte Carlo methods as uniform sampling in model space, Simulated Annealing or Neighbourhood algorithms avoid derivatives of misfit functions, hence the numerical approximations on which linearized techniques are based. Although with finite dimensional model and data spaces, Monte Carlo methods get very computationally intensive with the number of dimensions, they are prone to a parallel computation implementation, so they can exploit the power of modern cluster computing.

Chapter 4

Probabilistic approach to the Theory of Inverse Problems and Monte Carlo solutions

While in IPs solution as a solution of an optimization problem, Monte Carlo techniques are possible alternative tools for searching a set of acceptable (or optimal) solutions in the model parameters space, with the Bayesian inference approach they are basically used as methods for sampling probability density functions (PDFs) which take the same role as misfit functions in optimization methods.

A probabilistic approach to the theory of IPs was introduced by A. Tarantola and B. Valette in 1982 [16] and further developed in the subsequent years by Tarantola [18] and Mosegaard [17, 19]. The papers by Mosegaard and Sambridge[22, 23] are exhaustive review papers in the use of Monte Carlo methods both as stochastic direct search techniques and sampling methods of posterior probability density functions.

This probabilistic approach is called Bayesian in the sense that the solution of an IP is formulated/constructed as a posterior probability density function that update the information on model parameters derived from an a-priori probability density function, exploiting the information derived from a probability density function defined over the data space (errors on measured data).
Below, the basic theory and definitions introduced by Tarantola[18] are introduced.

According to Kolmogorov axiomatic formulation of Probability Theory, given a manifold $\mathcal{X}$ (that can be identified with model parameter space $\mathcal{M}$ or observable parameter space $\mathcal{D}$ or their cartesian product), a mapping M between the set of all possible subsets of $\mathcal{X}$ and R^+ can be always defined such that given A and B two subsets of $\mathcal{X}$ (called events in Probability Theory)

$$M(A \cup B) \;=\; M(A) \;+\; M(B) \text{ if } A \cap B \;=\; \emptyset \tag{4.1}$$

and

$$M(\emptyset) \;=\; 0 \tag{4.2}$$

where $\emptyset$ is the null set.

A particular collection of subsets of $\mathcal{X}$ is called a σ-field if it is an algebraic field and satisfies the property that, given A as a general element of it, $\mathcal{X}/A$ is still an element of the same collection (the symbol / means the complement set operation). Then, M is defined as a measure over $\mathcal{X}$ and when $M(\mathcal{X})$ is finite it is called a **probability**

distribution over $\mathcal{X}$, $M \doteq P$. $P(A)$ is called the probability of the event A. Generally, probability distributions over a manifold $\mathcal{X}$ are normalized to one such that $P(\mathcal{X}) = 1$.

Given a set of local coordinates over $\mathcal{X}$, $\mathbf{x} = \{x_1, x_2, x_3, ...\}$, it has been demonstrated that $\forall P()$ over $\mathcal{X}$, $\exists\ f(\mathbf{x})$ such that

$$P(A) \ = \ \int_A d\mathbf{x}\ f(\mathbf{x}). \tag{4.3}$$

$f(\mathbf{x})$ is the probability density function over $\mathcal{X}$ associated to the probability distribution $P()$, given the set of coordinates $\mathbf{x}$.

The central postulate of the theory of Tarantola and Mosegaard is that the most general way for describing any state of information over $\mathcal{X}$ is by defining a probability distribution $P()$ over $\mathcal{X}$.

Among the different possible definitions of probability distributions (or generally normalizible measures), a key role is played by the *volume* distribution defined as follows: given $dV(\mathbf{x}) \ = \ v(\mathbf{x}) \cdot d\mathbf{x}$ as the volume element of the manifold $\mathcal{X}$ around the point $\mathbf{x}$ (where $v(\mathbf{x})$ is the density of volume of $\mathcal{X}$ in the coordinate system $\mathbf{x}$, $v(\mathbf{x}) \ = \ det(g_{ij}(\mathbf{x}))$, $g_{ij}(\mathbf{x})$ being the metric tensor of the manifold $\mathcal{X}$),

$$V(A) \ = \ \int_A d\mathbf{x}\ v(\mathbf{x}) \tag{4.4}$$

defines the volume (measure) of the event A, while

$$V(\mathcal{X}) \ = \ \int_{\mathcal{X}} d\mathbf{x}\ v(\mathbf{x}) \tag{4.5}$$

is the total volume of the manifold. Then

$$\frac{V(A)}{V} \ = \ \int_A d\mathbf{x}\ \frac{v(\mathbf{x})}{V} \tag{4.6}$$

introduces a particular probility density function $\mu(\mathbf{x}) \ = \ \frac{v(\mathbf{x})}{V}$ called the **homogeneous probability density function** over the manifold $\mathcal{X}$. When $\mathcal{X}$ is a linear space X and $\mathbf{x}$ is a Cartesian coordinate system, $\mu(\mathbf{x})$ is a constant.

Usually, homogeneous probability density functions are used as prior probability density functions in Bayesian inference. Note also that, the homogeneous probability distribution corresponds to the classical simplest way through which the concept of probability of an event is defined using Measure Theory.

Two fundamental concepts are then introduced:

- operation of conjuction of two (or more) probability density functions;
- operation of disjuction of two (or more) probability density functions.

These two operations have been introduced in order to *compose* the states of information.

Given $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ two probility density functions over $\mathcal{X}$

$$(f_1 \wedge f_2)(\mathbf{x}) \ = \ \frac{1}{\nu} \cdot \frac{f_1(\mathbf{x}) \cdot f_2(\mathbf{x})}{\mu(\mathbf{x})} \tag{4.7}$$

is a new probability density function called the conjuction of $f_1()$ and $f_2()$ (ν is a normalization-to-1 constant over $\mathcal{X}$), while

$$(f_1 \vee f_2)(\mathbf{x}) \;=\; \frac{1}{2} \cdot (f_1(\mathbf{x}) \;+\; f_2(\mathbf{x}) \tag{4.8}$$

is the disjunction probability density of $f_1()$ and $f_2()$.

A p-event (probability-event) is introduced by Tarantola[18] as a specific type of probability distribution function over $\mathcal{X}$: $\forall$ event A, $\exists$ $P_A()$ probability distribution such that

$$P_A(B) \;=\; \int_B d\mathbf{x}\; \mu_A(\mathbf{x}) \tag{4.9}$$

where

$$\mu_A(\mathbf{x}) \;= \begin{cases} k' \cdot \mu(\mathbf{x}) \text{if} \mathbf{x} \;\in\; A \\ 0 \qquad \text{otherwise} \end{cases}$$

and k' is a normalization-to-1 constant over $\mathcal{X}$.

With these definitions in mind, Tarantola has proposed a new way of defining a conditional probability: given a probability distribution $P()$ and a p-event associated to an event A,

$$(P \wedge M_A)(B) \;=\; \int_B \; d\mathbf{x}\; (f \wedge \mu_A)(\mathrm{x}) \tag{4.10}$$

is called the conditional probability of the event B given the event A as the conditonal one. Using Eq. 4.7 in Eq. 4.10, it can be demonstrated that

$$(P \wedge M_A)(B) \;=\; \frac{P(A \cap B)}{P(A)} \tag{4.11}$$

where the right hand term is the usual definition of the probability of the event B conditioned by the event A. Following Eq. 4.11, $(P \wedge M_A)(B)$ can then be expressed using probability densities as

$$f(\mathbf{x}|\mathbf{y}) \;=\; \frac{f(\mathbf{x}, \mathbf{y})}{f(\mathbf{y})} \tag{4.12}$$

where $\mathbf{x}$ and $\mathbf{y}$ are sets of random variables belonging to the the same space $\mathcal{X}$, $f(\mathbf{x}, \mathbf{y})$ is the joint probability density function and $f(\mathbf{y})$ is the marginal probability density ($f(\mathbf{y}) \;=\; \int_{\mathcal{X}} \; d\mathbf{y}\; f(\mathbf{x}, \mathbf{y})$).

Bayes inference is based on a simple theorem about conditional probability:
the probability of the conjuction of two events A and B, i.e. $A \cap B$, can be written as

$$P(A \cap B) \;=\; P(A|B) \cdot P(B) \tag{4.13}$$

or as

$$P(A \cap B) \;=\; P(B|A) \cdot P(A) \tag{4.14}$$

due to the commutativity of the conjucntion operation between events (sets in parameter space). That implies that

$$P(B|A) \;=\; \frac{P(A|B) \cdot P(B)}{P(A)}. \tag{4.15}$$

Eq. 4.15 can then be charged with a complex and important meaning: it expresses the probability that the event B is the cause of the event A. It is usually called a **posterior probability**.

The definition of a conditional probability with the use of the concept of conjunction of propability densities has suggested Tarantola et al.[16, 17, 18] that a **posterior probability** can be obtained as the conjuction of other probability densities.
As cited above, regarding the general theory of IPs, they have exploited this main idea: the solution of an IP essentially consists in updating some information on a physical system (i.e. on its parameters $\mathbf{m}$) with measured data $\mathbf{d}$, for obtaining new estimates of the system parameters (a new model $\mathbf{m}$). The new model might be consistent with **a priori** information (the old model) and measured data.

The updating of this information on model space needs three main ingredients:

- **a-priori information on model space**, expressed as a probability density function $\rho_{\mathcal{M}}(\mathbf{m})$;
- **a-priori information on data space**, expressed as a probability density function $\rho_{\mathcal{D}}(\mathbf{d})$ and obtained from experimental measurements and statistical analysis of its errors;
- **theoretical information** about the correlation between model parameters and observable parameters, expressed as a joint probability density function $\Theta(\mathbf{m}, \mathbf{d})$ and correlated with the forward operator F.

The reason for which the classical formulation of the relation between model parameters and observed parameters, expressed in Eq. 2.1, is substituted with a joint probability density on the parameters space $\mathcal{X} = \mathcal{M}$ x $\mathcal{D}$ is, as previously cited, that measurements uncertainties and model inperfections make difficult putting together measurements and physical predictions, so usually the problem of finding a model $\mathbf{m}$ satisfying the error free equation Eq. 2.1 doesn't exist. A probabilistic correlation between model parameters and observable ones is a more general formulation of the forward problem, it avoids the use of optimization methods for solving a problem as expressed in Eq. 3.1, and it leads to a full probabilistic treatment of the corresponding IP, exploiting the information over model and data spaces (on the form of a-priori probability densities) by the notion of conjunction of states of information as expressed in Eq. 4.7.

Defined the parameters space as $\mathcal{X} = \mathcal{M}$ x $\mathcal{D}$, then $\mathbf{x} = (\mathbf{m}, \mathbf{d})$. The posterior probability density over $\mathcal{X}$ is defined as

$$\begin{aligned} \sigma(\mathbf{x}) &= (\Theta \wedge \rho)(\mathbf{x}) \\ &= k \cdot \frac{\Theta(\mathbf{x}) \cdot \rho(\mathbf{x})}{\mu(\mathbf{x})} \end{aligned} \tag{4.16}$$

where $\mu(\mathbf{x})$ is the homogeneous probability density expressed as the product of the homogenous probability densities over the two spaces ($\mu_{\mathcal{M}}(\mathbf{m})$ and $\mu_{\mathcal{D}}(\mathbf{d})$ respectively), while k is a normalization-to-1 constant over $\mathcal{X}$.

The a-priori probability density $\rho(\mathbf{x})$ is also the product of the two a-priori densities $\rho_{\mathcal{M}}(\mathbf{m})$ and $\rho_{\mathcal{D}}(\mathbf{d})$ under the hypothesis of independence of the two states of information over the distinct spaces.

It results that

$$\sigma(\mathbf{x}) = \sigma(\mathbf{m}, \mathbf{d}) = k \cdot \rho_{\mathcal{M}}(\mathbf{m}) \cdot H(\mathbf{m}, \mathbf{d}) \tag{4.17}$$

where

$$H(\mathbf{m},\mathbf{d}) \;=\; \frac{\theta(\mathbf{d}|\mathbf{m})\cdot\rho_{\mathcal{D}}(\mathbf{d})}{\mu_{\mathcal{D}}(\mathbf{d})} \tag{4.18}$$

assuming a definition for the theoretical information as

$$\Theta(\mathbf{m},\mathbf{d}) \;=\; \theta(\mathbf{d}|\mathbf{m})\cdot\mu_{\mathcal{M}}(\mathbf{m}) \tag{4.19}$$

i.e. the product of the homogeneous probability density obtained over model space times a conditional probability density relating model and data parameters.

Starting from Eq. 4.20, one can evaluate the posterior marginal probability density over the model space,

$$\sigma_{\mathcal{M}}(\mathbf{m}) \;=\; \int_{\mathcal{D}} d\mathbf{d}\;\sigma(\mathbf{m},\mathbf{d}) \tag{4.20}$$

$$\;=\; k\cdot\rho_{\mathcal{M}}(\mathbf{m})\cdot L(\mathbf{m}) \tag{4.21}$$

where $L(\mathbf{m})$ is called the **likelihood function** and is an estimate of how good the model $\mathbf{m}$ is in explaining the data $\mathbf{d}$,

$$L(\mathbf{m}) \;=\; \int_{\mathcal{D}} d\mathbf{d}\;H(\mathbf{m},\mathbf{d}). \tag{4.22}$$

$\sigma_{\mathcal{M}}(\mathbf{m})$ is the solution of the model space using this probabilistic approach, which is based on the identification of the state of information as a probability density function. It may be a complicated posterior probability density over model space, multimodal or with divergent moments. However, it lets obtain a "picture"of the solution of the IP: sampling from it with Monte Carlo methods leads to information about the acceptable solutions of the IP, i.e. models $\mathbf{m}$ that, under the uncertainties, can lead to data parameters that are "close"to the measured ones.

If both model and data spaces are linear spaces with Cartesian coordinates and there aren't uncertainites in the model formulation, then

$$\Theta(\mathbf{m},\mathbf{d}) \;=\; const\cdot\delta(\mathbf{d} \;-\; F(\mathbf{m})) \tag{4.23}$$

while a Gaussian hypothesis for errors distribution in model formulation leads to

$$\Theta(\mathbf{m},\mathbf{d}) \;=\; const\cdot exp[0.5\cdot(\mathbf{d} \;-\; F(\mathbf{m}))^T\cdot(C_T)^{-1}\cdot(\mathbf{d} \;-\; F(\mathbf{m}))] \tag{4.24}$$

where C_T is the covariance matrix of the random vector $(\mathbf{d} \;-\; F(\mathbf{m}))$.

Different types of a-priori probability densities in both spaces, $\rho_{\mathcal{M}}(\mathbf{m})$ and $\rho_{\mathcal{D}}(\mathbf{d})$, can be assumed under different hypothesis, e.g. Gaussian errors for measured data and a-priori information on the model space, resulting in simple analytical forms for $\sigma_{\mathcal{M}}(\mathbf{m})$.

In any case, setting the IP as a problem of conjunction of states of information has led to its solution as a posterior probability density over $\mathcal{M}$. Sampling from $\sigma_{\mathcal{M}}(\mathbf{m})$ can not only give an heuristic "view"over the acceptable solutions but also let calculate interesting parameters as the probability that a solution model belongs to a cetain region A of the model space, mean value, likelihood value, etc. The analysis of uncertainties and resolution in the solution of the IP can be obtained, for example, calculating the posterior covariance matrix of the model parameters set,

$$\widetilde{C}_M \;=\; \int_{\mathcal{M}} d\mathbf{m}\;(\mathbf{m} \;-\; \langle\mathbf{m}\rangle)\cdot(\mathbf{m} \;-\; \langle\mathbf{m}\rangle)^T\cdot\sigma(\mathbf{m}), \tag{4.25}$$

and comparing it with the a-priori covariance matrix over model space,

$$C_M \;=\; \int_{\mathcal{M}} d\mathbf{m}\;(\mathbf{m} \;-\; \langle\mathbf{m}\rangle)\cdot(\mathbf{m} \;-\; \langle\mathbf{m}\rangle)^T \cdot \rho_M(\mathbf{m}). \tag{4.26}$$

Monte Carlo sampling has been used for sampling both from a-priori model space probability density and from posterior probability density, in order to make the update of accepatble solutions to the IP. Monte Carlo numerical integration techniques, e.g. importance sampling, has been used for obtaining information from the posterior probability over model space, in order to characteriza the solutions.

Examples of applications of this Bayesian inference approach to IPs and the use of Monte Carlo methods for their solutions can be found in [19](regarding synthetic problems occuring in Exploration Seismics), in [24](for a problem of estimating past surface temperature changes from measured temperature profile through deep ice boreholes in Paleoclimatogy), in [25, 26] (estimation of the longitudinal and shear-waves velocities dependence on depth from measured arrival times of seismic disturbances generfated on the Moon by moonquakes, meteorite impacts and artificial impacts).

Bibliography

[1] G.E. Backus and J.F. Gilbert. Numerical applications of a formalism for geophysical inverse problems. *Geophys. J. R. Astron. Soc.*, 1967.

[2] V.I. Keilis-Borok and T.B. Yanovskaya. Inverse problems of seismology. *Geophys. J.*, 1967.

[3] F. Press. Earth models obtained by monte carlo inversion. *J. Geophys. Res.*, 1968.

[4] F. Press. Earth models consistent with geophysical data. *Phys. Earth Planet. Inter.*, 1970.

[5] F. Press. Regionalized earth models. *J. Geophys. Res.*, 1970.

[6] S.C. Kirkpatrick, D. Gelatt, and M.P.Vecchi. Optimization by simulated annealing. *Science*, 1983.

[7] S. Geman and D. Geman. Stochastic relaxation, gibbs distribution and the bayesian restoration of images. *IEEE Trans. Patt. Anal. Mach. Int.*, 1984.

[8] *Genetic Algorithm and Simulated Annealing.* Pitman, 1987.

[9] *Adaptation in Natural and Artificial Systems.* MIT Press, 1990.

[10]

[11] J-B.H. Minster, N.P. Williams, T.G. Masters, J.F. Gilbert, and J.S. Haase. Application of evolutionary programming to earthquake hypocenter determination. In *Proceedings of 4th Annual Conference of Evol. Prog.*

[12] D. Cvijovic and J. Klinowski. Taboo search: An approach to the multiple minima problem. *Science*, 1995.

[13] R. Vinther and K. Moseegard. Seismic inversion through tabu search. *Geophys. Prospect.*, 1996.

[14] M. Sambridge. Geophysical inversion with a neighbourhood algorithm, i: searching a parameter space. *Geopys. J. Int.*, 1999.

[15] M. Sambridge. Geophysical inversion with a neighbourhood algorithm, ii: appraising the ensemble. *Geopys. J. Int.*, 1999.

[16] A. Tarantola and B. Valette. Inverse problems = quest for information. *J. of Geophys.*, 1982.

[17] *International Handbook of Earthquake & Engineering Seismology*, chapter Probabilistic Approach to Inverse Problems. Academic Press, 2002.

[18] *Inverse Problem Theory and Methods for Model Parameter Estimation*. Siam, 2004.

[19] K. Mosegaard and A. Tarantola. Monte carlo sampling of solutions to inverse problems. *J. Geophys. Res.*, 1995.

[20] *Lectures on Cauchy Problem in Linear Partial Differential Equations*. Yale University Press, 1923.

[21] *Regularization of Inverse Problems*. Kluwer Academic Publishers, 1996.

[22] K. Mosegaard and M. Sambridge. Monte carlo analysis of inverse problems. *Inverse Problems*, 2002.

[23] M. Sambridge and K. Mosegaard. Monte carlo methods in geophysical inverse problems. *Reviews of Geophysics*, 2002.

[24] D. Dahl-Jensen, K. Moseegard, N. Gundestrup, G.D. Clow, S.J. Johnsen, A.W. Hansen, and N. Balling. Past temperatures directly from the greenland ice sheet. *Science*, 1998.

[25] A. Khan, K. Moseegard, and K. L. Rasmussen. A new seismic velocity model for the moon from a monte carlo inversion of the apollo lunar seismic data. *Geophys. Res. Lett.*, 2000.

[26] A. Khan and K. Moseegard. New information on the deep lunar interior from an inversion of lunar free oscillation periods. *Geophys. Res. Lett.*, 2001.