Chapter 6
Probabilistic Methods of Inverse Problem Solution

Abstract This chapter considers the methods of solving the linear discrete inverse problems using the probabilistic approach. We review two major techniques—the maximum likelihood and the maximum a posteriori estimation methods. The Bayes estimation method makes it possible to introduce some a priori information about the properties of the solution in the inversion. We demonstrate that the numerical implementation of these methods is similar to the weighted least-squares and Tikhonov's regularization methods, respectively. A summary of the typical stochastic inversion techniques, e.g., Monte Carlo, genetic algorithm (GA), and simulated annealing (SA) methods, is also provided.

Keywords Maximum likelihood method · Bayes estimation · Stochastic methods · Monte Carlo · Genetic algorithm (GA) · Simulated annealing (SA)

In Chap. 5, we considered the methods of solving the linear discrete inverse problems using the deterministic approach based on Tikhonov regularization. However, there exists an alternative approach based on the ideas of the probability theory. Therefore, in this chapter presents several methods for inverse problem solutions using the probabilistic approach following Zhdanov (1993, 2002, 2015).

6.1 Maximum Likelihood Method

As discussed in Chap. 2, the probability distribution can be described by a very complicated function in general cases. However, according to the *central limit theorem*, a large sample of a random variable tends to a very simple distribution, the so-called *Gaussian (or normal) distribution* , as the size of the random sample increases.

The joint distribution of two independent Gaussian variables is just the product of two univariate distributions. When the data forming a vector $\mathbf{d}$ are correlated (with mean $\langle \mathbf{d} \rangle$ and covariance $\boldsymbol{\sigma} = [\sigma_{ij}]$), the appropriate distribution turns out to be as follows (Menke 2018):

M. S. Zhdanov, *Advanced Methods of Joint Inversion and Fusion of Multiphysics Data*, Advances in Geological Science,
https://doi.org/10.1007/978-981-99-6722-3_6

$$P(\mathbf{d}) = \frac{|\,\boldsymbol{\sigma}\,|^{-\frac{1}{2}}}{(2\pi)^{\frac{N}{2}}} \exp[-\frac{1}{2}(\mathbf{d} - \langle\mathbf{d}\rangle)^T \boldsymbol{\sigma}^{-1}(\mathbf{d} - \langle\mathbf{d}\rangle)]. \tag{6.1}$$

The idea that the model and data are related by an explicit relationship,

$$\mathbf{Am} = \mathbf{d}, \tag{6.2}$$

can now be reinterpreted in the sense that this relationship holds only for the mean data:

$$\mathbf{Am} = \langle\mathbf{d}\rangle. \tag{6.3}$$

Substituting (6.3) into (6.1), we can rewrite the distribution of the data as follows:

$$P(\mathbf{d}) = \frac{|\,\boldsymbol{\sigma}\,|^{-\frac{1}{2}}}{(2\pi)^{\frac{N}{2}}} \exp[-\frac{1}{2}(\mathbf{d} - \mathbf{Am})^T \boldsymbol{\sigma}^{-1}(\mathbf{d} - \mathbf{Am})]$$

$$= \frac{|\,\boldsymbol{\sigma}\,|^{-\frac{1}{2}}}{(2\pi)^{\frac{N}{2}}} \exp[-\frac{1}{2} f_\sigma(\mathbf{m})], \tag{6.4}$$

where

$$f_\sigma(\mathbf{m}) = (\mathbf{d} - \mathbf{Am})^T \boldsymbol{\sigma}^{-1}(\mathbf{d} - \mathbf{Am}).$$

Under this assumption, we can say that the optimum values for the model parameters are those that maximize the probability that the observed data are, in fact, observed. Thus, the method of maximum likelihood is based on maximization of the probability function (6.4)

$$P(\mathbf{d}) = \max. \tag{6.5}$$

Clearly, the maximum of $P(\mathbf{d})$ occurs when the argument of the exponential function has maximum or when $f_\sigma(\mathbf{m})$ has minimum:

$$f_\sigma(\mathbf{m}) = (\mathbf{d} - \mathbf{Am})^T \boldsymbol{\sigma}^{-1}(\mathbf{d} - \mathbf{Am}) = \min. \tag{6.6}$$

Let us calculate the first variation of functional f_σ:

$$\delta f_\sigma(\mathbf{m}) = -(\delta\mathbf{Am})^T \boldsymbol{\sigma}^{-1}(\mathbf{d} - \mathbf{Am}) - (\mathbf{d} - \mathbf{Am})^T \boldsymbol{\sigma}^{-1}(\delta\mathbf{Am}).$$

It can be shown that for symmetrical matrix $\boldsymbol{\sigma}^{-1}$, the following equality holds:

$$\mathbf{a}^T \boldsymbol{\sigma}^{-1}\mathbf{b} = \mathbf{b}^T \boldsymbol{\sigma}^{-1}\mathbf{a},$$

where $\mathbf{a}$ and $\mathbf{b}$ are two arbitrary column vectors. Therefore, we can write the necessary condition for the functional f_σ to have a minimum as follows:

$$\delta f_\sigma(\mathbf{m}) = -2(\mathbf{A}\delta\mathbf{m})^T\boldsymbol{\sigma}^{-1}(\mathbf{d}-\mathbf{A}\mathbf{m}) = -2(\delta\mathbf{m})^T\mathbf{A}^T\boldsymbol{\sigma}^{-1}(\mathbf{d}-\mathbf{A}\mathbf{m}) = 0. \quad (6.7)$$

From Eq. (6.7), we obtain at once the following equation:

$$\mathbf{A}^T\boldsymbol{\sigma}^{-1}(\mathbf{d}-\mathbf{A}\mathbf{m}) = 0.$$

The last formula provides the following normal system of equations for the "pseudo-solution" of the minimization problem (6.6):

$$\mathbf{A}^T\boldsymbol{\sigma}^{-1}\mathbf{A}\mathbf{m} = \mathbf{A}^T\boldsymbol{\sigma}^{-1}\mathbf{d}. \quad (6.8)$$

If the matrix $\left(\mathbf{A}^T\boldsymbol{\sigma}^{-1}\mathbf{A}\right)$ is non-singular, then we can multiply both sides of normal system (6.8) by inverse matrix, $(\mathbf{A}^T\boldsymbol{\sigma}^{-1}\mathbf{A})^{-1}$, and write the pseudo-solution of minimization problem (6.6) in the explicit form as follows:

$$\mathbf{m}_0 = (\mathbf{A}^T\boldsymbol{\sigma}^{-1}\mathbf{A})^{-1}\mathbf{A}^T\boldsymbol{\sigma}^{-1}\mathbf{d}. \quad (6.9)$$

Comparing the last formula with the corresponding equation for the weighted least-squares method (5.31), we see that we have obtained exactly the same result if we substitute matrix $\mathbf{W}^2$ for $\boldsymbol{\sigma}^{-1}$:

$$\mathbf{W}^2 = \boldsymbol{\sigma}^{-1}. \quad (6.10)$$

Note that if data happen to be uncorrelated, then the covariance matrix becomes diagonal:

$$\boldsymbol{\sigma} = [\mathbf{diag}(\sigma_i^2)], \quad (6.11)$$

and the elements of the main diagonal are the variances of the data. In this case, the weights are given by the following formula:

$$w_i^2 = \frac{1}{\sigma_i^2}. \quad (6.12)$$

The functional

$$f_w(\mathbf{m}) = \chi^2(\mathbf{m}) = \sum_{i=1}^{N}\left(\frac{r_i}{\sigma_i}\right)^2 = \sum_{i=1}^{N}\left(\frac{d_i - d_i^p}{\sigma_i}\right)^2 \quad (6.13)$$

is called a "chi-square."

In the cases where the measurement errors are normally distributed, the quantity χ^2 is a sum of N squares of normally distributed variables, each normalized to its variance. Thus, by applying the weighted least-squares method, we can select the smaller weights for data with bigger standard deviations (less accurate data) and the

bigger weights for data with smaller standard deviations (more certain data). If the data have equal variances, σ_0^2, then the weighting matrix becomes scalar:

$$\mathbf{W}^2 = \boldsymbol{\sigma}^{-1} = \frac{1}{\sigma_0^2}\mathbf{I},$$

and the chi-square functional becomes equal to the conventional misfit functional.

6.2 The Maximum a Posteriori Estimation Method (The Bayes Estimation)

Let us consider the regularization technique from the point of view of probability theory (Tarantola 1987). First of all, we introduce the following (normally distributed) densities of probability:
(1) $P(\mathbf{d}/\mathbf{m})$ is a conditional density of probability of the data $\mathbf{d}$, given the model $\mathbf{m}$. It means that it is the probability density of theoretical data $\mathbf{d}$ to be expected from a given model $\mathbf{m}$.
(2) $P(\mathbf{m}/\mathbf{d})$ is a conditional density of probability of a model $\mathbf{m}$, given the data $\mathbf{d}$. According to the Bayes theorem, the following equation holds:

$$P(\mathbf{m}/\mathbf{d}) = \frac{P(\mathbf{d}/\mathbf{m})P(\mathbf{m})}{P(\mathbf{d})}, \tag{6.14}$$

where $P(\mathbf{d})$ and $P(\mathbf{m})$ are unconditional probability densities for data and model parameters, respectively. It is assumed that

$$\langle \mathbf{m} \rangle = \mathbf{m}_{apr},$$

where $\mathbf{m}_{apr}$ is an a priori constrained expectation of the model, and

$$[\mathrm{cov}(m_i, m_j)] = \boldsymbol{\sigma}_m.$$

Thus, considering normally distributed parameters, we have the following probability distribution of the model, **m:**

$$P(\mathbf{m}) = \frac{|\,\boldsymbol{\sigma}_m\,|^{-\frac{1}{2}}}{(2\pi)^{\frac{L}{2}}} \exp[-\frac{1}{2}(\mathbf{m} - \mathbf{m}_{apr})^T \boldsymbol{\sigma}_m^{-1}(\mathbf{m} - \mathbf{m}_{apr})]. \tag{6.15}$$

Analogously, it is assumed that,

$$[\mathrm{cov}(d_i, d_j)] = \boldsymbol{\sigma}_d$$

and we can write for the conditional density of probability of the data $\mathbf{d}$

$$P(\mathbf{d}/\mathbf{m}) = \frac{|\,\boldsymbol{\sigma}_d\,|^{-\frac{1}{2}}}{(2\pi)^{\frac{N}{2}}} \exp[-\frac{1}{2}(\mathbf{d}-\mathbf{A}\mathbf{m})^T\boldsymbol{\sigma}_d^{-1}(\mathbf{d}-\mathbf{A}\mathbf{m})]. \tag{6.16}$$

The maximum likelihood method can now be used to find the model $\mathbf{m}_0$ which maximizes the conditional probability of a model, $P(\mathbf{m}/\mathbf{d})$:

$$P(\mathbf{m}/\mathbf{d}) = \frac{|\,\boldsymbol{\sigma}_d\,|^{-\frac{1}{2}}}{(2\pi)^{\frac{N}{2}}} \exp[-\frac{1}{2}(\mathbf{d}-\mathbf{A}\mathbf{m})^T\boldsymbol{\sigma}_d^{-1}(\mathbf{d}-\mathbf{A}\mathbf{m})] \times$$

$$\times \frac{|\,\boldsymbol{\sigma}_m\,|^{-\frac{1}{2}}}{(2\pi)^{\frac{L}{2}}} \exp[-\frac{1}{2}(\mathbf{m}-\mathbf{m}_{apr})^T\boldsymbol{\sigma}_m^{-1}(\mathbf{m}-\mathbf{m}_{apr})]P^{-1}(\mathbf{d}). \tag{6.17}$$

It is evident that, to maximize $P(\mathbf{m}/\mathbf{d})$, we have to minimize the sum of the expressions in the exponential factors in Eq. (6.17):

$$f_{Bayes} = (\mathbf{d}-\mathbf{A}\mathbf{m})^T\boldsymbol{\sigma}_d^{-1}(\mathbf{d}-\mathbf{A}\mathbf{m}) + (\mathbf{m}-\mathbf{m}_{apr})^T\boldsymbol{\sigma}_m^{-1}(\mathbf{m}-\mathbf{m}_{apr}). \tag{6.18}$$

Note that the minimization of the first term in the above equation gives the classical maximum likelihood or weighted least-squares method.

Let us calculate the first variation of f_{Bayes}:

$$\delta f_{Bayes} = -2(\mathbf{A}\delta\mathbf{m})^T\boldsymbol{\sigma}_d^{-1}(\mathbf{d}-\mathbf{A}\mathbf{m}) + 2(\delta\mathbf{m})^T\boldsymbol{\sigma}_m^{-1}(\mathbf{m}-\mathbf{m}_{apr}) = 0.$$

From the last equation, we have

$$(\delta\mathbf{m})^T[\mathbf{A}^T\boldsymbol{\sigma}_d^{-1}(\mathbf{d}-\mathbf{A}\mathbf{m}) - \boldsymbol{\sigma}_m^{-1}(\mathbf{m}-\mathbf{m}_{apr})] = 0.$$

Thus, the normal system of equations for minimization of f_{Bayes} can be written as follows:

$$\mathbf{A}^T\boldsymbol{\sigma}_d^{-1}(\mathbf{d}-\mathbf{A}\mathbf{m}) - \boldsymbol{\sigma}_m^{-1}(\mathbf{m}-\mathbf{m}_{apr}) = 0,$$

From the last formula, we have at once the following equation:

$$(\mathbf{A}^T\boldsymbol{\sigma}_d^{-1}\mathbf{A} + \boldsymbol{\sigma}_m^{-1})\mathbf{m} = \mathbf{A}^T\boldsymbol{\sigma}_d^{-1}\mathbf{d} + \boldsymbol{\sigma}_m^{-1}\mathbf{m}_{apr}. \tag{6.19}$$

We can write the solution of Eq. (6.19) in the closed form as follows:

$$\mathbf{m}_0 = (\mathbf{A}^T\boldsymbol{\sigma}_d^{-1}\mathbf{A} + \boldsymbol{\sigma}_m^{-1})^{-1}(\mathbf{A}^T\boldsymbol{\sigma}_d^{-1}\mathbf{d} +_m^{-1} \mathbf{m}_{apr}). \tag{6.20}$$

By comparing Eqs. (6.20) and (5.36), we see that

$$\boldsymbol{\sigma}_m^{-1} = \alpha\mathbf{W}_m^2, \tag{6.21}$$

so $\boldsymbol{\sigma}_m^{-1}$ plays the role of the regularization parameter and the model parameter weights simultaneously.

Let us assume now that we have uncorrelated data with equal variances,

$$\boldsymbol{\sigma}_d = \sigma_d^2 \mathbf{I},$$

and similarly for the a priori covariance of the model,

$$\boldsymbol{\sigma}_m = \sigma_m^2 \mathbf{I}.$$

Then Eq. (6.20) takes the following form:

$$\mathbf{m}_0 = (\mathbf{A}^T \mathbf{A} + k\mathbf{I})^{-1} (\mathbf{A}^T \mathbf{d} + k\mathbf{m}_{apr}), \tag{6.22}$$

where

$$k = \frac{\sigma_d^2}{\sigma_m^2} = \alpha \tag{6.23}$$

plays the role of the regularization parameter.

We can see from formula (6.23) that large values of the variance σ_m^2 of the model parameters correspond to a small regularization parameter α, and vice versa, large values of α correspond to a small variance σ_m^2. This means that, without regularization (α close to zero), the uncertainty in determining the inverse model is great, while with regularization, it becomes smaller. The last formula illustrates once again the close connection between the probabilistic (Tarantola 1987) and deterministic (Tikhonov and Arsenin 1977) approaches to regularization.

6.3 Stochastic Methods of Inversion

We have already discussed in this and previous chapters that there are two different major points of view in addressing the inverse problem:
(a) the algebraic (deterministic) point of view, dating back to the works of Lanczos (1961), Backus and Gilbert (1967), Backus (1970a, b, c), Marquardt (1963, 1970), Tikhonov and Arsenin (1977), etc.;
(b) the probabilistic (stochastic) point of view, formulated in the pioneering papers of Foster (1961), Franklin (1970), Jackson (1972), Tarantola and Valette (1982), Tarantola (1987, 2005), etc.

The stochastic point of view is widely used in literature because it is closely associated with the statistical nature of the noise in the data. At the same time, it has been demonstrated in many publications (e.g., the classical work by Sabatier (1977)) that in many cases, both points of view result in similar computational algorithms (see Sects. 6.1 and 6.2).

The Monte Carlo inversion methods represent a general approach based on the stochastic point of view (Metropolis and Ulam 1949; Metropolis et al. 1953). They are named after the famous Casino in Monaco. There are two major types of Monte Carlo methods. The first one is based on an extensive random search in the space M of the model parameters for a solution, which generates the predicted data from the data space, D, close to the observed data, realizing the global minimum of the corresponding misfit functional $f(\mathbf{m})$ (e.g., Cary and Chapman 1988; Khan et al. 2000; Khan and Mosegaard 2001). This method is suitable for problems with misfit functionals having multiple local minimums, where conventional gradient-type minimization methods may have difficulties getting out from a "deep" local minimum (see Chap. 7). The second type of Monte Carlo method uses an optimization algorithm in order to minimize the number of steps required by the random search methods. The most effective global optimization algorithms have been developed based on known physical or biological rules to evolve to the best solution. For example, the simulated annealing (SA) algorithm (Kirkpatrick et al. 1983; Corana et al. 1987) comes from annealing in metallurgy, a technique involving heating and controlled cooling of a material. It is known from physics that, in order to minimize the final lattice energy, one should apply a very slow cooling process. The SA method uses an analogy between the minimization of lattice energy in the framework of the physical process of annealing and numerical problem of determining the global minimum of a misfit functional, $f(\mathbf{m})$.

The genetic algorithm (GA) (Holland 1975; Goldberg 1989; Michalewicz and Schoenauer 1996; Whitley 1994; Mosegaard and Sambridge 2002) is a heuristic search method that mimics the process of natural evolution. In a pure genetic algorithm, a population of candidate solutions (individuals) for an optimization problem is evolved toward better solutions. Traditionally, the solutions are coded in binary form as strings of 0s and 1s to be mutated and altered. The evolution starts from a population of randomly generated solutions from the search space and proceeds as an iterative process. The population in each iteration is called a *generation*. In each generation, the fitness of every individual is evaluated by an objective functional (e.g., a misfit functional $f(\mathbf{m})$). The individuals who have low misfits are stochastically selected from the current population, and then they are chosen to form a new generation by applying genetic operations (*mutation* and *crossover*). The above steps run iteratively until the inversion process meets the termination conditions.

A detailed overview of the simulated annealing and genetic algorithms can be found, for example, in Zhdanov (2015).

The Monte Carlo methods are considered to be an effective optimization technique for many inverse problems where some general gradient-type methods fail. They can be applied for solving optimization problems with continuous or discrete parameters and with small sample intervals; there is no need to calculate the derivatives; the global minimization problem can be solved for the misfit functional with multiple local minima.

The Monte Carlo methods were first applied to the solutions of earth science problems by Keilis-Borok and Yanovskaya (1967) and Press (1968, 1970a, b). The paper by Sambridge and Mosegaard (2002) provides an excellent review of applications of the Monte Carlo methods to solving geophysical inverse problems.

References and Recommended Reading to This Chapter

Backus GE (1970a) Inference from inadequate and inaccurate data. I Proceedings of the National Academy of Sciences, vol 65, pp 1–7

Backus GE (1970b) Inference from inadequate and inaccurate data, II. Proceedings of the National Academy of Sciences, vol 65, pp 281–287

Backus, GE (1970c) Inference from inadequate and inaccurate data, III: Proceedings of the National Academy of Sciences, vol 67, pp 282–289

Backus GE, Gilbert TI (1967) Numerical applications of a formalism for geophysical inverse problems. Geophys J R Astr Soc 13:247–276

Cary PW, Chapman CH (1988) Automatic 1D waveform inversion of marine seismic refraction data. Geophys J 93:527–46

Corana A, Marchesi M, Martini C, Ridella S (1987) Minimising multimodal functions of continuous variables with the "Simulated Annealing" Algorithm. ACM Trans Math Soft 13:262–280

Foster M (1961) An application of the Wiener-Kolmogorov smoothing theory to matrix inversion. J Soc Ind Appl Math 9:387–392

Franklin JN (1970) Well-posed stochastic extensions of ill-posed linear problems. J Math Anal Appl 31:682–716

Goldberg DE (1989) Genetic Algorithms in search, optimization, and machine learning. Addison-Wesley

Holland JH (1975) Adaptation in natural and artificial systems. Ann Arbor: University of Michigan Press

Jackson DD (1972) Interpretation of inaccurate, insufficient and inconsistent data: Geophys J R Astronom Soc 28:97–110

Keilis-Borok VI, TB Yanovskaya (1967) Inverse problems of seismology. Geophys J 13:223–234

Khan A, Mosegaard K, Rasmussen KL (2000) A new seismic velocity model for the Moon from a Monte Carlo inversion of the Apollo Lunar seismic data. Geophys Res Lett 27:1591–1594

Khan A, Mosegaard K (2001) New information on the deep lunar interior from an inversion of lunar free oscillation periods. Geophys Res Lett 28:1791

Kirkpatrick, S. C., D. Gelatt and M. P. Vecchi, 1983, Optimization by simulated annealing: Science, 220, 671–680 Lanczos C (1961) Linear differential operators. D van Nostrand Co

Marquardt DW (1963) An algorithm for least-squares estimation of nonlinear parameters. J Soc Ind Appl Math 11:431–441

Marquardt DW (1970) Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. Technometrics 12:591–612

Menke W (2018) Geophysical data analysis: discrete inverse theory, 4th ed. Academic Press, 330 pp

Metropolis N, Ulam SM (1949) The Monte Carlo method. J Am Stat Assoc **44**:335–341

Metropolis N, Rosenbluth MN, Rosenbluth AW, Teller, Teller E (1953) Equation of state calculations by fast computing machines. J Chem Phys 21:1087–1092

Michalewicz Z, Schoenauer M (1996) Evolutionary algorithms for constrained parameter optimization problems. Evol Comput 4(1):1–32

Mosegaard K, Sambridge M (2002) Monte Carlo analysis of inverse problems. Inverse Probl 18:R29–R54

Press F 1968 Earth models obtained by Monte Carlo inversion. J Geophys Res 73:5223–34

Press F (1970a) Earth models consistent with geophysical data. Phys Earth Planet Inter 3:3–22

Press F (1970b) Regionalized Earth models. J Geophys Res 75:6575–81.

Sabatier PC (1977) On geophysical inverse problems and constraints. J Geophys Res 43:115–137

Sambridge, M, Mosegaard K (2002) Monte Carlo methods in geophysical inverse problems. Rev Geophys 40, 3:1–29

Tarantola A (1987) Inverse problem theory. Elsevier, Amsterdam, Oxford, New York, Tokyo, 613 pp

Tarantola A (2005) Inverse problem theory and methods for model parameter estimation. SIAM, 344 pp

Tarantola A, Valette B (1982) Generalized nonlinear inverse problem solved using the least squares criterion. Rev Geophys Space Phys 20:219–232

Tikhonov AN, Arsenin VY (1977) Solution of ill-posed problems. W H Winston and Sons

Whitley DL (1994) A genetic algorithm tutorial. Stat Comput 4:65–85

Zhdanov MS (1993) Tutorial: regularization in inversion theory. CWP-136, Colorado School of Mines, 47 pp

Zhdanov MS (2002) Geophysical inverse theory and regularization problems. Elsevier, 609 pp

Zhdanov MS (2015) Inverse theory and applications in geophysics. Elsevier, 704 pp