Oh, S., Lee, D. K., Yang, J., Youn, Y. H., Geostatistics for Bayesian interpretation of geophysical data.
International Symposium on the Fusion Technology of Geosystem Engineering, Rock Engineering and Geophysical Exploration
Seoul, Korea, Nov. 18-19, 2003

Geostatistics for Bayesian interpretation of geophysical data

Seokhoon Oh, Duk Kee Lee, Junmo Yang, Yong-Hoon Youn
Meteorological Research Institute, Seoul, Korea

Abstract: This study presents a practical procedure for the Bayesian inversion of geophysical data by Markov chain Monte Carlo (MCMC) sampling and geostatistics. We have applied geostatistical techniques for the acquisition of prior model information, and then the MCMC method was adopted to infer the characteristics of the marginal distributions of model parameters. For the Bayesian inversion of dipole-dipole array resistivity data, we have used the indicator kriging and simulation techniques to generate cumulative density functions from Schlumberger array resistivity data and well logging data, and obtained prior information by cokriging and simulations from covariogram models. The indicator approach makes it possible to incorporate non-parametric information into the probabilistic density function. We have also adopted the MCMC approach, based on Gibbs sampling, to examine the characteristics of a posteriori probability density function and the marginal distribution of each parameter. This approach provides an effective way to treat Bayesian inversion of geophysical data and reduce the non-uniqueness by incorporating various prior information.

1. Prior information by geostatistical methods

Geostatistics is largely based upon the random function model, whereby the set of unknown values is regarded as a set of spatially dependent random variables. Such presentation reflects our imperfect knowledge of the unsampled value *z*(u) and more generally, of the distribution of *z* within the area (Goovaerts, 1997). Our ultimate goal is to get prior information from various geophysical data and to infer the uncertainty structure of an interpreted region.

Inference of uncertainties of blocks

The points where the geostatistical simulations may estimate are denser than the number of the parameterized blocks for inversion. Therefore, by assessing the values on denser points, we derive values by following indirect method.

We consider the problem of evaluating the block ccdf (conditional cumulative density function) $F_V(\mathrm{u};z|(n))$ that models the uncertainty about an average z-value over the block *V*(u) conditional to (*n*) neighborhood data:

$$F_V(\mathrm{u};z|(n)) = \mathrm{Prob}\{\mathrm{u} \le z|(n)\} \tag{1}$$

Because of the non-linearity of the indicator transform, the block ccdf cannot be derived simply as a linear combination of point ccdfs:

$$[F_V(\mathrm{u};z|(n))]^* \neq \frac{1}{J}\sum_{j=1}^{J}[F(\mathrm{u}'_j;z|(n))]^* \tag{2}$$

with the point-ccdf $F_V(\mathrm{u}'_j;z|(n))$ being defined at *J* points u'_j discretizing the block *V*(u). $[F_V(\mathrm{u};z|(n))]^*$ is the inference of ccdf for the block *V* by conditional to *n* data around the point u. In the absence of block data $z_V(\mathrm{u}_\alpha)$ and corresponding block statistics and block indicator data, the block ccdf (eq.1) can be numerically approximated by the cumulative distribution of many simulated block values $z_V^{(l)}(\mathrm{u})$ (Isaaks, 1990; Gomez-Hernandez, 1991; Deutsch and Journel, 1992; Glacken, 1996):

$$[F_V(\mathrm{u};z|(n))]^* = \frac{1}{L}\sum_{l=1}^{L} i_V^{(l)}(\mathrm{u};z) \quad l = 1,\ldots,L \tag{3}$$

with the block indicator value defined $i_V^{(l)}(\mathrm{u};z) = 1$ if $z_V^{(l)}(\mathrm{u}) \le z$,and zero otherwise. Each simulated block value $z_V^{(l)}$ is obtained by averaging a set of z-values simulated at *J* points u'_j discretizing the block *V*(u) :

$$z_V^{(l)}(\mathrm{u}) = \frac{1}{J}\sum_{j=1}^{J} z^{(l)}(\mathrm{u}'_j). \tag{4}$$

Incorporation of DC and logging data as prior information

For the Bayesian inversion of dipole-dipole array DC data, we have deduced the prior information from Schlumberger and well logging data by geostatistical simulation of which schematic process is shown in Fig.1. Indicator covariogram modelings were performed for each threshold with resistivity sounding and well logging data, then on unsampled points, where the density of estimation is finer than the parameterized blocks, sequential indicator joint simulation was conducted. Many realizations were provided for assessing the uncertainty of the studied region. As described in previous section, the simulated estimations now are used in the block CDF making process.

Because the resolution of each data is different, careful variogram modeling is required, but indicator transforming makes the steps more stable by data-classified level.

2. MCMC approach to solution

Tarantola (1987) showed that the complete solution of geophysical inverse problem is posterior information given by the conjunction of the two states of information, prior and theoretical or likelihood information, and this is an accordant form to the generalized Bayes' theorem. That is,

$$\sigma(\mathrm{m}) = k\rho(\mathrm{m})L(\mathrm{m})$$

where, $\sigma(\mathrm{m})$ is the posterior PDF, $\rho(\mathrm{m})$ is the prior PDF, and $L(\mathrm{m})$ is a likelihood PDF. Thus the simple multiplicity of prior PDF and likelihood PDF is the solution to the geophysical inverse problem in the viewpoint of probabilistic approach. But the high-multidimensional PDF itself gives no choice, and when we are determining the most probable model for consulting or inferring the characteristics of the neighborhood of maximum posterior PDF

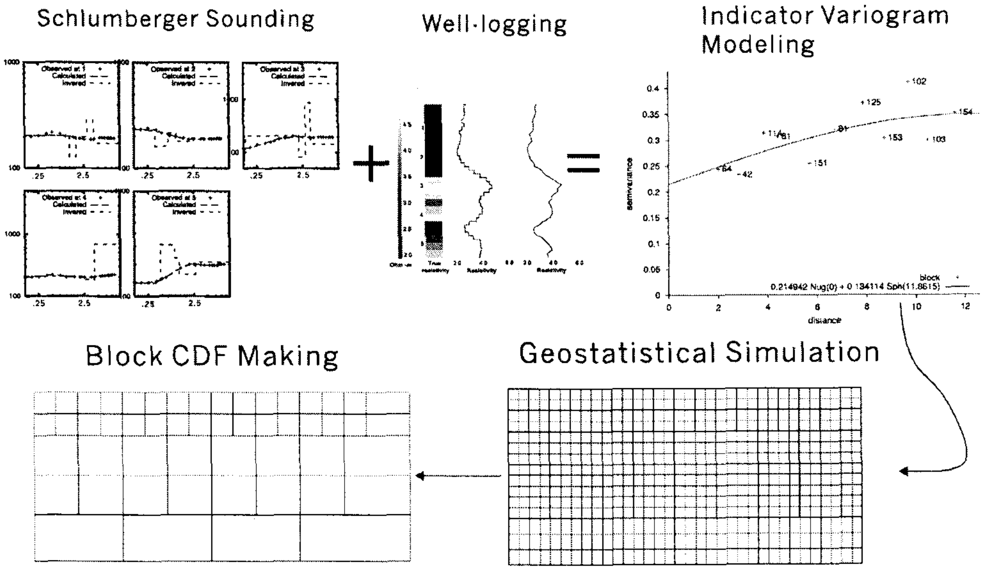

Fig. 1. Schematic process to generate prior information. Schlumberger array sounding and resistivity well logging data are put in covariogram modeling for a specified threshold for indicator kriging or simulation. Variogram model shows the data are rough from the high nugget value (Deutsch and Journel, 1992) but they have correlation and sustain a tendency. Then denser estimations contribute to the block CDF making, because it is impossible to estimate the resistivity of each block directly. So indirect estimation for blocks are made by the method proposed in previous section.

solving a high dimensional integral or an optimization problem called marginalization process is required. Unfortunately, the kernel of the integral is very highly nonlinear, so the analytic solution is very rare and we should attack it by numerical analysis. This study approaches the marginalization process by the Markov Chain Monte Carlo method based on the Gibbs sampler that utilizes the modified SA algorithm.

Sampling from prior information

To sample a candidate from the prior information for each parameterized block, cumulative density functions for the prior PDF were generated and interpolated between the discrete points. Here, the more sample points that decide the degree of discretization are selected, the better the inverse sampling from the CDF is performed. However, the maximum level of discretization is constrained by the number of the conditional data and in the indicator case, the order relaxation (Isaaks and Srivastava, 1989) degrades the information for excessive discretization. We performed 500 simulations and dicretized the PDF by 20 thresholds, and then random numbers between 0 and 1 were used to get sampled values.

Modified Gibbs sampling algorithm

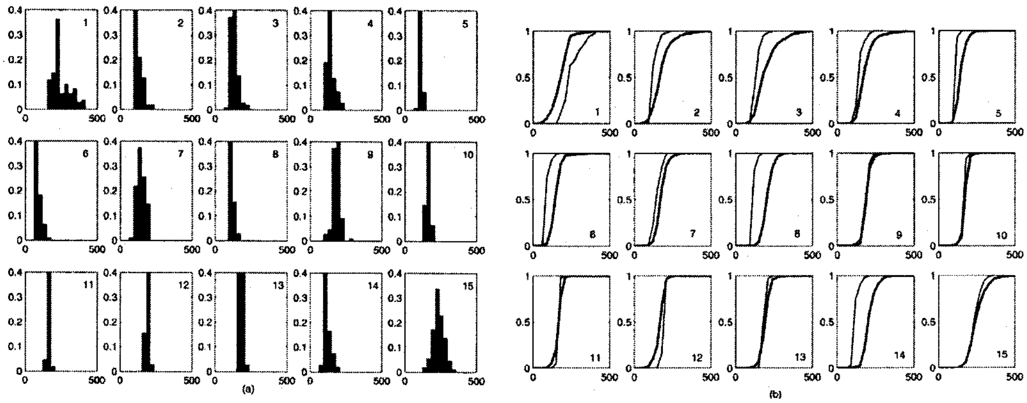

Fig. 2. (a) Posterior information given by probability density functions. The horizontal axes mean resistivity in ohm-m, and the vertical axes present the probability. (b) Cumulative Density functions for prior information (bold line) and posterior information (thin line) for upper blocks. The posterior CDF is concentrated on specific region representing low uncertainties. The vertical axes present CDF.

To overcome the problems of the previous MCMC sampling, we approached the sampler with the prior information. Actually, the full conditioning distribution in Gibbs sampling means to select the most probable parameter conditional on remaining parameters. Therefore when we have a marginalized prior PDF, we can apply it to the full conditional distribution. That is, the prior cumulative density function (CDF) is created by the prior PDF, then candidate sampling is conducted by the inverse CDF. Then, a component that updates the likelihood function is accepted and if not, the component is accepted with the probability,

$$P_{accept} = \exp\left(- \frac{(E_{accepted} - E_{current})}{T} \right)$$

where, E is an objective function generally given by discrepancy between calculated and observed data and T is a tuning parameter referred to as temperature in Simulated Annealing (Sen and Stoffa, 1995). And if the component is rejected, sampling is again conducted on the same CDF until the maximum number of allowances is reached This approach effectively samples the neighbourhood of the maximum posterior PDFs that fit the prior and posterior PDF after burn-in trials. Unlike the Rothman's heat bath algorithm (1986), ours prepares the model perturbation in advance and updates the component that fits the prior and posterior PDFs. This makes the process remarkably effective especially when the prior information reflect the likelihood functions well, though in the worst case searching the global model space should be guaranteed.

Marginal distribution of posterior information

Fig.2 (a), (b) present comparisons of prior and posterior PDFs produced by the above Gibbs sampling MCMC algorithm for the upper 15 blocks. The core of the Bayesian inference is based on the point of how we can reduce the uncertainty in posterior PDF as well as the error of discrepancy, because it is important to reduce the uncertainty in observed data by prior information. The update of the upper blocks shown in Fig.2 is remarkable in terms of reduction of uncertainty. But deeper blocks do not have significant reduction of uncertainty. These results suggest that blocks having low sensitivity from observed data tend to converge to prior information. But we can find that the convergence is stable, so we need not have any constraints for the stable convergence of inversion.

From these marginal distributions, various informations may be available, such as the confidence interval, mean of resistivity, etc., which were mentioned in numerous previous studies without any explanations of how to get the distribution. And posterior covariance and resolution matrix may be also obtained by indirect sampling of posterior PDF.

References

Deutsch, C. V. and Journel, A. G., 1992, GSLIB: Geostatistical Software Library and User's Guide: Oxford University Press, New York.

Glacken, I., 1996, Change of support by direct conditional block simulation: Master's thesis, Stanford Univ., Stanford, CA.

Gomez-Hernandez, J., A., 1991, A stochastic approach to the simulation of block conductivity fields conditional upon data measured at a smaller scale: Doctoral Dissertation, Stanford University, Stanford, CA.

Goovaerts, P., 1997, Geostatistics for Natural Resources Evaluation: Oxford University Press, New York.

Isaaks E. H., 1990, The application of Monte Carlo methods to the analysis of spatially correlated Data: Doctoral Dissertation, Stanford University, Stanford, CA.

Isaaks E. H., and Srivastava R. M., 1989, An Introduction to Applied Geostatistics: Oxford University Press, New York.

Rothman, D. H., 1986, Automatic estimation of large statics corrections, Geophysics, 51, 332-346.

Sen, M., and P. L. Stoffa, 1995, Global optimization methods in geophysical inversion, Advances in Exploration Geophysics, v.4, Elsevier, Amsterdam.

Tarantola, A., 1987, Inverse problem theory: Methods for data fitting and model parameter estimation, Elsevier