Tu SBT4 12

Most Frequent Value Based Factor Analysis of Engineering Geophysical Sounding Logs

N.P. Szabo* (MTA-ME Geoeng. Res. Group, Univ. of Miskolc) & G.P. Balogh (University of Miskolc)

SUMMARY

A multivariate statistical approach is presented to estimate water saturation in shallow heterogeneous formations. An improved factor analysis algorithm is developed to process engineering geophysical sounding data in a more reliable way. Resistivity and nuclear data acquired by cone penetration tools equipped with geophysical sensors are processed simultaneously to give an estimate to factor logs. The new factor analysis procedure is based on the iterative reweighting of data prediction errors using the highly robust most frequent value method, which improves the accuracy of factor scores in case of non-Gaussian data sets. A strong exponential relationship is detected between water saturation and the first factor log. Tests made on penetration logs measured from a Hungarian well demonstrate the feasibility of the most frequent value based factor analysis approach, which is verified by the results of local inverse modeling.

Introduction

Cone penetration tests complemented by geophysical measurements can be effectively used for the in-situ investigation of shallow sedimentary structures. Engineering geophysical sounding (EGS) data collected by nuclear and electric sondes provide information on the composition, water and gas saturation and geotechnical parameters of subsoils (Fejes and Jósa, 1990). Analogously to well-logging methods, several types of physical parameters can be observed in penetration holes with a specialty that only a steel tube isolates the probe from the soil, there is no invasion of drilling fluid into the formations, and the measured data are transferred through the rods pushed into the ground. In practice, data processing mostly incorporates deterministic or inversion methods adapted from well-log analysis (Drahos, 2005).

Factor analysis is applicable to reduce the dimensionality of statistical problems and extract not directly measurable information from the sample (Lawley and Maxwell, 1962). Factor analysis of EGS data allows the estimation of water saturation and the simulation of neutron log to unmeasured intervals (Szabó *et al*., 2012). Jöreskog (2007) suggested a non-iterative factor analysis technique, which gives optimal results for Gaussian distributed data. The above method works as a relatively noise sensitive statistical procedure in dissimilar cases. The algorithm of classical factor analysis has been further developed for giving a robust solution, which is based on the iterative improvement of the deviation between the measured and calculated data (prediction error). The core of the method is the use of an iterative re-weighting process using the most frequent value (MFV) method (Steiner, 1991). During the procedure of factor analysis, optimal weights are calculated for each component of the vector of prediction errors to estimate the logs of the extracted statistical variables (i.e. factors) more accurately. Szegedi and Dobróka (2014) applied the Steiner's weights for the establishment of a robust inversion-based Fourier transformation method, which showed high noise rejection capability. In this study, we test the most frequent value based factor analysis procedure on real EGS data and the result of it is compared to that of independent inverse modeling.

Local inversion of EGS data

The matrix of shallow sediments is composed of coarse and fine grain components, while their pore space is saturated with freshwater and/or gas (normally air). The model vector of inversion defined in a given depth is $\mathbf{m}=[V_{cl},V_s,V_w,V_g]^T$ (T is symbol of transpose), which includes the fractional volumes of clay, sand, water and gas. Water saturation quantifies the pore volume occupied by water

$$S_w = V_w/(V_w + V_g)\,. \tag{1}$$

The following data types are normally measured by EGS tools such as natural gamma-ray intensity (*GR*), density (*DEN*), neutron-porosity (*NPHI*) and resistivity (*RES*). The petrophysical model of freshwater formations is related to EGS data by the undermentioned probe response functions

$$GR = V_{cl}GR_{cl} + V_sGR_s\,, \tag{2}$$

$$DEN = V_w\rho_w + V_{cl}\,\rho_{cl} + V_s\rho_s\,, \tag{3}$$

$$NPHI = V_w\Phi_{N,w} + V_{cl}\Phi_{N,cl} + V_s\Phi_{N,s}\,, \tag{4}$$

$$RES = a\left(V_w + V_g + V_{cl}\right)^{-m}\left(\frac{V_{cl}/(V_w+V_{cl})}{R_{cl}} + \frac{1-[V_{cl}/(V_w+V_{cl})]}{R_w}\right)^{-1}\left(\frac{V_w+V_{cl}}{V_w+V_g+V_{cl}}\right)^{-n}, \tag{5}$$

where the physical constants of rock constituents and pore-filling fluids are indicated by *cl* (clay), *s* (sand), *w* (water), *g* (gas), ρ denotes density, Φ_N is neutron porosity, and parameters *m*, *a*, *n* represent the cementation exponent, tortuosity factor and saturation exponent, respectively. The column vector of observed data is $\mathbf{d}=[GR,DEN,NPHI,RES]^T$. A set of local inverse problems is to be solved depth-by-depth. The objective function is chosen as the Euclidean norm of the difference between the measured and calculated data. Since the EGS data are of different magnitudes, each deviation is

normalized by the measured data. The inverse problem is overdetermined; therefore, a stable inversion procedure can be achieved by using the Gaussian least squares method (Turai, 2011). By the inversion process, the components of vector **m** with their standard deviations are estimated.

Procedure of robust factor analysis

The input of factor analysis is the *N*-by-*K* matrix of standardized EGS data denoted by **D**, which is decomposed into two terms

$$\mathbf{D} = \mathbf{F}\mathbf{L}^{\mathrm{T}} + \mathbf{E}, \tag{6}$$

where **F** is the *N*-by-*M* matrix of factor scores, **L** is the *K*-by-*M* matrix of factor loadings and **E** is the matrix of residuals (*M* is the number of extracted factors, *N* is the number of sampled depths and *K* is the number of observed variables). The factor loadings and the factor scores are normally estimated simultaneously by the maximum likelihood method (MLM). Jöreskog (2007) proposed a fast non-iterative algorithm for calculating the factor loadings followed by an MLM estimation for the factor scores, which is optimal only for normally distributed data. To improve the method of Jöreskog, we properly modify the classical model of factor analysis defined in Eq. (6)

$$\mathbf{d} = \widetilde{\mathbf{L}}\mathbf{f} + \mathbf{e}, \tag{7}$$

where **d** denotes the *KN* column vector of input data, $\widetilde{\mathbf{L}}$ is the *NK*-by-*NM* matrix of factor loadings, **f** is the *MN* length vector of factor scores, **e** is the *KN* length vector of prediction error. In the first step of the procedure, we give an estimate to the initial values of factor loadings and scores by Jöreskog's method. Then, we apply an iterative algorithm, which takes the form in the *q*-th iteration step as

$$\mathbf{L}^{\mathrm{T}(q)} = \left(\mathbf{F}^{\mathrm{T}(q-1)}\mathbf{F}^{(q-1)} + \alpha^2\mathbf{I}\right)^{-1}\mathbf{F}^{\mathrm{T}(q-1)}\mathbf{D}, \tag{8}$$

$$\mathbf{f}^{(q)} = \left(\widetilde{\mathbf{L}}^{\mathrm{T}(q-1)}\mathbf{W}\widetilde{\mathbf{L}}^{(q-1)}\right)^{-1}\widetilde{\mathbf{L}}^{\mathrm{T}(q-1)}\mathbf{W}\mathbf{d}, \tag{9}$$

where α is a properly chosen damping factor. The elements of the *NK*-by-*NK* diagonal weighting matrix **W** are proportional to the deviation of the measured (**d**) and calculated data ($\widetilde{\mathbf{L}}\mathbf{f}$)

$$\mathrm{W}_{\mathrm{ii}} = \frac{\varepsilon^2}{\varepsilon^2 + (\mathrm{e}_{\mathrm{i}})^2} \qquad (\mathrm{i} = 1,2,...\mathrm{KN}), \tag{10}$$

where parameter ε called dihesion is automatically calculated by the MFV method (Steiner, 1991). According to Eq. (10), the larger the distance between the observed and predicted data, the less weight given to the relevant datum. The above iterative factor analysis procedure is named MFV-FA method.

Comparative study

The MFV-FA procedure is tested in a penetration hole drilled in Bátaapáti, south-west Hungary. The following EGS logs were measured in a shallow sedimentary complex: cone resistance (*RCPT*), natural gamma-ray intensity (*GR*), density (*DEN*), neutron-porosity (*NPHI*) and resistivity (*RES*). The values of zone parameters are fixed after Szabó *et al*. (2012). The MFV-FA procedure improves the results of Jöreskog's algorithm using Eqs. (8)–(10). The relative distance between the measured and calculated data decreases gradually with the number of iterations. The factor loadings and factor scores are updated in 15 iteration steps. In addition, in each step of the iterative procedure the Steiner weights are recalculated in further 30 steps. In this inner loop, dihesion is automatically decreased and changed differently for each EGS log (Figure 1-a). Beside the same value of ε, the larger the distance between the measured and calculated data, the smaller the weight. With the decrease of ε, bigger deviations contribute less to the solution. The optimal values of diagonal elements of the weighting matrix **W** are shown in Figure 1-b, where the weighting coefficients are represented as a function of the prediction error.

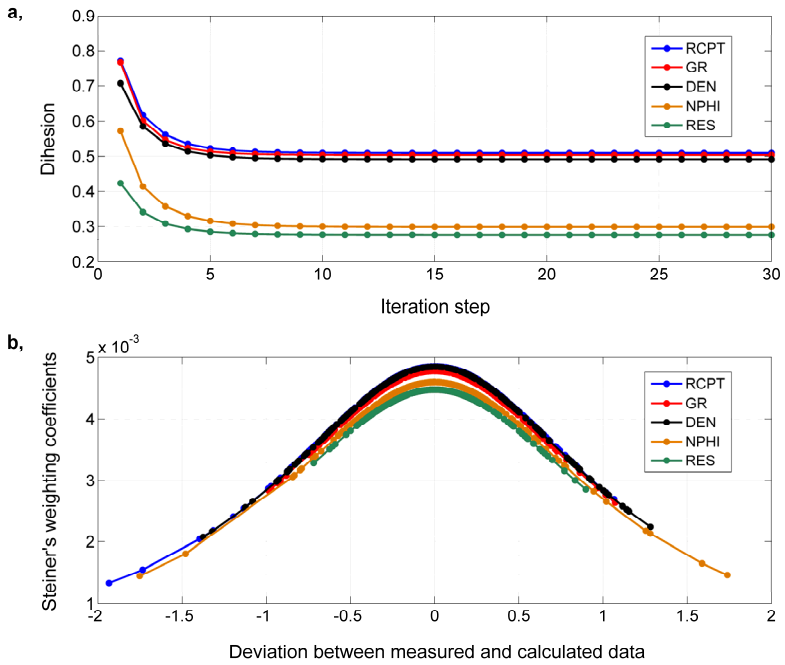

Figure 1 *Development of convergence of dihesion during the optimization of Steiner weights (a), Steiner weights applied to EGS data in the MFV-FA procedure (b).*

Two independent factors are calculated by the MFV-FA procedure. The first one explains 76.2 % of the total variance of the EGS data the loadings of which are –0.41 (*RCPT*), 0.31 (*GR*), 0.88 (*DEN*), 0.93 (*NPHI*), –0.98 (*RES*). The strongest correlation is indicated between the first factor and water saturation-sensitive logs (*NPHI, RES*). The first factor (F_1) is connected to water saturation (S_w) in regression analysis (Figure 2-a). The regression coefficients of the exponential relation are estimated with their 95 % confidence bounds (a=0.40±0.04, b=–0.33±0.03, c=0.20±0.04). The Pearson's correlation coefficient (R) between the first factor and water saturation shows strong correlation.

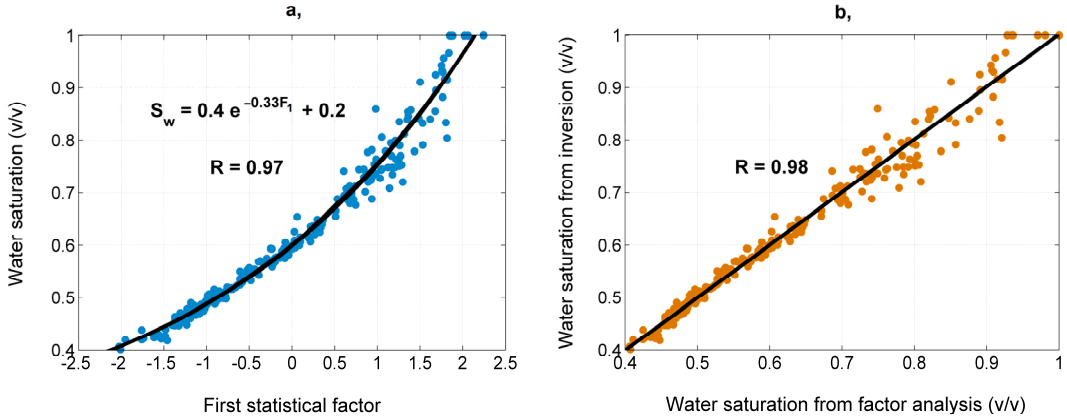

Figure 2 *Regression relation between the first factor and water saturation (a), connection between water saturations estimated separately by the MFV-FA procedure and local inverse modeling (b).*

Those of inverse modeling confirm the results of the MFV-FA procedure (Figure 2-b). The forward problem is solved by Eqs. (2)–(5), then a set of local inverse problems are solved in adjacent depths. The measured and calculated EGS logs show a good agreement (tracks 1-4 in Figure 3), where the average RMS between the observed and predicted data (distinguished by suffix TH) is 3.84 %. The estimated values of volumetric parameters in vector **m** is plotted in the last track. Their mean estimation error is 2.1 volume percent. The two factor logs are given in track 5, while the water

saturation logs estimated by robust factor analysis (*SW_MFV-FA*) and inverse modeling (*SW_INV*) are in track 6. The average RMS between the water saturation logs is 2.43 %.

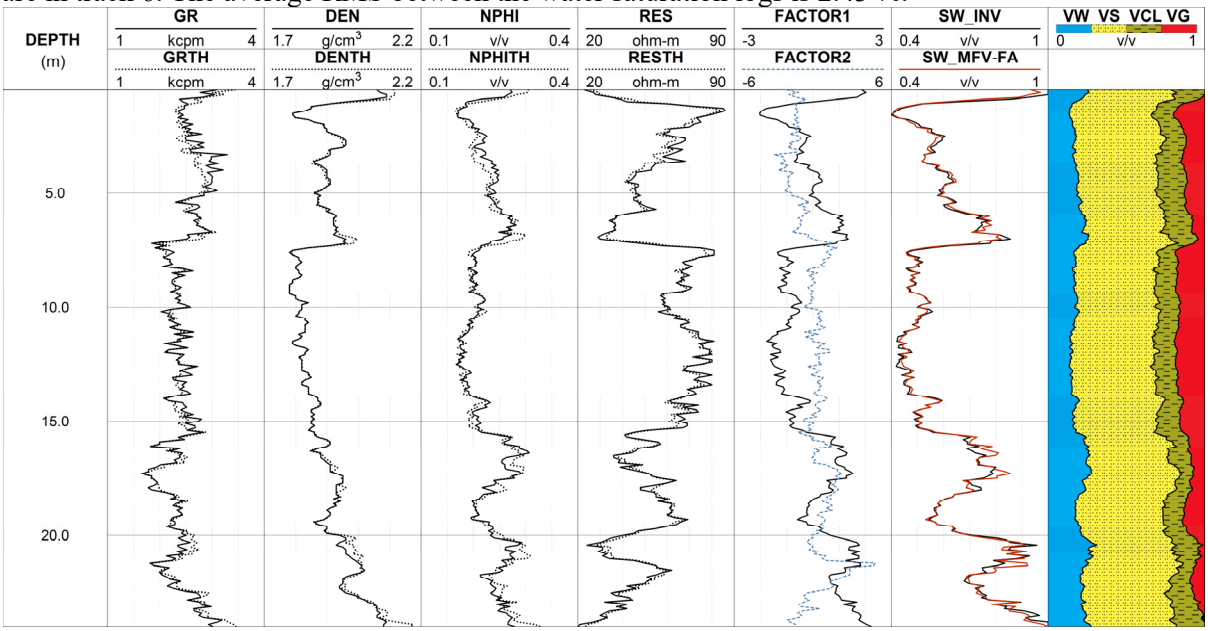

Figure 3 *Results of MFV-FA processing and local inversion of EGS data in a Hungarian well.*

Conclusions

An improved factor analysis algorithm is developed and applied to EGS data. As a result, the well logs of factor scores are estimated in a robust way. The strong correlation between the first factor and water saturation allows a more reliable estimation of water content and other derived quantities such as air saturation and dry density. The suggested method provides the petrophysical modeling of heterogeneous formations with in-situ information for solving engineering/environmental problems.

Acknowledgements

The first author as the leading researcher of OTKA project PD-109408 thanks to the support of the Hungarian Scientific Research Fund. Special thanks go to Dezső Drahos (Eötvös Loránd University), Prof. Mihály Dobróka (University of Miskolc), János Stickel (Elgoscar-2000 Ltd.) for cooperation.

References

Drahos, D. [2005] Inversion of engineering geophysical penetration sounding logs measured along a profile. *Acta Geodetica et Geophysica Hungarica*, **40**, 193-202.

Fejes, I., Jósa, E. [1990] The engineering geophysical sounding method. Principles, instrumentation, and computerised interpretation. In: S.H. Ward (Ed.) *Geotechnical and environmental geophysics*, 2, Environmental and groundwater: SEG, 321-331.

Jöreskog, K.G. [2007] Factor analysis and its extensions. In: R. Cudeck and R.C. MacCallum (Eds) *Factor analysis at 100, historical developments and future directions.* Lawrence Erlbaum Associates, Publishers, 47-77.

Lawley, D.N., Maxwell A.E. [1962] Factor analysis as a statistical method. *The Statistician*, **12**, 209-229.

Steiner F. [1991] *The most frequent value. Introduction to a modern conception of statistics*. Academic Press, Budapest.

Szabó N.P., Dobróka M., Drahos D. [2012] Factor analysis of engineering geophysical sounding data for water saturation estimation in shallow formations. *Geophysics*, **77**, WA35-WA44.

Szegedi H. and Dobróka M. [2014] On the use of Steiner's weights in inversion-based Fourier transformation: robustification of a previously published algorithm. *Acta Geodaetica et Geophysica*, **49**, 95-104.

Turai, E. [2011] Data processing method developments using tau-transformation of time domain IP data (II) - Interpretation results of field measured data. *Acta Geodaetica et Geophysica*, **46**, 391-400.