Robust estimation of reservoir shaliness by iteratively reweighted factor analysis

Norbert Péter Szabó[1,2], Mihály Dobróka[1]

Right running head: Iteratively reweighted factor analysis

[1]University of Miskolc, Department of Geophysics, Miskolc, Hungary.

[2]MTA-ME, Geoengineering Research Group, Miskolc, Hungary

E-mail: norbert.szabo.phd@gmail.com, dobroka@uni-miskolc.hu

Date of submission: 18 November 2016.

ABSTRACT

We suggest a statistical method for the simultaneous processing of electric, nuclear and sonic logging data using a robust iteratively reweighted factor analysis. After giving a first estimate by Jöreskog's approximate method, we refine the factor loadings and factor scores jointly in an iterative procedure, during which the deviation between the measured and calculated data is weighted in proportion to its magnitude for giving an outlier-free solution. We show a strong nonlinear relation between the first factor and shale volume of multi-mineral hydrocarbon formations. We test the noise rejection capability of the new statistical procedure by making synthetic modeling experiments. The iteratively reweighted factor analysis of simulated well-logging data including high amount of noise gives the well log of shale volume purified from large errors. Case studies from Hungary and the USA show that the results of factor analysis are consistent with that of independent deterministic modeling and core data. The statistical workflow can be effectively used for the processing of not normally distributed and extremely noisy well-logging data sets to evaluate the shale content and derived petrophysical properties more accurately in reservoir rocks.

INTRODUCTION

In order to improve the efficiency of modern hydrocarbon exploration, sophisticated and robust geophysical data processing methods are being currently developed. In well log analysis, multivariate statistical approaches such as factor analysis seem to be a powerful tool in the evaluation of hydrocarbon reservoirs. Factor analysis is generally applied to reduce the dimensionality of multivariate statistical problems (Lawley and Maxwell, 1962). Additionally, in geosciences, it allows the exploration of latent variables dependent upon lithological characteristics and petrophysical properties of rocks not directly measurable by geophysical instruments. This capacity has been recently utilized in mineral exploration (da Silva Pereira et al., 2010; Qian et al., 2011; Merdith and Müller, 2015), applied geochemistry (Sun et al., 2009; Asadi et al., 2014; Mongelli et al., 2014) and hydrogeological studies (Charfi et al., 2013; Derby et al., 2013; Szabó, 2015; Krogulec and Zabłocki, 2015). The idea of applying factor analysis for the interpretation of wireline logs dates back to the 1980's. Buoro and Silva (1994) published a factor analysis-based technique for studying the ambiguity in the inversion of well-logging data. By detecting the most ambiguous parameters, they increased the stability and uniqueness of the inverse problem. Goncalves et al. (1995) presented the statistical method as an effective pre-processing tool for lithofacies identification and classification. Factor analysis was first applied to estimate the shaliness of clastic formations by Szabó (2011). Shale volume as a key parameter in formation evaluation was related to the first factor by regression analysis. The exponential connection between the above variables has proved to be valid in several wells in Hungary and the USA (Szabó and Dobróka, 2013). Asfahani (2014) successfully applied the same technique to lithology identification in a basaltic area of Southern Syria. Seth et al. (2015) estimated the shale volume of siliciclastic and diatomic sediments of the Bering Sea. Factor analysis was applied to direct-push logs for water saturation estimation in shallow formations and simulation of neutron-porosity data to missing depth intervals (Szabó et al., 2012). Dry density of soils as an important geotechnical parameter was also derived by factor analysis for 2D case by Szabó (2012). Factor analysis was recently applied to the lithologic characterization of Paleozoic heterogeneous shale gas sediments (Wawrzyniak-Guz et al., 2016).

The quality of estimated parameters is highly dependent on the level and distribution of measurement noise. Thence, statistical method-based interpretation techniques should pay attention to the proper handling of data uncertainty. The asymmetry of the distribution of data noise or the presence of outliers, caused by the logging instruments or caverns and other borehole irregularities, may have great impact on the accuracy of the estimation results. One can easily exclude outliers from the analysis by removing them from the data set.

However, it is not always recommendable, because they may carry useful information, e.g. cycle skipping in acoustic logging is indicative of pore-filling gas or fractures, or anomalous well log readings may be linked to thin layers or rare minerals. Outliers associated to lithological/petrophysical variations are favorable to be weighted in accordance with their relative importance in the interpretation procedure. In factor analysis, the observed variables are developed as a linear combination of new statistical variables called factors, which explain the variance of the observed information in different amounts. The classical method of Bartlett (1937) based on the hypothesis of linearity gives a fast and optimal solution for Gaussian distributed measurement data. For other types of distributions, this procedure is rather sensitive to data errors and outliers. Although, in practice, the maximum likelihood method has been found to be highly efficient against the departures from normality (Jöreskog, 2007), several attempts have been made for the robustification of factor analysis. Croux et al. (1999) solved the problem of robust factor analysis with a regression technique using the criterion of least absolute deviations, which was especially useful for dealing with missing and outlying data. Pison et al. (2003) suggested a robust estimation of the data covariance matrix and a subsequent application of the maximum likelihood method for extracting factors. This technique has been successfully applied for compositional data analysis by Filzmoser et al. (2009), and Hoseinpoor and Aryafar (2014). Luttinen et al. (2012) modelled the data noise using Student's t-distribution and evaluated the posterior probability density function by Bayesian approximate methods to introduce a new probabilistic model for robust factor analysis. In this study, we offer a different approach for the improvement of factor analysis using an iterative reweighting process adapted from geophysical inverse theory.

In well-logging practice, the physical quantities are observed with different accuracies. Therefore, inversion methods normally minimize a weighted norm of the deviation between the observed and predicted data normalized by the standard deviation of data (Mayer and Sibbit, 1980). Because of the propagation of errors, data variances can also be employed to derive the estimation errors of model parameters (Menke, 2012). Analogously to inverse problems, in factor analysis, it is of high importance to distinguish the data by their uncertainty. The traditional methods of factor analysis generally apply weighting on the specific variances, representing the parts of the total variance of the observed data not explained by the common factors, but do not take account of the accuracy of a given datum. In the paper, we suggest an improved algorithm of factor analysis, which performs an automatic weighting process using the prediction errors (i.e. distance between the observed and calculated data) for estimating the factor scores more accurately. Similarly to inversion techniques based on automatic weighting (Drahos, 2008; Gyulai et al., 2014), the newly developed method called Iteratively Reweighted Factor Analysis (IRFA) updates the factor scores and loadings in an iterative procedure, while it effectively reduces the misfit between the observations and predictions. The proper use of Cauchy weights assures a high noise rejection capability of the procedure and its resistance against outliers. By using the IRFA method, we derive a more accurate regression relation between the first factor, which explains the major part of data variance, and shale volume of geological formations. In the paper, an IRFA-based statistical approach is presented to calculate the distribution of shale volume along a borehole. We numerically test the performance of the IRFA method, which is compared to traditional (non-iterative) factor analysis using simulated and observed well logs and core data.

THEORETICAL OVERVIEW

General model of factor analysis

Physical quantities observed in boreholes are simultaneously processed by factor analysis to derive statistical variables called factors. In extracting the factors, the measured data are first scaled to zero mean and unit standard deviation. All suitable input data are collected in a matrix

$$\mathbf{D} = \begin{pmatrix} d_{11} & d_{12} & \cdots & d_{1k} & \cdots & d_{1K} \\ d_{21} & d_{22} & \cdots & d_{2k} & \cdots & d_{2K} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nk} & \cdots & d_{nK} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{N1} & d_{N2} & \cdots & d_{Nk} & \cdots & d_{NK} \end{pmatrix}, \qquad (1)$$

where d_{nk} is the standardized data measured in the *n*-th depth with the *k*-th logging instrument (*N* is the number of sampled depths and *K* is the number of applied sondes). We decompose the data matrix in equation 1 according to the model of factor analysis

$$\mathbf{D} = \mathbf{F}\mathbf{L}^{\mathrm{T}} + \mathbf{E}, \qquad (2)$$

where **F** is the *N*-by-*M* matrix of factor scores, **L** is the *K*-by-*M* matrix factor loadings and **E** is the *N*-by-*K* matrix of residuals (T denotes the operator of matrix transpose). The number of factors (*M*) is less than that of the input variables (*K*). Factor scores of the *l*-th extracted statistical variable are represented in the *l*-th column of the matrix **F** (*l*=1,2,..,*M*). The first factor (given by *l*=1) explains the largest part of variance of the observed data. Subsequent factors represent decreasingly lower contributions of the total variance. The degree of correlation between the various types of data and factors are given in matrix **L**. The range of factor loadings is between −1 and 1, which is similar to the domain of the Pearson's correlation coefficient. The larger the absolute value of factor loadings, the stronger the correlation between the factors and observed data. Since the factors are assumed to be linearly independent ($N^{-1}\mathbf{F}^{\mathrm{T}}\mathbf{F} = \mathbf{I}$, where $\mathbf{I}$ is the unity matrix), the correlation (or covariance) matrix of standardized data is

$$\mathbf{R} = N^{-1}\mathbf{D}^{\mathrm{T}}\mathbf{D} = \mathbf{L}\mathbf{L}^{\mathrm{T}} + \boldsymbol{\Psi}, \qquad (3)$$

where $\boldsymbol{\Psi}$ is the *K*-by-*K* matrix of specific variances representing the portion of data variances not explained by the factors given in matrix **F**. Among different approaches, the factor loadings can be simultaneously estimated with the specific variances by optimizing the following objective function (Jöreskog, 2007)

$$\Gamma(\mathbf{L}, \boldsymbol{\Psi}) = \mathrm{tr}\left(\mathbf{R} - \mathbf{L}\mathbf{L}^{\mathrm{T}} - \boldsymbol{\Psi}\right)^2 = \min . \qquad (4)$$

The factor loadings are usually rotated for a more efficient physical interpretation of factors. When all factor loadings are close to 0 or 1, the factors can be easily interpreted. In other cases, an orthogonal transformation is applied to simplify the diversified structure of factor loadings, which allows the recalculation of factors to obtain ones that are more spectacular. In this study, the *varimax* algorithm suggested by Kaiser (1958) is used, which specifies few data types to which the factors strongly correlate.

We assume that $\mathbf{L}$ and $\boldsymbol{\Psi}$ are known and observed data are normally distributed. The factors can be determined by the maximum likelihood method. An unbiased estimate of the factor scores can be given (Bartlett, 1937)

$$\mathbf{F}^{\mathrm{T}} = \left(\mathbf{L}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\mathbf{L}\right)^{-1}\mathbf{L}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\mathbf{D}^{\mathrm{T}}. \qquad (5)$$

The optimal number of factors can be set by statistical tests (Bartlett 1950), a model selection-based approach (Preacher et al., 2013) or a non-iterative method presented in the next section. Equation 3 shows that the common factors are responsible only for a part of data correlations. Singular value decomposition of the reduced correlation matrix $\hat{\mathbf{R}} = \mathbf{R} - \boldsymbol{\Psi} = \mathbf{L}\mathbf{L}^{\mathrm{T}}$ can be used to quantify the proportions of the total variance of observed data explained by the factors. We decompose the above correlation matrix as $\hat{\mathbf{R}} = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathrm{T}}$, where $\mathbf{U}$ and $\mathbf{V}$ are K-by-K orthogonal matrices and $\mathbf{S}$ is the K-by-K diagonal matrix of singular values arranged in descending order. The total variance of data is given by the trace of matrix $\mathbf{S}$, while the relative percentage of variance explained by the l-th factor is

$$\sigma_l^2 = \frac{S_{ll}}{\mathrm{tr}(\mathbf{S})} \cdot 100\ (\%). \qquad (6)$$

The well logs of rotated factors can be directly applied to extract hidden petrophysical information on reservoir rocks from the well-logging data set.

Non-iterative approximate solution

The appearance of matrix $\mathbf{E}$ in equation 2 causes some inconvenience in solving the problem of factor analysis. In principal component analysis, the matrix of residuals is neglected and a set of linear equations is solved uniquely. If matrix $\mathbf{E}$ is treated as unknown, an appropriate estimate to the specific variances $\mathbf{\Psi}$ must be given. Jöreskog (2007) suggested a non-iterative approximate algorithm for calculating the factor loadings and specific variances. Jöreskog's method gives an objective estimate also to the number of extracted factors, which makes it practical for geophysical applications. Correlation matrix $\mathbf{R}$ in equation 3 can be developed with the factor loadings and specific variances, the diagonal elements of which is composed of the variances of standardized measured variables. The diagonal elements of the reduced correlation matrix $\hat{\mathbf{R}}$ are called communalities, which account for the proportions of data variance explained by the common factors. The matrix of specific variances is related to the *K*-by-*K* matrix of communalities as $\mathbf{H}^2 = \mathbf{I} - \mathbf{\Psi}$. If the communalities h_{kk}^2 are far less than unity, the factors contain only poor information on the observed data.

The counterpart of the *k*-th communality is $u_{kk}^2 = 1 - h_{kk}^2 = r_{kk}^{-1}$, where r_{kk} is the *k*-th diagonal element of the inverse of correlation matrix **R**. Jöreskog (1963) suggested the approximate formula $u_{kk}^2 = \theta r_{kk}$, where the parameter θ was specified as less than unity. This formulation gives the following model of factor analysis

$$\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^{\mathrm{T}} + \theta\left(\operatorname{diag} \mathbf{\Sigma}^{-1}\right)^{-1}, \tag{7}$$

where the covariance matrix **Σ** appears as an implicit function of the factor loadings and parameter θ. By analogy with equation 3, the second term of equation 7 approximates the matrix of specific variances. The modified matrices $\mathbf{\Sigma}^* = \left(\operatorname{diag} \mathbf{\Sigma}^{-1}\right)^{1/2} \mathbf{\Sigma} \left(\operatorname{diag} \mathbf{\Sigma}^{-1}\right)^{1/2}$ and $\mathbf{L}^* = \left(\operatorname{diag} \mathbf{\Sigma}^{-1}\right)^{1/2} \mathbf{L}$ give

$$\mathbf{\Sigma}^* = \mathbf{L}^* \mathbf{L}^{*\mathrm{T}} + \theta\, \mathbf{I}\,, \tag{8}$$

where the second term represents the simplified matrix of specific variances. Let the sample covariance matrix **C** be an estimate of matrix **Σ**. Similarly to the previous case, the matrix $\mathbf{C}^* = \left(\operatorname{diag}\mathbf{C}^{-1}\right)^{1/2} \mathbf{C} \left(\operatorname{diag}\mathbf{C}^{-1}\right)^{1/2}$ is also a consistent estimate of matrix **Σ***. Jöreskog (2007) gave an estimate to the matrix of factor loadings using the eigenvalues and eigenvectors of matrix $\mathbf{C}^*$

$$\mathbf{L} = \left(\operatorname{diag} \mathbf{C}^{-1}\right)^{-1/2} \mathbf{\Omega}_M \left(\mathbf{\Gamma}_M - \theta\, \mathbf{I}\right)^{1/2} \mathbf{U}\,, \tag{9}$$

where $\mathbf{\Gamma}_M$ denotes the diagonal matrix of the first *M* number of sorted eigenvalues, $\mathbf{\Omega}_M$ is the matrix of the first *M* number of eigenvectors (given in its columns) and **U** is an arbitrarily chosen *M*-by-*M* orthogonal matrix. This algorithm suggests that we should choose the smallest number of factors for satisfying the inequality

$$\theta = (K - M)^{-1}(\lambda_{M+1} + \lambda_{M+2} + \ldots + \lambda_K) < 1. \qquad (10)$$

The factor scores can be derived by equation 5. The above solution is very fast to compute and does not require the calculation of communalities, but the procedure is relatively noise sensitive especially if we have outliers in the analyzed data set.

Iteratively reweighted factor analysis

In order to improve the efficiency of traditional factor analysis, which solves a least squares problem using the specific variances as weights in equation 5, we suggest the use of a more sophisticated weighting process implemented in the data space. In the workflow of the Iteratively Reweighted Factor Analysis procedure, the factor loadings and scores calculated by equations 5 and 9 serve as starting values for the iterative procedure. We estimate the new values of factor scores by minimizing the weighted norm of prediction errors (i.e. deviation between the observed data and theoretical data calculated by the actual factor model) in each iteration step. The IRFA method gradually refines the well logs of factor scores by rejecting the measurement noise with high efficiency.

In order to avoid the use of multidimensional arrays, we reformulate the model of factor analysis defined in equation 2

$$\mathbf{d} = \tilde{\mathbf{L}}\mathbf{f} + \mathbf{e}, \qquad (11)$$

where **d** denotes the *NK* column vector of standardized (observed) data, $\tilde{\mathbf{L}}$ is the *NK*-by-*NM* matrix of factor loadings, **f** is the *NM* column vector of factor scores and **e** is the *NK* length vector of prediction errors ($\tilde{\mathbf{L}}\mathbf{f}$ gives the vector of calculated data). Consider the *NK*-by-*NK* diagonal weighting matrix **W**, the non-zero elements of which is chosen in relation with the standard deviation of measured data. The weights quantify the relative importance of each data to the solution. In geoscience problems, the Cauchy weights can be used for giving an outlier-resistant solution (Tarantola, 1987)

$$W_{kk} = \frac{\varepsilon^2}{\varepsilon^2 + e_k^2}, \qquad (12)$$

where ε^2 is a properly set scale parameter, the value of which can be chosen by trial-and-error technique or automatically by the most frequent value method (Steiner, 1991). A feasibility study and application of the most frequent value method in hydrogeological modeling can be found in Szűcs et al. (2006). According to equation

12, the larger the distance between the observed and calculated data, the less weight is given to the relevant data. In the IRFA procedure, the vector of factor scores **f** is calculated by the minimization of the weighted norm of prediction errors

$$\mathbf{e}^{\mathrm{T}}\mathbf{W}\mathbf{e} = \left(\mathbf{d} - \tilde{\mathbf{L}}\mathbf{f}\right)^{\mathrm{T}}\mathbf{W}\left(\mathbf{d} - \tilde{\mathbf{L}}\mathbf{f}\right) = \min , \qquad (13)$$

where matrix **W** contains also the factor scores. The nonlinear weighted least squares problem is solved in the framework of the Iteratively Reweighted Least Squares (IRLS) method (Scales and Gersztenkorn, 1988). In the first step of the procedure, we substitute $\tilde{\mathbf{L}} = \tilde{\mathbf{L}}_0$ into equation 13, where $\tilde{\mathbf{L}}_0$ is the matrix of Jöreskog's factor loadings. It results in the set of normal equations

$$\tilde{\mathbf{L}}_0^{\mathrm{T}}\mathbf{W}\tilde{\mathbf{L}}_0\,\mathbf{f}_1 = \tilde{\mathbf{L}}_0^{\mathrm{T}}\mathbf{W}\mathbf{d} , \qquad (14)$$

the solution of which is equivalent with that of the minimization of the weighted norm of specific variances in equation 5. In the solution of equation 14, the weight matrix **W** is recalculated by the IRLS technique through requisite number of inner-loop iterations. In the next step, the vector of updated factor scores $\mathbf{f}_1$ is used as a constant vector $\hat{\mathbf{f}}$ to improve the factor loadings in $\tilde{\mathbf{L}}_1$. By doing this, we neglect the non-Gaussian nature of the data and solve the problem of the Damped Least Squares method (Marquardt, 1959)

$$\left(\mathbf{d}-\tilde{\mathbf{L}}_1\hat{\mathbf{f}}\right)^{\mathrm{T}}\left(\mathbf{d}-\tilde{\mathbf{L}}_1\hat{\mathbf{f}}\right)+\alpha^2\tilde{\mathbf{L}}_1^{\mathrm{T}}\tilde{\mathbf{L}}_1=\min\,, \qquad (15)$$

where α is the damping constant. Since equation 15 is unweighted, we can return to the original matrix formulation used in equation 2 and write

$$(\mathbf{D}-\hat{\mathbf{F}}\mathbf{L}_1^{\mathrm{T}})^{\mathrm{T}}(\mathbf{D}-\hat{\mathbf{F}}\mathbf{L}_1^{\mathrm{T}})+\alpha^2\mathbf{L}_1^{\mathrm{T}}\mathbf{L}_1=\min\,. \qquad (16)$$

The new values of factor loadings are given by the well-known formula

$$\mathbf{L}_1^{\mathrm{T}}=(\hat{\mathbf{F}}^{\mathrm{T}}\hat{\mathbf{F}}+\alpha^2\,\mathbf{I})^{-1}\hat{\mathbf{F}}^{\mathbf{T}}\mathbf{D}\,. \qquad (17)$$

Following the iterative procedure, matrix $\mathbf{L}_1^{\mathrm{T}}$ (and so $\tilde{\mathbf{L}}_1$) can be inserted into equation 14 to calculate the new values of factor scores in $\mathbf{f}_2$ and so on, until we find the optimal values of the prediction errors. In the q-th step of the iteration procedure, the following set of equations is solved for updating the matrix of factor loadings $\mathbf{L}_q^{\mathrm{T}}$ and the vector of factor scores $\mathbf{f}_q$

$$\left.\begin{aligned}\mathbf{L}_q^{\mathrm{T}}&=(\mathbf{F}_{q-1}^{\mathrm{T}}\mathbf{F}_{q-1}+\alpha^2\,\mathbf{I})^{-1}\mathbf{F}_{q-1}^{\mathrm{T}}\mathbf{D}\\ \mathbf{f}_q&=\left(\tilde{\mathbf{L}}_{q-1}^{\mathrm{T}}\mathbf{W}\tilde{\mathbf{L}}_{q-1}\right)^{-1}\tilde{\mathbf{L}}_{q-1}^{\mathrm{T}}\mathbf{W}\mathbf{d}\end{aligned}\right\}. \qquad (18)$$

At end of the IRFA procedure, a sorting method is applied to the elements of vector $\mathbf{f}$ to represent the factors as depth dependent quantities. In this study, we refer to the Jöreskog’s method combined with the Bartlett’s solution as Traditional Factor Analysis (TFA), and we make a comparison between the TFA and IRFA methods using synthetic and real well-logging data.

Shale volume estimation using factor analysis

The amount of shaliness is classically determined from the natural gamma-ray intensity log. Larionov (1969) suggested an empirical relation between the gamma-ray log reading and shale volume for both unconsolidated and compacted rocks by assuming that radioactive minerals other than clays are not present in the rock matrix. This method can be improved by using the spectral gamma-ray logs, which allow the identification of clay type and estimation of their relative amounts (Serra, 1984). Shale volume can be determined more reliably by the simultaneous processing of suitable logs. In the evaluation of hydrocarbon reservoirs, the following well logs sensitive to shale volume are normally utilized such as natural gamma-ray intensity (*GR* in API), spontaneous potential (*SP* in mV), bulk density (ρ_b in g/cm^3), neutron-porosity (Φ_N in v/v), sonic traveltime (Δt in μs/ft), photoelectric absorption index (P_e in barn/electron) and deep resistivity (R_d in ohm-m). For instance, Kamel and Mabrouk (2003) used neutron-porosity, density and acoustic traveltime data for a more accurate determination of formation shaliness by the presence of radioactive materials other than shale. Several advanced formation evaluation methods are based on statistical principles such as inverse modeling or multivariate statistical procedures. Szabó (2011) showed a strong connection between the first factor and shale volume of sedimentary sequences using oilfield well logs. Szabó and Dobróka (2013) proposed the following regression function for Hungarian and North-American hydrocarbon reservoirs

$$V_{sh} = ae^{bF_1} + c \,, \tag{19}$$

where V_{sh} (v/v) denotes the shale volume, F_1 is the first factor and *a*, *b*, *c* are site-specific constants. It was shown that equation 19 gave a consistent solution for near Gaussian or moderately skewed data distributions. However, in case of larger asymmetry of input data there is a need to use a more robust algorithm of factor analysis. Equation 19 also describes the connection between the first factor and shale volume of unconsolidated North-Hungarian aquifers (Szabó et al., 2014). The hydraulic conductivity of aquifers with primary and secondary porosity as a related quantity to shale volume was successfully correlated to the relevant factor by Szabó (2015).

SYNTHETIC AND FIELD RESULTS

Forward modeling

The aim of synthetic modeling is to quantify the accuracy of the IRFA method. Well-logging data simulated using an exactly known petrophysical model are processed to test the noise sensitivity of the IRFA procedure by measuring the strength of correlation between the exact and estimated shale volume logs. For calculating the well logs in complex reservoirs, we apply the following set of probe response equations (Alberty and Hashmy, 1984)

$$GR=\rho_b^{-1}\left(V_{sh}GR_{sh}\rho_{sh}+\sum_{i=1}^{n}V_{ma,i}GR_{ma,i}\rho_{ma,i}\right), \tag{20}$$

$$SP=V_{sh}SP_{sh}-K^{*}\lg\left(R_{mf}/R_w\right)\left(1-V_{sh}\right), \tag{21}$$

$$\begin{aligned}\rho_b=\mathit{\Phi}\left[\rho_{mf}-1.07\left(1-S_{x0}\right)\left(\alpha_0\rho_{mf}-1.24\rho_h\right)\right]+V_{sh}\rho_{sh}\\+\sum_{i=1}^{n}V_{ma,i}\rho_{ma,i}\end{aligned}\quad, \tag{22}$$

$$\mathit{\Phi}_N=\mathit{\Phi}\begin{Bmatrix}\mathit{\Phi}_{N,mf}-\left(1-S_{x0}\right)C_{cor}\\-2\mathit{\Phi}\left(1-S_{x0}\right)S_{hf}\left(1-2.2\rho_h\right)\\\cdot\left[1-\left(1-S_{x0}\right)\left(1-2.2\rho_h\right)\right]\end{Bmatrix}+V_{sh}\mathit{\Phi}_{N,sh}+\sum_{i=1}^{n}V_{ma,i}\mathit{\Phi}_{N,ma,i}\ , \tag{23}$$

$$\mathit{\Delta}t=\mathit{\Phi}\left[\mathit{\Delta}t_{mf}S_{x0}+\left(1-S_{x0}\right)\mathit{\Delta}t_h\right]+V_{sh}\mathit{\Delta}t_{sh}+\sum_{i=1}^{n}V_{ma,i}\mathit{\Delta}t_{ma,i}\ , \tag{24}$$

$$P_e=\frac{1.07}{\rho_b+0.19}\begin{bmatrix}\mathit{\Phi}S_{x0}U_{mf}+\mathit{\Phi}\left(1-S_{x0}\right)U_h+V_{sh}U_{sh}\\+\sum_{i=1}^{n}V_{ma,i}U_{ma,i}\end{bmatrix}, \tag{25}$$

$$R_d = \left[\frac{\Phi^m S_w^{n^*}}{t \cdot R_w \left(1 - V_{sh}\right)} + \frac{V_{sh} S_w}{R_{sh}} \right]^{-1}, \tag{26}$$

$$\Phi + V_{sh} + \sum_{i=1}^{n} V_{ma,i} = 1, \tag{27}$$

where *Φ* (v/v) denotes the fractional volume of shale-free pore space, $V_{ma,i}$ (v/v) is the relative volume of the *i*-th matrix constituent, *n* is the total number of mineral components, S_{x0} (v/v) and S_w (v/v) are water saturation in the invaded and uninvaded zone, respectively. The zone parameters appearing in equations 20–26 represent the physical properties of mud filtrate (*mf*), hydrocarbon (*h*), shale (*sh*) and the rock matrix (*ma*), which are treated as constant in the forward problem. The denotations and actual values of zone parameters can be found in section Nomenclature and Table 1. The zone properties can be estimated from crossplot techniques, drilling information, laboratory measurements (Jarzyna et al., 2016) or some of them alternatively by interval inversion (Dobróka and Szabó, 2011). Shale volume can also be derived from the material balance formula given in equation 27, which is a constraint equation in solving the well-logging inverse problem. In complex reservoirs, more than one type of minerals form the rock matrix, the volumes of which can be determined from the joint inversion of well logs, e.g. by an interval inversion procedure (Dobróka et al., 2012).

Test using synthetic data

We test the IRFA method using the well logs calculated by equations 20–26 for the petrophysical model plotted in Figure 1. The matrix of the gas-bearing formation is composed of quartz and calcite, while the pore space is occupied by water, irreducible (S_{hirr}=1–S_{x0}) and movable hydrocarbon (S_{hm}=S_{x0}–S_w). In synthetic tests, we assume the model parameters to be exactly known quantities. The average and standard deviation calculated for the well logs of porosity is 0.15±0.05 v/v, for water saturation in invaded zone is 0.92±0.08 v/v, for water saturation in virgin zone is 0.69±0.27 v/v, for shale volume is 0.24±0.25 v/v, for quartz volume is 0.42±0.16 v/v and for calcite volume is 0.19±0.07 v/v. The average skewness (−0.1) and kurtosis (−0.84) of model parameters show slightly flatter and asymmetric data distribution compared to Gaussian.

Factor logs are estimated separately by the TFA and IRFA procedures, which are directly correlated to the known values of shale volume. In Well-1, we simulate the measurement by adding different amount of noise to the synthetic data. We contaminate each datum by a number randomly chosen from Gaussian distribution with zero mean and standard deviation proportional to the required noise level. In the first test, synthetic well logs including 5 % Gaussian noise serves as input for factor analysis. The histogram of the random noise added to the standardized data is shown in Figure 2. At first, we show that the IRFA method gives proper results for normally distributed data (which is a condition for the use of TFA). The average of Pearson's correlation coefficients of the observed variables is 0.55, which shows moderate correlation between the well logs (Table 2). The kurtosis of the sample is 0.12, which also confirms the Gaussian statistics. The result of the TFA method is given by equations 5 and 9, which is used as an initial model for the IRFA method. We set the scale parameter to ε^2=2 for calculating the 1757-by-1757 weight matrix in equation 12. The weight coefficients as a function of the difference between the measured and calculated (standardized) data (vector **e** in equation 11) are shown in Figure 2. By using equation 18, we progressively improve the estimates of the loadings and scores of two uncorrelated factors (suggested by equation 10) over 10 iterations. The magnitude of weight coefficients continuously decreases as the procedure progresses. We use the smallest possible value for regularization parameter α, which ensures a stable iterative process and does not disturb considerably the physical solution, i.e. the values of factor loadings. The damping factor, the initial value of which is α=1, is decreased by 90 % of its actual value in each iteration. In Table 3, the factor loadings inform about the correlation relations between the factors and quasi-measured variables. The largest impact on the first factor is put by shale-sensitive logs like *SP*, *GR* and Φ_N, while the second factor is influenced considerably by seismic properties such as ρ_b and Δt. The use of equation 6 shows that the first factor explains the 91 % of the total variance of the well-logging data, which correlates highly with the fraction of shale (Figure 2). Despite of the high value of the Pearson's correlation coefficient (0.98), the functional relation is slightly nonlinear (Figure 2). The regression coefficients of equation 19 estimated with 95 % confidence bounds are a=1.12±0.29, b=0.20±0.05, c=–0.90±0.28. The input logs and the results of IRFA are in Figure 3. Shale volumes estimated by traditional ($V_{sh,TFA}$) and iteratively weighted factor analysis ($V_{sh,IRFA}$) show good agreement with their theoretical values. The procedure of IRFA gives a slightly smoother solution. The root-mean-square error (RMS) between the exact and TFA-derived shale volume logs is 7.1 %, while it is 4.8 % for the IRFA method. The relative percentage decrease is 32 %, which shows better result for the iterative approach in case of normally distributed data.

In the next step, the effect of outliers is tested. For simulating a non-Gaussian data distribution, six times higher amount of noise is randomly added to the 1/10 part of the Gaussian distributed data. The RMS (normalized) distance between the noisy and noiseless data is 7 %. The average correlation (0.36) shows weaker correlation between the well logs than in the previous case, and the kurtosis (5.49) indicates a leptokurtic data distribution including outliers. We conduct statistical tests of normality for proving the non-Gaussian distribution of the input data. In Figure 4, the ordinate axes are scaled by the standard Gaussian distribution function. The deviation of data points from the straight-line shows that the (quasi-) measured variables are not normally distributed. Two factors are extracted from the data set by using ε^2=1 and α=1. The maximal number of iterations is 10. The components of vector **e** and the Cauchy weights calculated in the first iteration are plotted in Figure 5. The first 251 elements in the main diagonal of matrix **W** represent the weights of the *GR* log, the second 251 elements corresponds to the *SP* log, and so on. We normalize each weighting coefficient by the sum of the weights of the given well log type. Thus, the maximal weight put on small prediction errors for each well log is different. The estimated scores of the two factors are cross-plotted in Figure 6, which shows that the TFA method is quite sensitive to outliers, while the IRFA procedure is outlier-resistant. The first factor explains the 94 % part of variance of the observations. The loadings of the same factor are consistent to those of the purely Gaussian case: $L^{(SP)}$=–0.97, $L^{(GR)}$=0.98, $L^{(\rho_b)}$=0.05, $L^{(\Phi_N)}$=0.89, $L^{(\Delta t)}$=0.18, $L^{(Rd)}$=–0.79, $L^{(Pe)}$=0.58. The first factor is still strongly correlated to *SP*, *GR*, Φ_N and R_d logs, but the relatively noisy ρ_b, Δt and P_e logs have negligible impact on it. One can make a comparison between the factor vs. shale volume relations estimated by the TFA and IRFA procedures in Figure 7. The Pearson’s correlation coefficient between the exact values of shale volume and the first factor is 0.95 for TFA, while it is 0.98 for IRFA. We conclude that IRFA gives a more accurate solution with better correlation. The coefficients of the exponential function obtained by IRFA are a=0.84±0.17, b=0.27±0.05, c=–0.63±0.16. The noisy (input) well logs and the results of factor analysis are plotted in Figure 8. Factor and shale volume logs calculated by the TFA procedure include erroneous peaks mostly in the interval of 0–10 m. In contrast with the traditional method, the IRFA procedure gives a smoother estimate to the factor variables and the derived shale volume log. The RMS misfit between the exact and TFA-based shale volume log is 7.2 %, while that for the IRFA method is 4.4 %. It is concluded that the IRFA method can be advantageously used to process data sets of non-Gaussian distributions and extreme noises. The synthetic modeling experiments show that the factors can be completely purified from outliers as well as significant improvement can be made in the estimation accuracy of shale volume by using the robust algorithm.

Test using real well logs

We chose a well-logging data set originated from the Powder River Basin Province, Wyoming, USA, from the literature (Anna, 2009). In the processed depth of Well-2, the Minnelusa formation of Paleozoic age is composed of shales as potential source rocks and sandy dolomites as low porosity hydrocarbon reservoirs. We apply the IRFA method to analyze the spontaneous potential (*SP*), natural gamma-ray intensity (*GR*), acoustic traveltime (Δt) and deep resistivity (R_d) logs measured with a sampling interval of one foot. The average correlation of the well logs is 0.42 and the data are nearly Gaussian distributed (skewness is 0.33 and kurtosis is −0.77). We extract one factor from four observed well logs. The calculation of factor loadings is made in a stable procedure by α=0 in equation 18. The maximal number of iterations is 10 and the scale parameter ε^2 is 0.3. At the end of the IRFA procedure the factor loadings are $L^{(\Delta t)}$=0.09, $L^{(Rd)}$=–0.17, $L^{(SP)}$=–0.72, $L^{(GR)}$=0.79. The first factor is mainly sensitive to lithologic effects, which was demonstrated in the synthetic modeling experiments, too. We calculate the reference values of shale volume by the Larionov formula suggested for older than Tertiary rocks (Larionov, 1969). Figure 9 shows a strong relation between the first factor and shale volume in Well-2, which is indicated by the correlation coefficient of 0.98. The magnitude and sign of the regression coefficients are consistent with those of the synthetic modeling tests. The estimated values of the constants in equation 19 are a=0.57±0.12, b=0.33±0.06, c=–0.34±0.12. In the upper part of the section, the *GR* log shows the presence of shales (Figure 10). Around the depth of 10,300 feet, high resistivities indicate a hard dolomite, below which sandstones and sandy dolomites are deposited. The first factor and shale volume are plotted in the last two tracks. The upper part of the section shows that IRFA properly resolves thin shale layers. The RMS distance between the shale volumes estimated by the Larionov formula ($V_{sh,LAR}$) and the IRFA method is 4.9 %.

Test using core data

The investigated borehole (Well-3) was originally drilled for hydrocarbon exploration in Baktalórántháza, Great Hungarian Plain, North-East-Hungary. We process an interval of 80 m, where high porosity (unconsolidated) shaly sands of Pleistocene age were deposited. In the rock matrix we find carbonate cement, which amounts to 1.5–9.6 % according to core tests. In the processed depth, the reservoirs are fully freshwater saturated (S_{x0}=S_w=1). We utilize the self-potential (*SP*), natural gamma-ray intensity (*GR*), caliper (*CAL*), gamma-gamma intensity (γ–γ), neutron-(thermic) neutron (*NN*) and shallow resistivity (R_s) logs. Shale volume is available from the grain-size analysis of 29 core samples, and estimated by the Larionov formula used in younger than Tertiary rocks (Larionov, 1969). The porosity of the formation is calculated from the combination of γ–γ and *NN* logs. The observed variables practically follow Gaussian distribution (the skewness and kurtosis are nearly 0, the average correlation is 0.42), thus, we show the results only of the IRFA procedure. We extract one factor from the six well logs by the same parameter settings as in Well-2. The first factor is mainly sensitive to the lithologic logs as the factor loadings show: $L^{(Rs)}$=–0.40, $L^{(SP)}$=0.49, $L^{(GR)}$=0.96, $L^{(\gamma-\gamma)}$=0.44, $L^{(NN)}$=–0.34, $L^{(CAL)}$=–0.13. The factor scores are directly correlated to core measurements. Figure 9 shows the local regression relation between the first factor and shale volume in Well-3. The correlation coefficient (0.92) indicates a strong relation between the above quantities. The coefficients of equation 19 are close to those of Well-2. The estimated values of constants and their standard deviations are a=0.43±0.19, b=0.36±0.13, c=–0.17±0.19. The result of IRFA is shown in Figure 11. The *GR* image shows the aquifers with light grey color, while shales are represented by darker colors. The cyclic variation of sediments is well observable in the *SP* and R_s logs. The IRFA-derived shale volume log fits well to those of calculated by the Larionov formula and core information. The RMS error between the shale volumes estimated by the Larionov formula and core measurements ($V_{sh,CORE}$) is 4.1 %, while that of computed between the IRFA method and laboratory measurements is only 3.1 %.

DISCUSSION

As opposed to classical methods using a single log for the interpretation, we utilize all wireline logs sensitive to shaliness for factor analysis to give a more reliable estimate to shale volume. The validity of equation 19 has been verified both by simulated and real well-logging data. The results of field experiments show that shale volume estimated by the IRFA method agrees properly with that of independent well log analysis and laboratory methods (Figure 12). The factor analysis of synthetic data is useful also in setting the control parameters of the IRFA procedure. The choice of the scale parameter of weighting function can be made automatically or by preliminary tests. With the decrease of ε, bigger deviations contribute less to the solution. Another important question is the setting of the number of extracted factors. The optimal number of factors can be classically determined by statistical tests. Jöreskog (2007) gives a more practical solution to this problem by applying equation 10. For studying the effect of the number of factors, we generate a data set composed of GR, SP, ρ_b, Φ_N, Δt, R_d, P_e logs contaminated with 5 % Gaussian noise. In Table 4, parameter θ is less than and closest to 1, if the number of factors is two. (Similarly, in our previous synthetic tests we extracted two factors in the optimal case.) In this experiment, we study the IRFA procedure by maximum five factors. Table 4 shows that the increase of the number of factors improves the fit between the measured and calculated data (decreases the L_2 norm of prediction error), but the factor loadings related to the lithologic logs (e.g. GR) and the relative variance explained by the first factor significantly decreases. It is obvious that the formulation of IRFA allows the quality check of the results of factor analysis in data space. In factor analysis, the measured variables are transformed into less number of factors, which implies the loss of some observed information. Figure 13 shows the misfit between the calculated (standardized) and quasi-measured resistivity logs beside different number of factors. It must be mentioned that computed well log types with higher factor loadings such as *GR*, *SP* and Φ_N show even better fit to observations than the resistivity log for the cases of less extracted factors. By increasing the number of factors, we neglect relatively smaller amount of information as well as the data misfit reduces. For more number of factors, the singular values for the rest of the factors usually equal to zero and we obtain zero factor loadings. In case of high number of factors, the information is shared more greatly by the factors and the correlation between the first factor and shale volume is reduced. In order to concentrate the information on the lithology, we are advised to use less number of factors. For the specification of the number of factor logs, we suggest the use of equation 10 on the condition that we have high-valued factor loadings related to determinative well logs and a tolerable level of data misfit.

CONCLUSIONS

We present an improved data processing approach to make a more robust determination of shale volume using the factor analysis of well logs. The test results, presented in this paper, confirm the feasibility of the IRFA method, which gives optimal results not only for normally distributed well-logging data. The weighting process of IRFA efficiently reduces the harmful impact of extreme noises caused by the measurement tools or other environmental effects, which can be a useful tool, e.g. in the re-processing of old well logs. The IRFA method insensitive to outliers allows a more reliable estimation of shale content variation along the borehole. According to our experience, at least 30−40 % relative improvement of estimation accuracy can be achieved compared to traditional factor analysis depending on the noise level of the well logs. For data sets including higher number of outliers than in the presented study, even better quality improvement can be reached. It must be mentioned that the caliper log is strongly connected to lithologic characteristics of formations, but we cannot describe these connections explicitly. Thus, caliper log cannot be used in inversion procedures. In accordance with inverse modeling, there is nothing to prevent using the caliper log in factor analysis, which may give further information for a better interpretation of shaly formations. Moreover, the IRFA method allows the quality check of the results in data space. As a conclusion, we must take a trade-off between the value of prediction error and the amount of information explained by the extracted factors.

We currently study the possibility of using global optimization methods such as Genetic Algorithms in solving the problem of factor analysis to improve the fit between the measured and predicted well logs. The essence of the method is the use of a consistent exponential relation between the first factor and shaliness of reservoirs rocks. Shale volume and derived quantities such as effective porosity, permeability, water and hydrocarbon saturations can be extracted reliably from the factor scores by the IRFA procedure, which may improve the results of reservoir modeling. This statistical method can be easily further developed to crosswell applications. The simultaneous processing of well-logging data sets acquired from several neighboring boreholes gives multi-dimensional sections of the factor variables and petrophysical properties. In this framework, the usage of a large statistical sample makes significant improvement in the estimation accuracy of the derived reservoir parameter. The presented statistical methodology can be employed as a useful data processing tool in oilfield applications, the basic idea of which may be applied fruitfully also in the study of unconventional (especially shale gas) reservoirs.

NOMENCLATURE

a, b, c = Regression coefficients of factor vs. shale content relation

$\mathbf{C}$ = Sample covariance matrix of standardized data

C_{cor} = Mud-filtrate correction coefficient in neutron response equation

$\mathbf{d}$ = Vector of standardized well-logging data in IRFA procedure

$\mathbf{D}$ = Standardized data matrix as input for factor analysis

$\mathbf{e}$ = Vector of data prediction errors in IRFA procedure

$\mathbf{E}$ = Matrix of residuals in the model of factor analysis

$\mathbf{f}$ = Vector of factor scores in IRFA procedure

$\mathbf{F}$ = Matrix of factor scores

F_1 = Score of the first statistical factor

$\mathbf{H}^2$ = Matrix of communalities

K = Number of log types involved in factor analysis

K^* = Temperature factor used for calculation of SP (mV)

$\mathbf{L}$ = Matrix of factor loadings

m = Cementation exponent in resistivity response equation

M = Number of extracted statistical factors

n = Number of mineral types in material balance equation

n^* = Saturation exponent in resistivity response equation

N = Number of investigated depths along the borehole

P_e = Photoelectric absorption index (barn/electron)

$\mathbf{R}$ = Correlation matrix of standardized data

$\hat{\mathbf{R}}$ = Reduced correlation matrix calculated from the factor loadings

R_d = Resistivity measured by deep penetration tool (ohm-m)

R_{mf} = Resistivity of mud-filtrate (ohm-m)

R_{sh} = Resistivity of shale (ohm-m)

R_w = Resistivity of pore-water (ohm-m)

S_{hf} = Residual hydrocarbon coefficient in neutron response equation

S_{hirr} = Irreducible hydrocarbon saturation (v/v)

S_{hm} = Movable hydrocarbon saturation (v/v)

S_{w} = Water saturation in uninvaded zone (v/v)

S_{x0} = Water saturation in invaded zone (v/v)

t = Tortuosity factor in resistivity response equation

U_{h} = Volumetric photoelectric absorption index of hydrocarbon (barn/cm^3)

U_{ma} = Volumetric photoelectric absorption index of rock matrix (barn/cm^3)

U_{mf} = Volumetric photoelectric absorption index of mud-filtrate (barn/cm^3)

U_{sh} = Volumetric photoelectric absorption index of shale (barn/cm^3)

V_{lm} = Fractional volume of limestone (v/v)

V_{ma} = Fractional volume of rock matrix (v/v)

V_{sd}	=	Fractional volume of sandstone (v/v)
V_{sh}	=	Volume of shale relative to total rock volume (v/v)
$\mathbf{W}$	=	Diagonal matrix of Cauchy weight coefficients
α	=	Damping factor used for calculating factor loadings
α_0	=	Mud-filtrate coefficient in density response equation
ε	=	Scale parameter of Cauchy weight function
Φ	=	Shale-free porosity (v/v)
Φ_N	=	Neutron-porosity (v/v)
$\Phi_{N,ma}$	=	Neutron-porosity of rock matrix (v/v)
$\Phi_{N,mf}$	=	Neutron-porosity of mud-filtrate (v/v)
$\Phi_{N,sh}$	=	Neutron-porosity of shale (v/v)
$\mathbf{\Psi}$	=	Matrix of specific variances
θ	=	Jöreskog’s constant used to give the number of factors
ρ_b	=	Bulk density (g/cm^3)
ρ_h	=	Density of hydrocarbon (g/cm^3)

ρ_{ma} = Density of rock matrix (g/cm^3)

ρ_{mf} = Density of mud-filtrate (g/cm^3)

ρ_{sh} = Density of shale (g/cm^3)

σ^2_l = Relative variance explained by the *l-th* factor

$\mathbf{\Gamma}$ = Matrix of eigenvalues of sample covariance matrix

$\mathbf{\Omega}$ = Matrix of eigenvectors of sample covariance matrix

$\mathbf{\Sigma}$ = Covariance matrix of standardized data

Δt = Acoustic interval-time (μs/ft)

Δt_h = Acoustic interval-time of hydrocarbon (μs/ft)

Δt_{ma} = Acoustic interval-time of rock matrix (μs/ft)

Δt_{mf} = Acoustic interval-time of mud-filtrate (μs/ft)

Δt_{sh} = Acoustic interval-time of shale (μs/ft)

γ-γ = Gamma-gamma logging data (cpm)

CAL = Observed caliper of borehole (inch)

GR = Natural gamma-ray intensity (API)

GR_{ma}	=	Natural gamma-ray intensity of rock matrix (API)
GR_{sh}	=	Natural gamma-ray intensity of shale (API)
IRFA	=	Method of Iteratively Reweighted Factor Analysis
NN	=	Neutron-neutron logging data (cpm)
RMS	=	Root mean square error (deviation between well logs)
SP	=	Spontaneous potential (mV)
SP_{sh}	=	Spontaneous potential of shale (mV)
TFA	=	Traditional algorithm of factor analysis

ACKNOWLEDGMENTS

The research was supported by the GINOP-2.3.2-15-2016-00010 "Development of enhanced engineering methods with the aim at utilization of subterranean energy resources" project in the framework of the Széchenyi 2020 Plan, funded by the European Union, co-financed by the European Structural and Investment Funds. The research were partly supported by the National Research Development and Innovation Office (project No. K109441), as the leader of the project, the second author thanks the support. The authors thank to Minnelusa.com for permitting the use of the data set of Well-2, and to

Béla Latrán from Geokomplex Ltd. for the well-logging and grain-size data of Well-3.

REFERENCES

Alberty, M., and K. Hashmy, 1984, Application of ULTRA to log analysis: SPWLA Symposium Transactions, paper Z, 1–17.

Anna, L. O., 2006, Geologic assessment of undiscovered oil and gas in the Powder River Basin Province: U.S. Geological Survey Digital Data Series, DDS–69–U.

Asadi, H., Lu, Y-j., and T. C., McCuaig, 2014, Exploratory data analysis and C–A fractal model applied in mapping multi-element soil anomalies for drilling: A case study from the Sari Gunay epithermal gold deposit, NW Iran: Journal of Geochemical Exploration, **145**, 233–241.

Asfahani, J., 2014, Statistical factor analysis technique for characterizing basalt through interpreting nuclear and electrical well logging data (case study from Southern Syria): Applied Radiation and Isotopes, **84**, 33–39.

Bartlett, M. S., 1937, The statistical conception of mental factors: British Journal of Psychology, **28**, 97–104.

Bartlett, M. S., 1950, Tests of significance in factor analysis: British Journal of Statistical Psychology, **3**, 77–85.

Buoro, A. B., and J. B. C. Silva, 1994, Ambiguity analysis of well-log data: Geophysics, **59**, P. 336−344.

Charfi, S., Zouari, K., Feki, S., and E. Mami, 2013, Study of variation in groundwater quality in a coastal aquifer in north-eastern Tunisia using multivariate factor analysis: Quaternary International, **302**, 199−209.

Croux, C., Filzmoser, P., Pison, G., and P. J. Rousseeuw, 1999, Fitting factor models by robust interlocking regression: Technical Report, Vienna University of Technology.

Da Silva Pereira, J. E., Strieder, A. J., Pereira Amador, J., Da Silva, J. L., and L. L. Volcato Descovi Filhoc, 2010, A heuristic algorithm for pattern identification in large multivariate analysis of geophysical data sets: Computers and Geosciences, **36**, 83–90.

Derby, N. E., Korom, S. F., and F. X. M. Casey, 2013, Field-scale relationships among soil properties and shallow groundwater quality: Groundwater, **51**, 373–384.

Dobróka, M., and N. P. Szabó, 2011, Interval inversion of well-logging data for objective determination of textural parameters: Acta Geophysica, **59**, 907–934.

Dobróka, M., Szabó, N. P., and E. Turai, 2012, Interval inversion of borehole data for petrophysical characterization of complex reservoirs. Acta Geodaetica et Geophysica, **47**, 172–184.

Drahos, D., 2008, Determining the objective function for geophysical joint inversion: Geophysical Transactions, **45**, 105–121.

Filzmoser, P., Hron, K., Reimann, C., and R. Garrett, 2009, Robust factor analysis for compositional data: Computers & Geosciences, **35**, 1854–1861.

Goncalves, C. A., Harvey, P. K., and M. A. Lovell, 1995, Application of a multilayer neural network and statistical techniques in formation characterisation: SPWLA 36th Annual Logging Symposium, 1–12.

Gyulai, Á., Baracza, M. K., and N. P. Szabó, 2014, On the application of combined geoelectric weighted inversion in environmental exploration: Environmental Earth Sciences, **71**, 383–392.

Hoseinpoor, M. K., and A. Aryafar, 2014, The use of robust factor analysis of compositional geochemical data for the recognition of the target area in Khusf 1:100000 sheet, South Khorasan, Iran: International Journal of Mining and Geo-Engineering, **48**, 191−199.

Jarzyna, A. J., Krakowska, P. I., Puskarczyk, E., Wawrzyniak-Guz, K., Bielecki, J., Tkocz, K., Tarasiuk, J., Wronski, S., and M. Dohnalik, 2016, X-ray computed microtomography - a useful tool for petrophysical properties determination, Computational Geosciences, **20**, 1155–1167.

Jöreskog, K. G., 1963, Statistical estimation in factor analysis: Almqvist & Wiksell.

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, 47–77.

Kaiser, H. F., 1958, The varimax criterion for analytical rotation in factor analysis: Psychometrika, **23**, 187–200.

Kamel, M. H., and Mabrouk W. M., 2003, Estimation of shale volume using a combination of the three porosity logs: Journal of Petroleum Science and Engineering, **40**, 145–157.

Krogulec, E., and S. Zabłocki, 2015, Relationship between the environmental and hydrogeological elements characterizing groundwater-dependent ecosystems in central Poland. Hydrogeology Journal, **23**, 1587–1602.

Larionov, V. V., 1969, Radiometry of boreholes (in Russian). Nedra Moscow.

Lawley, D. N., and A. E. Maxwell, 1962, Factor analysis as a statistical method: The Statistician, **12**, 209–229.

Luttinen, J., Ilin, A., and J. Karhunen, 2012, Bayesian robust PCA of incomplete data: Neural Processing Letters, **36**, 189−202.

Marquardt, D. W., 1959, Solution of non-linear chemical engineering models: Chemical Engineering Progress, **55**, 65–70.

Mayer, C., and A. Sibbit, 1980, GLOBAL, a new approach to computer-processed log interpretation: 55th SPE Annual Fall Technical Conference and Exhibition, paper 9341, 1–14.

Menke, W., 1984, Geophysical data analysis: Discrete inverse theory: Academic Press Inc.

Merdith, A., and D. Müller, 2015, Prospectivity of Western Australian iron ore from geophysical data using a reject option classifier: Ore Geology Reviews, **71**, 761–776.

Mongelli, G., Boni, M., Buccione, R., and R. Sinisi, 2014, Geochemistry of the Apulian karst bauxites (southern Italy): Chemical fractionation and parental affinities: Ore Geology Reviews, **63**, 9–21.

Pison, G., Rousseeuw, P. J., Filzmoser, P., and C. Croux, 2003, Robust factor analysis: Journal of Multivariate Analysis, **84**, 145–172.

Preacher, K. J., Zhang, G., Kim, C., and G. Mels, 2013, Choosing the optimal number of factors in exploratory factor analysis: A model selection perspective: Multivariate Behavioral Research, **48**, 28–56.

Qian, J., Chen, H., and Y. Meng, 2011, Geological characteristics of the Sizhuang gold deposit in the region of Jiaodong, Shandong Province - A study on tectono-geochemical ore prospecting of ore deposits: Chinese Journal of Geochemistry, **30**, 539−553.

Scales, J. A., and A. Gersztenkorn, 1988, Robust methods in inverse theory: Inverse Problems, **4**, 1071−1091.

Serra, O., 1984, Fundamentals of well-log interpretation: Elsevier.

Seth, V., Srivardhan, V., and S. Maiti, 2015, Evaluation of formation shaliness using factor analysis of site-U1344A of IODP expedition 323 in the Bering Sea: 77th EAGE Conference and Exhibition, Extended Abstracts, paper Tu SP 114, 1–4.

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

Sun, X., Deng, J., Gong, Q., Wang, Q., Yang, L., and Z. Zhao, 2009, Kohonen neural network and factor analysis based approach to geochemical data pattern recognition: Journal of Geochemical Exploration, **103**, 6–16.

Szabó, N. P., 2011, Shale volume estimation based on the factor analysis of well-logging data: Acta Geophysica, **59**, 935–953.

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

Szabó, N. P., 2012, Dry density derived by factor analysis of engineering geophysical sounding measurements: Acta Geodaetica et Geophysica, **47**, 161–171.

Szabó, N. P., and M. Dobróka, 2013, Extending the application of a shale volume estimation formula derived from factor analysis of wireline logging data: Mathematical Geosciences, **45**, 837–850.

Szabó, N. P., Dobróka, M., Turai, E., and P. Szűcs, 2014, Factor analysis of borehole logs for evaluating formation shaliness: a hydrogeophysical application for groundwater studies: Hydrogeology Journal, **22**, 511–526.

Szabó, N. P., 2015, Hydraulic conductivity explored by factor analysis of borehole geophysical data: Hydrogeology Journal, **23**, 869–882.

Szűcs, P., Civan, F., and M. Virág, 2006, Applicability of the most frequent value method in groundwater modeling: Hydrogeology Journal, **14**, 31–43.

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

Wawrzyniak-Guz, K., Jarzyna, J. A., Zych, M., Bała, M., Krakowska, P. I., and E. Puskarczyk, 2016, Analysis of the heterogeneity of the Polish shale gas formations by factor analysis on the basis of well logs: 78th EAGE Conference and Exhibition, Extended Abstracts, paper Tu SBT3 07, 1–5.

LIST OF FIGURE CAPTIONS

Figure 1. Exactly known petrophysical model and calculated well logs in Well-1. Model parameters are shale-free porosity (Φ), water saturation in the invaded and uninvaded zone (S_{x0} and S_w), shale volume (V_{sh}), sand volume (V_{sd}), limestone volume (V_{lm}), irreducible and movable hydrocarbon saturation (S_{hirr} and S_{hm}). Noiseless

calculated logs are natural gamma-ray intensity (GR), spontaneous potential (SP), photoelectric absorption index (P_e), bulk density (ρ_b), neutron-porosity (Φ_N), sonic traveltime (Δt) and deep resistivity (R_d).

Figure 2. Gaussian noise added to standardized synthetic well logs in Well-1 (top left panel). Regression relation between the first factor extracted from well logs contaminated with 5 % Gaussian noise and the exact values of shale volume (top right panel). Cauchy weights calculated automatically for each well log in the IRFA procedure vs. data prediction errors (bottom panel). Input well logs are natural gamma-ray intensity (GR), spontaneous potential (SP), bulk density (ρ_b), neutron-porosity (Φ_N), sonic traveltime (Δt), deep resistivity (R_d) and photoelectric absorption index (P_e).

Figure 3. Result of factor analysis in Well-1. Input well logs contaminated with 5 % Gaussian noise are natural gamma-ray intensity (GR), spontaneous potential (SP), photoelectric absorption index (P_e), bulk density (ρ_b), neutron-porosity (Φ_N), sonic traveltime (Δt) and deep resistivity (R_d). Well logs of the first and second factors extracted by traditional factor analysis ($F_{1,TFA}$ and $F_{2,TFA}$) and iteratively reweighted factor analysis ($F_{1,IRFA}$ and $F_{2,IRFA}$). Well logs of exactly known shale volume ($V_{sh,exact}$) and shale volume estimated by traditional factor analysis ($V_{sh,TFA}$) and iteratively reweighted factor analysis ($V_{sh,IRFA}$).

Figure 4. Normal probability plots of well logs simulated in Well-1. Frequency of standardized Gaussian distributed statistical variable (solid line). Frequency of synthetic well-logging data contaminated by Gaussian noise and outliers (circle symbol). Well logs are natural gamma-ray intensity (*GR*), spontaneous potential (*SP*), bulk density (ρ_b), neutron-porosity (Φ_N), sonic traveltime (Δt), deep resistivity (R_d) and photoelectric absorption index (P_e).

Figure 5. Cauchy weights used in the first iteration step of the IRFA procedure in Well-1. Prediction error as deviation between the observed and calculated data (top panel). Cauchy weight coefficients are inversely proportional to the deviation between measured and calculated well-logging data and are of different maximal values for each log type (bottom panel).

Figure 6. Crossplots of the two statistical factors estimated in Well-1 by the TFA and IRFA procedures, respectively. The result of the TFA procedure is highly sensitive to outliers (left panel), while they are effectively rejected using the IRFA method (right panel).

Figure 7. Regression relation between the first factor extracted from noisy synthetic well logs including outliers and the exact values of shale volume in Well-1. Regression analysis is highly influenced by extreme values of factors estimated by the TFA procedure (left panel), while the IRFA method is resistant against outliers (right panel). The

Pearson's correlation coefficient between the shale volume and first factor is 0.95 for TFA, while it is 0.98 for IRFA.

Figure 8. Result of factor analysis in Well-1. Well logs including Gaussian noise and outliers are natural gamma-ray intensity (*GR*), spontaneous potential (*SP*), photoelectric absorption index (P_e), bulk density (ρ_b), neutron-porosity (Φ_N), sonic traveltime (Δt) and deep resistivity (R_d). Well logs of the first and second factors extracted by traditional factor analysis ($F_{1,TFA}$ and $F_{2,TFA}$) and iteratively reweighted factor analysis ($F_{1,IRFA}$ and $F_{2,IRFA}$). Well logs of exactly known shale volume ($V_{sh,exact}$) and shale volume estimated by traditional factor analysis ($V_{sh,TFA}$) and iteratively reweighted factor analysis ($V_{sh,IRFA}$).

Figure 9. Regression relation between the first factor extracted from real well-logging data and shale volume. Nonlinear (exponential) relation between the first factor and shale volume calculated deterministically by the Larionov method in Well-2 (left panel). Regression connection between the first factor and shale volume obtained from core analysis in Well-3 (right panel).

Figure 10. Result of factor analysis in Well-2. The investigated formation is a North-American shaly-sandy carbonate of Pennsylvanian age. Input well logs are natural gamma-ray intensity (*GR*), spontaneous potential (*SP*), acoustic traveltime (Δt) and deep resistivity (R_d). The first factor log is extracted by traditional factor

analysis ($F_{1,TFA}$) and iteratively reweighted factor analysis ($F_{1,IRFA}$). Shale volume is estimated by traditional factor analysis ($V_{sh,TFA}$), iteratively reweighted factor analysis ($V_{sh,IRFA}$) and the Larionov method ($V_{sh,LAR}$).

Figure 11. Result of factor analysis in Well-3. The unconsolidated formation is a Hungarian shaly-sandy (carbonate cemented) sequence of Pleistocene age. Observed well logs are natural gamma-ray intensity (*GR*), caliper (*CAL*), spontaneous potential (*SP*), gamma-gamma intensity (γ–γ), neutron-neutron (*NN*) and shallow resistivity (R_s). The parameters of the compositional analysis are shale-free porosity (Φ), shale volume (V_{sh}) and volume of rock matrix (V_{ma}). The first factor log is extracted by iteratively reweighted factor analysis ($F_{1,IRFA}$). Shale volume is estimated by iteratively reweighted factor analysis ($V_{sh,IRFA}$), the Larionov method ($V_{sh,LAR}$) and core analysis ($V_{sh,CORE}$).

Figure 12. Regression relation between shale volumes derived by different methods using real well logs. Linear connection between shale volumes estimated separately by the IRFA and Larionov method in Well-2 (left panel). Linear connection between shale volumes estimated separately by the IRFA method and core analysis in Well-3 (right panel).

Figure 13. Results of factor analyses given by different number of factors (*M*) in Well-1. Deviation between the deep resistivity (R_d) logs calculated from the estimated factors (dashed line) and the quasi-measured resistivity logs (solid line) characterizes the misfit in data space at the end of the IRFA procedure.