PRINCIPAL COMPONENT ANALYSIS AS A TOOL FOR ENHANCED WELL LOG INTERPRETATION

Gina ANDREI *, Bogdan Mihai NICULESCU *

* *University of Bucharest, Faculty of Geology and Geophysics, Department of Geophysics*
(*gina.andrei@g.unibuc.ro*; *bogdan.niculescu@gg.unibuc.ro*)

Introduction

Principal Component Analysis (PCA) (Pearson, 1901; Hotelling, 1933; Jolliffe, 2002) is a multivariate data dimensionality reduction technique, used to simplify a data set to a smaller number of factors which explain most of the variability (variance). PCA aims to convert a set of correlated variables to a number of uncorrelated orthogonal *principal components* (PCs). Besides dimensionality reduction, this analysis may also be employed to discover and interpret the dependencies and relationships possibly existing among the original variables. In practice, PCA is carried out by computing the covariance matrix of the data set, then the eigenvalues and eigenvectors of the covariance matrix are computed and sorted according to decreasing eigenvalues, i.e. decreasing amounts of data variability. For a meaningful interpretation of the principal components it is important to determine which original variables are associated with particular components.

PCA has been successfully used for a variety of well logging data applications, such as: identification and characterization of pressure seals / low permeability intervals (Moline et al., 1992), delineation of lithostratigraphic units, identification of aquifer formations and distinction between hydraulic flow units (Kassenaar, 1991; Barrash and Morin, 1997), interdependency and correlation between some hydraulic properties and geophysical/petrophysical parameters (Morin, 2006), well-to-well correlation by pattern recognition etc. We investigate and discuss the potential usefulness of PCA in providing meaningful petrophysical information in the case of hydrocarbon exploration wells, in addition to the results obtained via conventional log interpretation, or to constrain and validate such results.

Summary of Principal Component Analysis method

Taking into account a multivariate data set X consisting in p random variables x_1, x_2, …, x_i, …, x_p (i.e., geophysical well logs, each log consisting in n measurements of a specific subsurface property), the p principal components z_1, z_2, …, z_i, …, z_p of the data set are given by the linear combinations:

$$z_i = a_i^T X = a_{i1} x_1 + a_{i2} x_2 + \ldots + a_{ip} x_p \; ; \; i = 1, 2, \ldots, p \qquad (1)$$

where a_i are the column vectors of an orthogonal p-by-p transformation matrix A ($A^TA = AA^T = I$, with T denoting the transpose and I the p-by-p identity matrix). Besides the normalization expressed by $a_i^T a_i = 1$ ($i = 1, 2, \ldots, p$) and the orthogonality of the PCs, a condition imposed when extracting the PCs is $var(z_1) \geq var(z_2) \geq \ldots \geq var(z_p)$, where *var* stands for the variance. The first PC is $a_1^T X$, subject to $a_1^T a_1 = 1$, that maximizes $var(a_1^T X)$; the second PC is $a_2^T X$ that maximizes $var(a_2^T X)$, subject to $a_2^T a_2 = 1$ and covariance $cov(a_1^T X, a_2^T X) = 0$ (uncorrelated principal components) and so on. For each PC, the variance that has to be

maximized subject to $a_i^T a_i = 1$ (i.e., $a_i^T a_i - 1 = 0$) can be expressed as $var(z_i) = var\,(a_i^T X) = a_i^T S\, a_i \rightarrow maximum$, where S is the p-by-p sample covariance matrix of the data.

The constrained maximization problem can be solved by creating a function $L = a_i^T S\, a_i - \lambda (a_i^T a_i - 1)$, where λ stands for a Lagrange multiplier. By cancelling the partial derivatives of function L with respect to the unknown a_i vectors, i.e. $\partial L / \partial a_i = 0$, one obtains the matrix equation

$$(S - \lambda I)\, a_i = 0. \qquad (2)$$

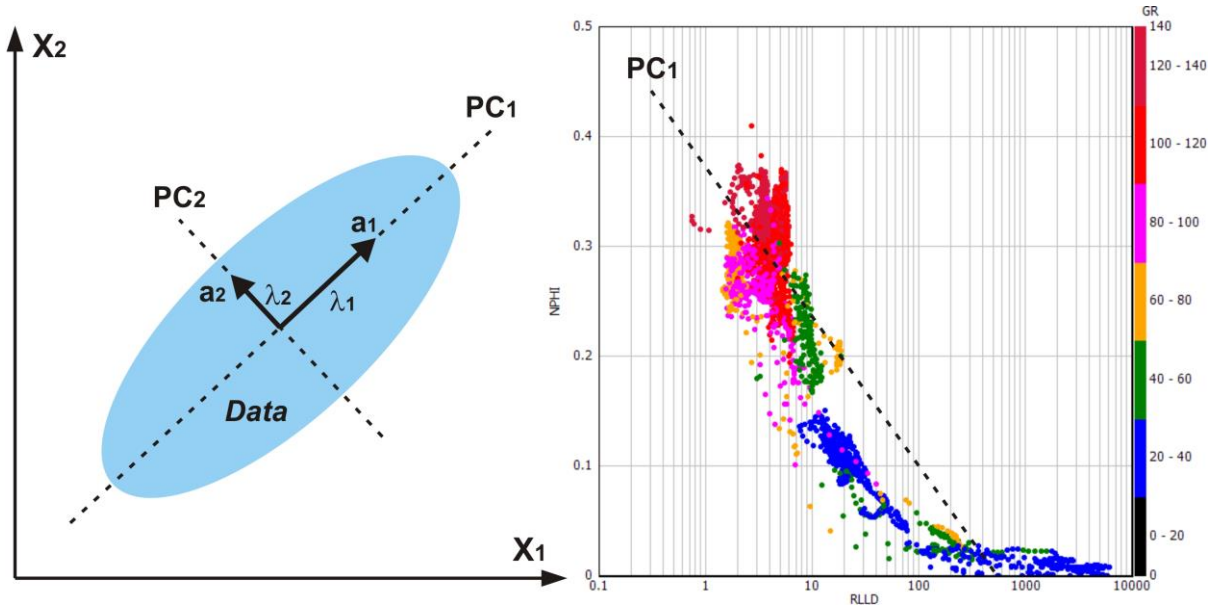

Figure 1 *Left*: *Idealized illustration of the PCA method for the case of two random variables x_1 and x_2. PCA finds the main variability directions in the data "cloud" and defines a new coordinate system, using optimal rotations. The axes of this system are defined by the eigenvectors a_1 and a_2. The eigenvalues λ_1 and λ_2 ($\lambda_1 \geq \lambda_2$) correspond to the data variance in the newly defined coordinate system. Right*: *Interdependency between two real random variables* (*geophysical logs recorded in the example well - compensated neutron porosity vs. deep resistivity*). *The main variability direction shown corresponds to the first principal component* (PC_1).

The characteristic equation $det(S - \lambda I) = 0$ has p roots (eigenvalues) λ_i, $i = 1, 2, \ldots, p$, such that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$. Once the eigenvalues λ_i are determined, the corresponding eigenvectors a_i can be computed by solving Eq. (2). For a p variables data set X, each a_i is a p-by-1 vector defining the axes of a new, rotated coordinates system that maximizes data variability along each axis (Fig. 1). PCA's results are usually expressed and interpreted in terms of *component scores* (z_i values corresponding to particular data points) and *loadings* (the components of each eigenvector a_i, i.e. $a_{i1}, a_{i2}, \ldots, a_{ip}$ from Eq. (1), acting as weighting factors of the original variables).

Case study: Gas exploration well, Moldavian Platform - Romania

An example of PCA applied to a suite of geophysical well logs recorded in a gas exploration well is presented in Table 1 and Fig. 2. From the complete wireline data set, the following logs were used for PCA: *GR* (I_γ) - total gamma ray intensity [API], *RLLD* (ρ_{LLD}) - Dual Laterolog deep resistivity [Ω·m], *RMLL* (ρ_{MLL}) - Microlaterolog resistivity [Ω·m], *NPHI* (Φ_N) - compensated neutron porosity [V/V], *DEN* (δ) - compensated bulk density [g/cm^3], *DT* (Δt) - compensated sonic transit time [μs/ft].

Table 1. Principal components of the geophysical logs covariance matrix.

	Variances explained by principal components (*eigenvalues*) [*% of total data variance*]					
	PC_1	PC_2	PC_3	PC_4	PC_5	PC_6
	81.69	12.26	2.81	1.54	1.01	0.69
	Component loadings (eigenvectors)					
	PC_1	PC_2	PC_3	PC_4	PC_5	PC_6
GR	0.37268	0.62829	-0.22511	0.06773	-0.38833	-0.51019
RLLD	-0.42330	0.22487	0.51642	0.54463	-0.43241	0.14128
RMLL	-0.40748	0.43385	0.30454	-0.20671	0.60248	-0.38377
NPHI	0.42738	0.25964	-0.03796	0.60030	0.53286	0.32279
DEN	-0.39694	0.47338	-0.52716	-0.19476	-0.06154	0.54657
DT	0.41912	0.27380	0.55727	-0.50772	-0.10731	0.41173

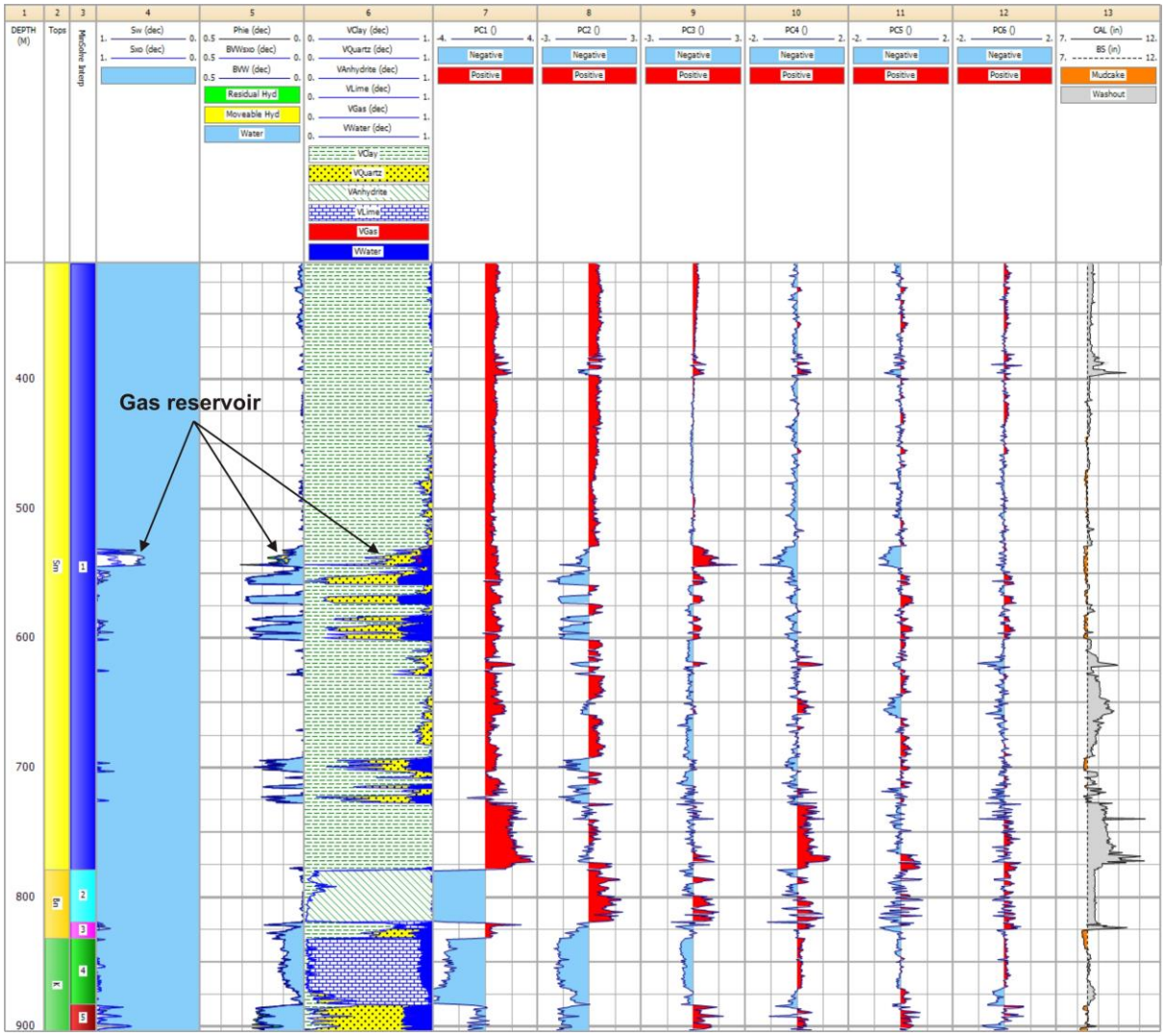

Figure 2 *Comparison between the results of conventional log interpretation* (*track 4 - uninvaded and flushed zones water saturations, track 5 - effective porosity and bulk volumes of fluids, track 6 - volumetric formation analysis*) *and the PCA results* (*tracks 7 to 12 - score logs of the principal components* PC_1 *to* PC_6). *Track 2 shows the actual formation tops* (*Sm - Sarmatian* (*Upper Miocene*), *Bn - Badenian* (*Middle Miocene*), *K - Cretaceous*) *and track 3 shows the zoning used for interpretation. The caliper log and bit size are presented in track 13, to show hole conditions. Log interpretation results were confirmed by flow tests, the uppermost Sarmatian sand producing dry gas. All other Sarmatian sands located below are water-bearing.*

The first principal component (PC_1) explains the largest part (81.69%) of the total variability in the data set, with approximately equal loadings (weights) for all logs. RLLD, RMLL and DEN (negative weights) are inversely correlated with GR, NPHI and DT (positive weights). In track 7 from Fig. 2, the depth intervals with positive PC_1 score log correspond to formations with high GR, NPHI and DT, but low RLLD, RMLL and DEN, while the negative PC_1 scores delineate the opposite case. PC_1 acts as a major lithological cut-off, separating the younger and/or less compact formations (Sarmatian and Badenian shales, sands and slightly cemented sandstones) from the older and/or compact, low-porosity and resistive formations (Badenian anhydrites, Cretaceous limestones and highly cemented sandstones). The second principal component (PC_2 - track 8) accounts for 12.26% of the total variability in the log suite, being dominated by the contribution of GR and subordinately by DEN. PC_2 may be interpreted as an accurate separator of porous-permeable intervals (negative score values), no matter their lithological composition, with respect to impermeable formations (positive score values) - shales or very compact rocks.

Higher-order components, like PC_3 to PC_5, respond more to fluids type and volume (or other fluid-related influences), as a result of significant RLLD and RMLL loadings in their eigenvectors structure. PC_3 (track 9) has higher loadings for RLLD, DEN and DT, DEN being inversely correlated with RLLD and DT. The positive PC_3 score values correspond to formations showing relatively high resistivity, low bulk density and high sonic transit time, i.e. good indicators of hydrocarbons presence (particularly light hydrocarbons, such as gas). The strong positive PC_3 "anomaly" noticeable in the 529-545 m interval (Sarmatian gas-bearing sand) correlates extremely well with the quantitative log interpretation and flow tests results. Note also the significant negative "anomalies" of PC_4 and PC_5 score logs in the same interval (especially the PC_5 one, negative only in the gas-bearing sand and positive in all other sands, which are water-bearing); most likely, this is related to the large neutron log contribution in the corresponding eigenvectors and to the low hydrogen index (neutron porosity) of gas with respect to formation water.

References

Barrash W., Morin R.H. (1997). Recognition of units in coarse, unconsolidated braided-stream deposits from geophysical log data with principal components analysis. *Geology*, **25** (8), 687-690.

Hotelling H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, **24** (6), 417-441.

Jolliffe I.T. (2002). *Principal Component Analysis, Second Edition*. Springer Series in Statistics, Springer-Verlag.

Kassenaar J.D.C. (1991). An application of principal components analysis to borehole geophysical data. *4th International MGLS / KEGS Symposium on Borehole Geophysics for Minerals, Geotechnical and Groundwater Applications*, Proceedings.

Moline G.R., Bahr J.M., Drzewjecki P.A., Shepherd L.D. (1992). Identification and characterization of pressure seals through the use of wireline logs: A multivariate statistical approach. *Log Analyst*, **34**, 362–372.

Morin R.H. (2006) Negative correlation between porosity and hydraulic conductivity in sand-and-gravel aquifers at Cape Cod, Massachusetts, USA. *Journal of Hydrology*, **316**, 43-52.

Pearson K. (1901). On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine*, **2**, 559-572.