Principal component analysis for assessing oil and gas production (the case of the Kogalym field)

G R Igtisamova¹, N N Soloviev², F A Ikhsanova¹, D Sh Nosirov¹ and A A Abdulmanov¹

¹Ufa State Petroleum Technological University, Branch of the University in the City of Oktyabrsky, 54a, Devonskaya St., Oktyabrsky, Republic of Bashkortostan, 452607, Russia
²LLC Gazprom VNIIGAZ
Staraya Basmannaya str., 20, b. 8, Moscow, Russia

Email: Igtisamova_galiy@mail.ru

Abstract. When analysing the criteria measuring the volume of oil deposits with wide gaps in configuration of reservoir parameters and physicochemical properties of fluid, it is necessary to group and characterize objects under study. Classification makes it possible to adjust conformity and distinctive features of deposits, and explain research theories. The analysis of information according to the subjects determined by the parameters measured or evaluated is difficult to carry out. It requires a lot of time and effort. Therefore, it is necessary to reduce the data volume, compress initial information to the smallest number of characteristics. Parameters can be selected from initial data or calculated and modified (i.e. minimum loss of data on the objects under study). The effective analysis tool capable of identifying the problems is the principal component analysis (PCA) which is a method for reducing the data volume. The principal component can be found in almost every text using the multivariate analysis.

1. Introduction
When analysing the criteria measuring the volume of oil deposits [1] with wide gaps in the configuration of the reservoir parameters of reservoirs and physicochemical properties of fluid, it is necessary to group and characterize objects according to homogeneous groups [2, 3]. Classification makes it possible to adjust the conformity and distinctive feature of deposits, explain research theories. The analysis of information according to the subjects determined by the parameters measured or evaluated is difficult to carry out. It requires a lot of time and effort [4]. Therefore, it is necessary to reduce the data volume, compress initial information to the smallest number of characteristics. Parameters can be selected from initial data or calculated and modified (i.e. minimum loss of data on the objects under study). The effective analysis tool capable of identifying the problems is the principal component analysis (PCA) which is a method for reducing the data volume. The principal component analysis (PCA) is a method for reducing data volume. The principal component can be found in almost every text using the multivariate analysis [5]. The examples of its applications are data compression, image processing, visualization, search data analysis, pattern recognition and time series prediction.
The most common finding of the principal component analysis is the standardized linear projection which maximizes variance in the projected space.

For a set of observed $d$-dimensional vector data $\{t_n\}, n \in \{1 \ldots N\}$, $q$ of principal axes $w_j$, $j \in \{1 \ldots q\}$, are the orthonormal axes on which the maximum retained dispersion in the projection is maximum. Vectors $w_j$ are set by $q$ which are dominant vectors (i.e., they have maximum related values $\lambda_j$) of the correlation matrix $S = E[(t - \mu)(t - \mu)^T]$ so that $Sw_j = \lambda_j w_j$. $q$ principal components of observed vector $t_n$ are set by vector $x_n = W^T(t_n - \mu)$, where $W^T = (w_1, w_2, \ldots, w_q)^T$ [6]. Variables $x_j$ are interchanged so that the diagonal of the correlation matrix $E[xx^T]$ consists of elements $\lambda_j$.

An additional characteristic of the PCA related to the Pearson coefficient is the fact that the projection of the principal component minimizes only the square error permutation $\Sigma_n \|t_n - \tilde{t}_n\|^2$, rather than all orthogonal linear projections $x_n = W^T(t_n - \mu)$. The optimal linear permutation $t_n$ is determined by formula $\tilde{t}_n = Wx_n + \mu$.

One of the disadvantages of the PCA is a lack of the probability density model and a related likelihood measure [7]. The PCA based on density estimation has a number of advantages, including:

- the use of a likelihood measure allows for comparison with other density estimation methods and facilitates statistical testing;
- Bayesian inference methods can be used (for example, to compare models) by combining probability with the previous one;
- if the PCA is used to simulate a conditional class of densities in the classification problem, one can calculate probabilistic probabilities of class membership;
- the probability density function gives a measure of novelty of the new data point;
- a single PCA model can be expanded to combination of these models.

2. Results and Discussion

The key result of this work is to show that the PCA is based on the probabilistic model. This follows from inclusion of $W$ in a model of hidden variable density which is closely related to the analysis of statistical factors. Using an iterative algorithm, maximum likelihood estimate $W$ [8] is more computational than the standard estimate with its own decomposition. However, the use of $W$ can be calculated in a standard way using differential equations [9-11] and be included in the model to implement the advantages listed above.

We will present the models with a hidden variable and with a factor analysis. Then, it will be shown how the PCA arises from parameterization of the model applied when calculating and classifying wells according to their specific parameters.

Hidden model variables

The hidden variable model links a set of dimensional observable vectors $\{t_n\}$ of data with a set of $q$-dimensional hidden variables $\{x_n\}$:

$$t = y(x; \theta) + \varepsilon,$$

where $y(x)$ is the function of hidden variable $x$; parameters $\theta$ and $\varepsilon$ characterize an independent noise process. As a rule, $q < d$, hidden variables provide a more detailed description of the data. Through preliminary distribution by $x$, equation (1) induces corresponding distribution in the data space, and the model parameters can be determined by maximum likelihood.

In the standard factor analysis, mapping $y(x; \theta)$ is linear:

$$t = Wx + \mu + \varepsilon,$$

Where hidden variables $x \sim N(0, I)$ have a single isotropic Gaussian distribution. The error or noise model is Gaussian, so that $\varepsilon \sim N(0, D)$, with $D$ diagonal, the matrix of parameters $d \times q$ $W$ contains factor loads and $\mu$ is a constant whose maximum likelihood estimate is the average of the data. Given this expression, the model for t is also the norm $N(\mu, C)$, where correlation $C = D + WW^T$. 

2
The main initial material for determining hidden variables is the dispersion-correlation matrix. Let us assume that we know the characteristics of the wells at deposits N determined by the spatial vector \( X \{x_1, x_2, x_3, ..., x_n \} \). According to the dispersion-correlation matrix (interconnected matrix), we find the vectors. The principal components establish the total dispersion and its significance characterizing the impact (risk) of any factor. The number of shaded conditions is determined by the number of eigenvalues (Figure 1).

![Scree plot](image)

**Figure 1.** Dependence of eigenvalues on variability and aggregate.

![Cumulative variability](image)

**Figure 2.** Dependence of eigenvalues on variability and aggregate.

Indeed, a key assumption for this model: because of the diagonal \( D \), the observable variables \( t \) are conditionally independent, given the values of hidden variables \( x \). Thus, the dimensional distribution \( x \) is designed to simulate the dependencies between the observed variables and is independent noise. This contrasts with the PCA which is used for the dependencies between different variables and independent noise. In the factor analysis, \( W \) columns usually do not correspond to the main data subspace. In addition, unlike the PCA, for \( W \), there is no analytical solution, and therefore their values should be determined by iterative procedures. Because of the term \( WW^\tau \) correlation \( C \) and probability are invariant with respect to the orthogonal after multiplying \( W \). That is, \( WR \), where \( R \) is an arbitrary orthogonal matrix, gives equivalent \( C \).

Due to the diagonal noise model \( D \), factor loads \( W \) are generated by the principal axes (even if arbitrary rotation is taken into account). According to many researchers, the principal components arise when it is assumed that the data contain a systematic component and an independent error term for each variable with a common variance \( \sigma^2 \). This means that the diagonal elements of the error matrix \( D \) in the factor analysis should be equivalent. Indeed, similarity of factor loads and principal axes is observed in places where the elements \( D \) are approximately equal.

These observations do not take into account the maximum likelihood context. Considering the model represented by formula (2) with an isotropic noise structure \( D = \sigma^2 I \), we show that even when the correlation model is approximate, the maximum (likelihood estimator \( W_{ML} \) is a matrix whose columns are scaled and rotated in the primary eigenvectors of the covariance matrix \( S \). The important consequence is that the PCA can be expressed in terms of the density model.

For the isotropic noise model \( \epsilon \sim N(0, \sigma^2 I) \) from equation (2), distribution of probabilities by \( t \) can be determined by formula

\[
p(t|x) = (2\pi \sigma^2)^{-d/2} \exp \left\{ -\frac{1}{2\sigma^2} \| t - Wx - \mu \|^2 \right\}.
\]

(3)

By hidden variables
\[ p(x) = (2\pi)^{-d/2} \exp \left\{ -\frac{1}{2} x^T x \right\}, \]  
we have limit distribution \( t \) in the form
\[ p(t) = \int p(t|x) p(x) dx, \]
\[ p(t) = (2\pi)^{-d/2} |C|^{-1/2} \exp \left\{ -\frac{1}{2} (t - \mu)^T C^{-1} (t - \mu) \right\}, \]
where model correlation
\[ C = \sigma^2 I + WW^T. \]  
Using the Bayes’ rule, it is possible to calculate the subsequent distribution of hidden variables, taking into account the observed \( t \):
\[ p(t|x) = (2\pi)^{-d/2} \sigma^{-2} M^{\frac{1}{2}} \exp \left\{ -\frac{1}{2} \{x - M^{-1} W^T (t - \mu)\}^T (\sigma^{-2} M) \{x - M^{-1} W^T (t - \mu)\} \right\}, \]
where the back correlation matrix is determined by formula
\[ \sigma^2 M^{-1} = \sigma^2 (\sigma^2 I + W^T W)^{-1}. \]
\( M \) is \( q \times q \), and \( C \) - \( d \times d \).
Logarithmic probability of observing data from this model is
\[ L = \sum_{n=1}^{N} \ln(p(t_n)), \]
\[ L = -\frac{Nd}{2} \ln(2\pi) - \frac{N}{2} \ln|C| - \frac{N}{2} \text{tr}[C^{-1} S], \]
\[ S = \frac{1}{N} \sum_{n=1}^{N} (t_n - \mu)(t_n - \mu)^T. \]

Maximum Likelihood Estimation Properties

Log-likelihood (10) is maximized when columns \( W \) cover the main data subspace. To show this, let us consider the derivative of (10)
\[ \frac{DL}{DW} = N(C^{-1} SC^{-1}W - C^{-1}W), \]
which can be obtained from the results of the standard matrix discharge. It also shows that the only non-zero stationary points (12) take place for:
\[ W = U_q\left(A_q - (\sigma^2 I)^{\frac{1}{2}}R\right), \]

Where vectors \( q \) of the column in \( U_q \) are eigenvectors from \( S \) with eigenvalues in the diagonal matrix \( A_q \), and \( R \) is the arbitrary \( q \times q \) matrix of the orthogonal rotation. In addition, the stationary point corresponding to the global maximum of probability occurs when \( U_q \) contains the principal eigenvectors \( S \) and all other combinations of the eigenvectors represent the beginning (likelihood surface points). Thus, from (13), columns of estimation of maximum probability \( W_{ML} \) contain eigenvectors \( S \) with a scale determined by the corresponding eigenvalue and parameter \( \sigma^2 \), with arbitrary rotation.

At \( W = W_{ML} \) maximum probability for \( \sigma^2 \) is determined by formula
\[ \sigma^2_{ML} = \frac{1}{d-q} \sum_{j=q+1}^{d} \lambda_j, \]
which is interpreted as dispersion of the “loss” in the projection, averaged by lost measurements.

Columns \( W_{ML} \) are not orthogonal as
\[ (W_{ML})^T W_{ML} = R^T (A_q - (\sigma^2 I) R), \]
which is not diagonal for \( R \neq I \). Other PCA algorithms are elements of the rotational ambiguity. The orthonormal basis for the principal subspace can be easily extracted using standard methods. In addition, the actual principal axes can be determined if equation (15) is an expansion by eigenvector \( (W_{ML})^T W_{ML} \), where transposed rotation matrix \( R^T \) is a matrix whose columns are eigenvectors of matrix \( q \times q \) \( (W_{ML})^T W_{ML} \).
However, referring to the optimal property of PCA recovery, further parameter processing is not needed. (8) shows that back middle projection $t_n$ is determined by $\langle x_n \rangle = M^{-1}W^T(t_n - \mu)$. When $\sigma^2 \rightarrow 0, M^{-1} \rightarrow (W^TW)^{-1}$ and $WM^{-1}W^T$ becomes an orthogonal projection. Therefore, the probability model for the PCA is recovered. However, the density model becomes singular and indexed. At $\sigma^2 > 0$ the diversity projection is distorted in the beginning of coordinates due to $x$. $W(x_n)$ is not orthogonal projection $t_n$. However, each data point can be recovered from a hidden variable. At $W = W_{ML}$ the value is determined by formula

$$t_n = W_{ML}\{W_{ML}^TW_{ML}\}^{-1}M(x_n).$$

(16)

**Table 1.** Kogalym field formation parameters.

| Objects | $m_g$, % | $m_k$, % | $K_n$, m$^3$/days·MPa | $\mu_n$, mD | $\mu_0$, mD | $\rho_n$, kg/m$^3$ |
|---------|----------|----------|------------------------|-------------|-------------|----------------|
| 1       | 14.9     | 14.9     | 0.73                   | 155.3       | 119.7       | 914            |
| 2       | 14.5     | 14.5     | 0.73                   | 72.3        | 52.3        | 902            |
| 3       | 14.4     | 14.6     | 0.69                   | 40.4        | 30          | 894            |
| 4       | 14.1     | 14       | 0.77                   | 6.8         | 5           | 848            |
| 5       | 21.7     | 22       | 0.78                   | 10.3        | 7.2         | 857            |
| 6       | 17       | 16.2     | 0.76                   | 10          | 7           | 868            |
| 7       | 16.7     | 16.7     | 0.83                   | 3.2         | 2.5         | 813            |
| 8       | 19.1     | 18.9     | 0.86                   | 4           | 3.2         | 805            |
| 9       | 12.8     | 12.9     | 0.68                   | 36.9        | 23          | 886            |
| 10      | 11.2     | 11.2     | 0.72                   | 22          | 15          | 877            |
| 11      | 11.5     | 11.3     | 0.78                   | 14.7        | 10.5        | 870            |
| 12      | 6.9      | 6.2      | 0.7                    | 12.8        | 8.8         | 869            |
| 13      | 11.5     | 12.9     | 0.84                   | 22.8        | 13.9        | 892            |
| 14      | 10.6     | 9.9      | 0.81                   | 6.6         | 5.0         | 844            |
| 15      | 12.5     | 12.1     | 0.87                   | 5.2         | 4.2         | 835            |
| 16      | 12       | 12.3     | 0.83                   | 2.2         | 1.9         | 785            |
| 17      | 4.4      | 4        | 0.79                   | 6.5         | 5           | 842            |
| 18      | 11.4     | 12.5     | 0.85                   | 2.4         | 1.9         | 793            |

**Table 2.** Formation parameters influencing the well flow rate (after the PCA)

| F1   | F2   | F3   | F4   | F5   | F6   | F7   |
|------|------|------|------|------|------|------|
| $N_t$, m | -0.256 | -0.038 | 0.773 | -0.534 | 0.218 | 0.055 | 0.007 |
| $m_g$, % | 0.071 | 0.686 | -0.082 | -0.132 | -0.008 | 0.687 | 0.171 |
m₉, %
0.062  0.688  -0.019  -0.158  0.051  -0.682  -0.173

K, m³/days ∙ MPa
-0.414  0.194  0.234  0.717  0.470  0.018  0.009

μₙ, mD
0.506  0.037  0.384  0.246  -0.142  -0.172  0.697

μ₀, mD
0.499  0.042  0.399  0.270  -0.183  0.172  -0.674

ρₙ, kg/m³
0.500  -0.121  -0.183  -0.158  0.822  0.033  0.024

**Figure 3.** The graph of correlation between principal components and parameters of the objects

**Figure 4.** The graph of correlation between the principle components and parameters of the objects

**Figure 5.** Distribution of research objects by the axes of principal components F1 – F2 (a), F1 – F3 (b), F1 – F4 (c)

Thus, the hidden variables convey the necessary information for optimal recovery of the original data vector, even at $\sigma^2 > 0$ (Tables 1, 2, Figures 3, 4, 5).

In addition, we developed an algorithm for determining parameters of the model. The algorithm plays a crucial role in transforming the approach to linear transformation into a new set of “independent” random variables which are arranged in an descending order of their variances. The PCA are determined by the eigenvectors of the correlation matrix (the iterative approach for large d
can be used for the standard PCA, since the algorithm requires inversion of the \( q \times q \) matrix, in contrast to the full decomposition of the \( d \times d \) correlation matrix. However, in such cases there are other iterative algorithms.

Instead of identifying the principal components, we emphasized the advantages of correlation of the probabilistic model with the PCA. In many applications, these advantages can be implemented by calculating \( U_q \) and \( A_q \) (decomposition of the correlation matrix and inclusion of these parameters in the probability model using equations (13) and (14)), avoiding the use of the algorithm.

In practice, correlation of isotropic noise \( \sigma^2 I \) within the model is more advantageous than diagonal correlation \( D \) used in the standard factor analysis. In the latter method, it is necessary to pay attention to the number of factors \( q \). Inappropriate choice can be misleading [4]). The main problem is that if the observations can be explained by two factors, a model that identifies only one factor cannot meet with one of two. This results from incorrect \( q \) specification which is compensated for in the factor loads \( W \), which is not typical of the PCA. In this case, the use of the isotropic noise model implies that the first two main axes will obviously include only one parameter.

3. Conclusion
The article showed that the PCA is a maximum likelihood procedure used for grouping oil deposits in the Kogalym field based on the probability density model of the observed data. The PCA based on the probability density model of the observed data allowed for selection of four relatively homogeneous groups using only 7 factors (principal components) (Figures 3, 4, 5) from the initial data set containing 18 research objects characterized by 12 parameters which contain more information than the individual parameters (Tables 1, 2 and Figure 2).

One more important advantage of the density simulation (using individual or mixed models) is the ability to control complexity of the model by choosing \( q \) by limiting the number of parameters used to denote the correlation structure. This makes it possible to build density models in high dimensional spaces and avoid inappropriate diagonal or spherical constraints. Thus, classification by simulating conditional densities can be used even when \( d \) is large.

In addition to the traditional PCA based on the general statistical basis, probabilistic formalism generates other density estimation methods with a wide range of possibilities for practical application. Illustrative examples from the previous paragraph emphasize that the model is promising.

References
[1] Anderson T W 1963 Asymptotic theory for principal component analysis Ann. Math. Stat. 34 122–48
[2] Bartholomew D J 1987 Latent Variable Models and Factor Analysis (London: Charles Griffin & co. Ltd)
[3] Bahtizin R N, Nurgaliev R Z, Fattakhov I G, Andreev V E and Safiullina A R 2018 On the question of the efficiency analysis of the bottom-hole area stimulation method Int. J. of Mechanical Engineering and Technology (IJMET) 9(6) 1035–44
[4] Bishop C M 1995 Neural Networks for Pattern Recognition (Oxford: Clarendon Press.)
[5] Dempster A P, Laird N M and Rubin D B 1977 Maximum likelihood from incomplete data via the EM algorithm J. R. Stat. Soc. B 39(1) 1–38
[6] Hotelling H 1933 Analysis of a complex of statistical variables into principal components J. Educ. Psychol. 24 417–41
[7] Igtisamova G R and Nosirov D Sh 2018 Peculiarities of problem solving at studying well drilling with the use of linear differential equations with constant coefficients Advances in Engineering Research (AER) (Int. conf. “Actual issues of mechanical engineering” (AIME 2018) vol 157) pp 211–14
[8] Tyncherov K T and Selivanova M V 2017 Modular wavelet approach for reinforced plastic pipelines operation quality evaluation Int. Conf. Industrial Engineering, Applications and Manufacturing (ICIEAM) (St. Petersburg: IEEE)
[9] Mukhametshin V V 2018 Rationale for trends in increasing oil reserves depletion in Western Siberia cretaceous deposits based on targets identification Bull. of the Tomsk Polytechnic Univer. Geo Assets Engineering 329(5) 117–24

[10] Tyncherov K T, Mukhametshin V Sh, Paderin M G, Selivanova M V, Shokurov I V and Almukhametova E M 2018 Thermoacoustic inductor for heavy oil extraction IOP Conf. ser. Mat. Sci. 327(4) 042111

[11] Yakupov R F, Mukhametshin V Sh and Tyncherov K T 2018 Filtration model of oil coning in a bottom water-drive reservoir Periodico Tche Quimica 15(30) 725–33