Adv Syst Sci Appl 2020; **02**:98–118
Published online at https://ijassa.ipu.ru.

Student Mixture and Its Machine Learning Applications to PVT Properties of Reservoir Fluids

Nikita Volkov[1,2*], Elizaveta Dakhova[1,2], Semen Budennyy[1,2], Alla Andrianova[3]

[1]*Moscow Institute of Physics and Technology, Moscow, Russia*

[2]*Center for engineering and technology of MIPT, Moscow, Russia*

[3]*Gazprom Neft Science and Technology Center, St.-Petersburg, Russia*

Abstract: Distribution mixture models are widely used in cluster analysis. Particularly, a mixture of Student t-distributions is mostly applied for robust data clustering. In this paper, we introduce EM algorithm for a mixture of Student distributions, where at the E-step, we apply variational Bayesian inference for parameters estimation. Based on the mixture of Student distributions, we construct a machine learning method that allows to solve regression problems for any set of features, clustering, and anomaly detection within one model. Each of these problems can be solved by the model even if there are missing values in the data. The proposed method was tested on real data describing the PVT properties of reservoir fluids. The results obtained by the model do not contradict the basic physical properties. In majority of conducted experiments our model gives more accurate results than well-known machine learning methods in terms of MAPE and RMSPE metrics.

Keywords: Student mixture, EM algorithm, variational Bayesian inference, clustering, regression, anomalies, missing values, PVT properties

1. INTRODUCTION

Normal distributions often occur in various data analysis problems and they are fairly well studied. However, their disadvantage for evaluating parameters is their light distribution tails. In the presence of outliers, as it usually occurs in real problems, the parameter estimates are strongly biased towards outliers. To eliminate this disadvantage, the Student distribution (or T-distribution) is often considered, because its properties are similar to those of the normal distribution, but it has heavy tails. Thus, the Student distribution has a certain degree of stability to emissions.

The properties of the Student distribution were first studied by William Gossett. The author has published his first results on that under the pseudonym *Student*. Gosset noted that the distribution of the standardized (centered and scaled) normal sample average where the unknown variance is replaced with its estimation is different from the normal one [1]. There are many other theoretical properties of the Student distribution. All the most important of them used in the paper are given in Section 2.

Mixtures of normal distributions are often used to describe data. Parameters of such a mixture are usually estimated using the EM algorithm [6]. For a description of the EM algorithm and some theoretical properties of a mixture of distributions, see Section 3. If there are outliers in data, it is natural to consider a mixture of Student distributions. Some ideas

[*] Corresponding author: volkov.na@cet-mipt.ru

for the parameters estimation of the Student mixture were described in [7], [8], [9] and [12]. In particular, [7] describes a conditional EM algorithm. In this paper, Section 4 provides a complete derivation of the parameter estimation method using a similar variation of the EM algorithm, which optimizes the variational Bayesian inference at the E-step (see [6]).

This probabilistic model for data description has many practical applications, which allow it to be flexibly configured and conduct extensive data analytics. Let us list the applications discussed in detail in Section 5:

1. Clustering with a predefined number of clusters.
2. Detecting anomalies.
3. Regression to predict any set of real features using any other set of features.
4. Filling missing values.

This large number of applications is due to the fact that the mixture model is generative, since it describes the joint distribution of all features. This distribution also allows to create artificial data.

The model of a mixture of distributions can be recommended to solve problems with expected continuous dependence of features between each other, for example, physical problems. In Sections 6 and 7, we apply our model to PVT properties of reservoir fluids, where it shows high quality relative to widely known machine learning models. It should be remarked that, the obtained experimental results do not contradict physical laws, unlike outcomes of many machine learning methods, in particular those based on decision trees.

2. DISTRIBUTIONS AND THEIR PROPERTIES

This section provides definitions and some basic properties of the normal and Student distributions that are in the basis of the developed model. The properties of the gamma distribution are also given as they are to represent Student random vector in a convenient form. All statements in this section can be found, for example, in the literature [1], [2], [3], [4], [5], [6].

2.1. Normal distribution

The density of the multidimensional normal distribution centered at a point $\mu \in \mathbb{R}^d$ and a symmetric positive definite covariance matrix Σ equals

$$q(x|\mu, \Sigma) = \frac{1}{(2\pi)^{d/2}\sqrt{\det \Sigma}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right),$$

where $det\Sigma$ means the determinant of matrix Σ.

We will use this notation for the normal distribution density throughout.

Let matrix B be the root of the matrix Σ that is satisfying the condition $BB^T = \Sigma$. Then, if ξ has distribution $\mathcal{N}(0, I_d)$, where I_d is identity matrix of dimension d, then $\eta = \mu + B\xi$ has distribution $\mathcal{N}(\mu, \Sigma)$.

2.2. Gamma distribution

The gamma distribution density $\Gamma(\alpha, \beta)$ is

$$\gamma(x|\alpha, \beta) = \frac{\alpha^\beta}{\Gamma(\beta)} x^{\beta-1} e^{-\alpha x} I\{x \geqslant 0\}.$$

We will use this notation for the density of gamma distribution throughout.

Let $\xi \sim \Gamma(\alpha, \beta)$. It is not difficult to make sure that $\mathsf{E}\xi = \beta/\alpha$ [1], $\mathsf{E}\ln\xi = \psi(\beta) - \ln\alpha$, where $\psi(x) = \frac{d\ln\Gamma(x)}{dx}$ is digamma function. If $\beta > 1$ we also get $\mathsf{E}\xi^{-1} = \frac{\alpha}{\beta-1}$.

2.3. Student distribution

The Student distribution has parameter ν indicating the number of degrees of freedom. The less the number of degrees of freedom, the heavier Student distribution tails, and the more resistant it is to outliers. Moreover, the Student distribution converges in distribution to the normal distribution when the nu converges to infinity.

We denote multidimensional Student distribution with $\nu > 0$ degrees of freedom, with mean vector $\mu \in \mathbb{R}^d$ and symmetric positive semi-definite and non-degenerate scale matrix Σ by $T_\nu(\mu, \Sigma)$. The density of this distribution in $x \in \mathbb{R}^d$ equals

$$p(x|\mu, \Sigma, \nu) = \frac{\Gamma\left(\frac{\nu+d}{2}\right)}{\Gamma(\nu/2)\nu^{d/2}\pi^{d/2}|\Sigma|^{1/2}} \left[1 + \frac{1}{\nu}(x-\mu)^T\Sigma^{-1}(x-\mu)\right]^{-\frac{\nu+d}{2}}.$$

The mathematical expectation and the covariance matrix of the distribution $T_\nu(\mu, \Sigma)$ equal $\mathsf{E}X = \mu$ if $\nu > 1$ and $\mathrm{Var}X = \frac{\nu}{\nu-2}\Sigma$ if $\nu > 2$ respectively.

Claim 2.1:
[5] Let random vector ξ and random variable η be independent and have distributions $\mathcal{N}(0, \Sigma)$ and $\Gamma(\nu/2, \nu/2)$ respectively. Let $\mu \in \mathbb{R}^d$ be a fixed vector. Than random vector $X = \mu + \xi/\sqrt{\eta}$ has Student distribution with ν degrees of freedom, mean vector μ and scale matrix Σ.

The Student distribution density can be represented in an integral form

$$p(x|\mu, \Sigma, \nu) = \int_0^{+\infty} q(x|\mu, \Sigma/y)\gamma(y|\nu/2, \nu/2)dy.$$

2.4. Marginal distributions

Let a, b be disjoint sets of indexes, and $a \sqcup b = \{1, ..., d\}$. Without any loss of generality, we set $a = \{1, ..., d_a\}, b = \{d_a + 1, ..., d\}, d_b = d - d_a$. Vectors and matrices are represented as follows

$$X = \begin{pmatrix} X_a \\ X_b \end{pmatrix}, \quad \mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ab}^T & \Sigma_{bb} \end{pmatrix},$$

where $X_a \in \mathbb{R}^{d_a}, X_b \in \mathbb{R}^{d_b}, \mu_a \in \mathbb{R}^{d_a}, \mu_b \in \mathbb{R}^{d_b}, \Sigma_{aa} \in \mathbb{R}^{d_a \times d_a}, \Sigma_{ab} \in \mathbb{R}^{d_a \times d_b}, \Sigma_{bb} \in \mathbb{R}^{d_b \times d_b}$. The following property is known [10].

Claim 2.2:
Let random vector X have normal distribution $\mathcal{N}(\mu, \Sigma)$. Then random vector X_a has normal distribution $\mathcal{N}(\mu_a, \Sigma_{aa})$. If random vector X has Student distribution $T_\nu(\mu, \Sigma)$, then random vector X_a has Student distribution $T_\nu(\mu_a, \Sigma_a)$.

2.5. Conditional distribution

Our model of probability distributions mixture allows to build a regression on arbitrary features using a conditional distribution. In this section, we recall relations for parameters of conditional distribution for a normal vector (Claim 2.3) and for a Student vector (Theorem 2.1).

Let random vector X have normal distribution $\mathcal{N}(\mu, \Sigma)$, where Σ is positive semi-definite scale matrix. Moreover let a, b, c are disjoint sets of indexes, and $a \sqcup b \sqcup c = \{1, ..., d\}$. Without loss of generality, set $a = \{1, ..., d_a\}, b = \{d_a + 1, ..., d_a + d_b\}, c = \{d_a + d_b + 1, ..., d\}$. Vectors and matrices are represented as follows

$$X = \begin{pmatrix} X_a \\ X_b \\ X_c \end{pmatrix}, \quad \mu = \begin{pmatrix} \mu_a \\ \mu_b \\ \mu_c \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} & \Sigma_{ac} \\ \Sigma_{ab}^T & \Sigma_{bb} & \Sigma_{bc} \\ \Sigma_{ac}^T & \Sigma_{bc}^T & \Sigma_{cc} \end{pmatrix}.$$

Denote also

$$\Lambda = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ab}^T & \Lambda_{bb} \end{pmatrix} = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ab}^T & \Sigma_{bb} \end{pmatrix}^{-1}.$$

Claim 2.3:
Vector X_a conditioned on X_b has distribution $\mathcal{N}(\widetilde{\mu}, \widetilde{\Sigma})$, where

$$\widetilde{\mu} = \mu_a - \Lambda_{aa}^{-1}\Lambda_{ab}(X_b - \mu_b), \quad \widetilde{\Sigma} = \Lambda_{aa}^{-1}.$$

The proof of this statement is given in [10]. It follows that the conditional distributions of the components of a normal vector are also normal. Let us now consider the case in which the vector X has a Student distribution $T_\nu(\mu, \Sigma)$ [5].

Theorem 2.1:
Vector X_a conditioned on X_b has distribution $T_{\widetilde{\nu}}(\widetilde{\mu}, \widetilde{\Sigma})$, for where

$$\widetilde{\nu} = \nu + d_b, \quad \widetilde{\mu} = \mu_a - \Lambda_{aa}^{-1}\Lambda_{ab}(X_b - \mu_b), \quad \widetilde{\Sigma} = \frac{\nu + \varphi(X_b)}{\nu + d_b}\Lambda_{aa}^{-1},$$

$$\varphi(x) = (x - \mu_b)^T \left(\Lambda_{bb} - \Lambda_{ab}^T\Lambda_{aa}^{-1}\Lambda_{ab}\right)(x - \mu_b).$$

Remark. Function $\varphi(x)$ is non-negative since all angular minors of the matrix $\Lambda_{bb} - \Lambda_{ab}^T\Lambda_{aa}^{-1}\Lambda_{ab}$ are positive. The latter fact follows from the block matrix determinant formula [11] and positive definiteness of the scale matrix.

3. MIXTURE MODEL

In this section, we formalize the concept of a mixture of probability distributions and give its properties that appear helpful for solving machine learning problems mentioned in Introduction. In addition, iterative methods for estimating parameters of a mixture of normal distributions and a mixture of Student distributions are provided.

3.1. Properties of a mixture model

Consider a mixture of probability distributions

$$\mathsf{P} = \sum_{j=1}^{k} w_j \mathsf{P}_j,$$

where P_j is a "simple" probability distribution that defines the component of the mixture and $w_j \in [0, 1]$ are components weights, $\sum_{i=1}^{n} w_j = 1$. These components are often called clusters. A random vector X obeys the model of a mixture of distributions if it is represented as

$$X = \sum_{j=1}^{k} X^j I\{T = j\},$$

where X^j is distributed as P_j and random variable T is equal to the cluster number, i.e. $\mathsf{P}(T = j) = w_j$. Moreover, variables X^j are independent of T. Next statement can be found in [12].

Claim 3.1:
Let $\mathsf{E}X^j = \mu_j, \mathrm{Var}X^j = \Sigma_j$ (their existence is assumed). Then the expectation and

covariance matrix for a mixture of distributions are equal to

$$\mathsf{E}X = \sum_{j=1}^{k} w_j \mu_j, \quad \mathrm{Var}X = \sum_{j=1}^{k} w_j \Sigma_j + \sum_{j=1}^{k} w_j \mu^j (\mu^j)^T - \sum_{j,s=1}^{k} w_j w_s \mu^j (\mu^s)^T.$$

Claim 3.2:
Let $X^T = \ldots \left(X_a^T, X_b^T, X_c^T\right)$. Then conditional probability of the cluster j conditioned on X_b equals

$$\widetilde{w}_j = \mathsf{P}(T = j \mid X_b) = \frac{w_j p_j^{(b)}(X_b)}{\sum_{j=1}^{k} w_j p_j^{(b)}(X_b)},$$

where $p_j^{(b)}$ is the density of the vector component X_b conditioned on P_j.

Let vector X_a has distribution $\widetilde{\mathsf{P}}_j^{(a|b)}$ conditioned on X_b and $I\{T = j\}$. Then vector X_a conditioned on X_b has a distribution of mixture of distributions $\widetilde{\mathsf{P}}_j^{(a|b)}$ with weights $\widetilde{w}_j$

$$\widetilde{\mathsf{P}}^{(a|b)} = \sum_{j=1}^{k} \widetilde{w}_j \widetilde{\mathsf{P}}_j^{(a|b)}.$$

3.2. Normal mixture distribution

The density in the model of a mixture of normal distributions equals

$$p(x) = \sum_{j=1}^{k} w_j q(x|\mu_j, \Sigma_j).$$

Let $X_1, \ldots, X_n$ be a sample from such a mixture of distributions. The estimation of mixture parameters is performed by solving the problem of maximizing model likelihood using an iterative EM algorithm [6]. This procedure consists in selecting some random initial approximation of the parameters and then alternating two steps. For a mixture of normal distributions they are conducted in the following way:

E-step. Compute the following auxiliary values

$$r_{ij} = \frac{w_j q(X_i|\mu_j, \Sigma_j)}{\sum\limits_{s=1}^{k} w_s q(X_i|\mu_s, \Sigma_s)}.$$

r_{ij} is the probability that the random vector X_i is obtained from the jth component of the mixture at the current approximation of the parameters w_j, μ_j, Σ_j.

M-step. Compute a new approximation of parameters

$$w_j = \frac{1}{n} \sum_{i=1}^{n} r_{ij}, \qquad \mu_j = \sum_{i=1}^{n} r_{ij} X_i \Bigg/ \sum_{i=1}^{n} r_{ij}, \qquad \Sigma_j = \sum_{i=1}^{n} r_{ij} (X_i - \mu_j)^2 \Bigg/ \sum_{i=1}^{n} r_{ij}.$$

Stopping criterion. Iterations of the method are made up to the convergence of the variational lower bound on the logarithmic likelihood function

$$\mathcal{L}(w, \mu, \Sigma, r) = \sum_{i=1}^{n} \sum_{j=1}^{k} r_{ij} \left[\ln w_j + \ln q(X_i|\mu_j, \Sigma_j)\right] - \sum_{i=1}^{n} \sum_{j=1}^{k} r_{ij} \ln r_{ij}.$$

The procedure terminates as $\mathcal{L}$ changes in no more than a pre-set small number $\varepsilon > 0$ [6].

3.3. Student mixture distribution

The density of a mixture of multidimensional Student distributions equals

$$p(x) = \sum_{j=1}^{k} w_j p(x|\mu_j, \Sigma_j, \nu),$$

where $p(x|\mu_j, \Sigma_j, \nu)$ is the density of multidimensional Student distribution with ν degrees of freedom centered at μ_j and the scale matrix Σ_j. Parameter ν is a hyperparameter of the model.

Let $X = (X_1, \ldots, X_n)$ be sample vectors from that mixture distribution. The proposed method for estimating the parameters of the mixture consists in selecting some initial approximation of the parameters and performing the next steps at each iteration.

E-step. Perform several iterations of the next two steps:

I. Compute the following auxiliary values

$$r_{ij} = \frac{w_j q\left(X_i\,|\mu_j, \Sigma_j a_i/b_i\right)}{\sum\limits_{s=1}^{k} w_s q\left(X_i\,|\mu_s, \Sigma_s a_i/b_i\right)},$$

As above, r_{ij} is the probability that the object X_i is obtained from the jth component of the mixture at the current approximation of the parameters w_j, μ_j, Σ_j.

II. Compute

$$a_i = \frac{\nu + d}{2}, \qquad b_i = \frac{\nu}{2} + \frac{1}{2}\sum_{j=1}^{k} r_{ij}\,(X_i - \mu_j)^T \Sigma_j^{-1}(X_i - \mu_j), \qquad c_i = b_i/a_i,$$

where d is the dimension of the feature space.

M-step. Compute a new approximation of parameters

$$w_j = \sum_{i=1}^{n} r_{ij} \Bigg/ \sum_{i,j=1}^{n,k} r_{ij}, \quad \mu_j = \sum_{i=1}^{n} r_{ij} c_i\, X_i \Bigg/ \sum_{i=1}^{n} r_{ij} c_i,$$

$$\Sigma_j = \frac{1}{n}\sum_{i=1}^{n} c_i\,(X_i - \mu_j)(X_i - \mu_j)^T.$$

Stopping criterion. Iterations of the method are performed up to convergence of

$$\mathcal{L}(w, \mu, \Sigma, r, a, b) = \sum_{i=1}^{n}\sum_{j=1}^{k} r_{ij}\left[\ln w_j - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln\det\Sigma_j - \right.$$
$$\left. -\frac{b_i}{2a_i}\left[\nu + (X_i - \mu_j)^T \Sigma_j^{-1}(X_i - \mu_j)\right] + \frac{\nu}{2}\ln\frac{\nu}{2} - \Gamma\left(\frac{\nu}{2}\right) + \left(\frac{\nu + d}{2} - 1\right)(\psi(b_i) - \ln a_i)\right] -$$
$$-\sum_{i=1}^{n}\sum_{j=1}^{k} r_{ij}\ln r_{ij} - \sum_{i=1}^{n}\left[b_i \ln a_i - \ln\Gamma(b_i) + (b_i - 1)\left(\psi(b_i) - \ln a_i\right) - b_i\right].$$

Using those iterations, the method approximates a local maximum of the logarithmic likelihood function, which makes it necessary to run the procedure several times from different initial points. For more information about method convergence see Section 4.

The hyperparameter ν can be chosen in a usual way by performing the algorithm for several values ν and choosing the one that maximizes $\mathcal{L}$.

4. FORMULAS DERIVATION FOR ESTIMATION THE PARAMETERS OF STUDENT MIXTURE

Let $X = (X_1, ..., X_n)$ be a sample of vectors from a mixture of Student distributions. The likelihood equals

$$L(w, \mu, \Sigma) = \prod_{i=1}^{n} \sum_{j=1}^{k} w_j p(X_i | \mu_j, \Sigma_j, \nu).$$

The family of distributions does not belong to the exponential class of distributions. Therefore, a maximization of likelihood is not straightforward. Let us use the representation of a Student random vector via a normal random vector and a gamma random vector. Next, we introduce hidden (i.e. unknown) values.

1. For each X_i, introduce the cluster number as $T_i = (T_{i1}, \dots, T_{ik}) \in \{0, 1\}^k$, with $\sum_{j=1}^{k} T_{ij} = 1$. The value $T_{ij} = 1$ if the object X_i is taken from the cluster j and $T_{ij} = 0$ otherwise. Denote $T = (T_1, \dots, T_n)$.
2. Also for each X_i, we introduce a random variable Y_i that has distribution $\Gamma(\nu/2, \nu/2)$ such that

$$X_i = \sum_{j=1}^{k} \left(\mu_j + \xi_{ij} \Big/ \sqrt{Y_i} \right) I\{T_{ij} = 1\},$$

where the random vector ξ_{Ij} has distribution $\mathcal{N}(0, \sigma_j)$ and is independent from Y_i. Denote $Y = (Y_1, \dots, Y_n)$.

Vector X_i has the normal distribution $\mathcal{N}(\mu_j, \Sigma_j/y)$ conditioned on $T_{ij} = 1$ and $Y_i = y$. The joint distribution of vectors (X, T, Y) has the density (the component T has discrete density)

$$p(x, t, y | w, \mu, \Sigma, \nu) = \prod_{i=1}^{n} \prod_{j=1}^{k} \left[w_j q(x_i | \mu_j, \Sigma_j / y_i) \gamma(y_i | \nu/2, \nu/2) \right]^{t_{ij}}.$$

Compute

$$\ln p(x, t, y | w, \mu, \Sigma, \nu) = \sum_{i=1}^{n} \sum_{j=1}^{k} t_{ij} \left[\ln w_j + \ln q(x_i | \mu_j, \Sigma_j / y_i) + \ln \gamma(y_i | \nu/2, \nu/2) \right] =$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{k} t_{ij} \left[\ln w_j - \frac{d}{2} \ln 2\pi - \frac{1}{2} \ln \det \Sigma_j + \frac{d}{2} \ln y_i - \frac{y_i}{2} (x_{ij} - \mu_j)^T \Sigma_j^{-1} (x_{ij} - \mu_j) + \right.$$

$$\left. + \frac{\nu}{2} \ln \frac{\nu}{2} - \ln \Gamma(\nu/2) + \left(\frac{\nu}{2} - 1 \right) \ln y_i - \frac{\nu}{2} y_i \right].$$

The estimation of the mixture parameters is performed by solving the problem of maximizing the model likelihood function using EM algorithm. At the E-step the distributions of T and Y are estimated iteratively using variational Bayesian inference. Independence of posteriori distributions (conditioned on X) is assumed. These results are in the following iterative scheme.

1. Selection a random initial values w_j, μ_j, Σ_j. Vectors μ_j can be obtained from normal distribution $\mathcal{N}(0, I_d)$, vectors w_j — from the uniform distribution on the simplex (Dirichlet distribution). The matrices Σ_j can be generated using the Wishart distribution [13].
2. E-step:
   (a) Select a random initial value for Y.
   (b) Having the current distribution of Y, approximate the distribution of T.
   (c) Having the current distribution of T, approximate the distribution of Y.
   (d) Repeat steps b-c until convergence of L.
3. M-step.
4. Repeat E and M steps until convergence of L.

4.1. E-step, Internal step I

At this step, the distribution of T conditioned on X is approximated using the relation $\ln r(t) \propto \mathsf{E}_\gamma \ln p(X, t, Y | w, \mu, \Sigma, \nu)$, where mathematical expectation E_γ is computed under the condition that the current distribution of Y was computed on the previous iteration of step II.

Remark. Everywhere below symbol $\propto$ means equality up to a multiplicative constant for probabilities and equality up to an additive constant for logarithms of probabilities.

Each T_i has a discrete distribution with values in a set of binary vectors that have exactly one unit. For this distribution, density logarithm is $\sum_{j=1}^k t_{ij} \ln r_{ij}$, where $r_{ij} = \mathsf{P}(T_{ij} = 1)$ and $\sum_{j=1}^k r_{ij} = 1$.

$$\ln r(t) \propto \mathsf{E}_\gamma \ln p(X, t, Y | w, \mu, \Sigma, \nu) \propto$$

$$\propto \sum_{i=1}^n \sum_{j=1}^k t_{ij} \left[\ln w_j - \frac{1}{2} \ln \det \Sigma_j - \frac{\mathsf{E}_\gamma Y_i}{2} (X_i - \mu_j)^T \Sigma_j^{-1} (X_i - \mu_j) \right] \propto$$

$$\propto \sum_{i=1}^n \sum_{j=1}^k t_{ij} \left[\ln w_j - \frac{1}{2} \ln \det \left(\Sigma_j / \mathsf{E}_\gamma Y_i \right) - \frac{1}{2} (X_i - \mu_j)^T \left(\Sigma_j / \mathsf{E}_\gamma Y_i \right)^{-1} (X_i - \mu_j) \right] \propto$$

$$\propto \sum_{i=1}^n \sum_{j=1}^k t_{ij} \left[\ln w_j + \ln q \left(X_i \left| \mu_j, \Sigma_j / \mathsf{E}_\gamma Y_i \right. \right) \right].$$

The resulting expression implies that the optimal approximation of conditional distribution of T (provided X) is such that the values $T_1, \ldots, T_n$ are independent and $r_{ij} \propto w_j q \left(X_i \left| \mu_j, \Sigma_j / \mathsf{E}_\gamma Y_i \right. \right)$. From the condition $\sum_{j=1}^k r_{ij} = 1$ we get

$$r_{ij} = \frac{w_j q \left(X_i \left| \mu_j, \Sigma_j / \mathsf{E}_\gamma Y_i \right. \right)}{\sum\limits_{s=1}^k w_s q \left(X_i \left| \mu_s, \Sigma_s / \mathsf{E}_\gamma Y_i \right. \right)}.$$

The expectation $\mathsf{E}_\gamma Y_i$ is taken of the current approximation of Y_i.

4.2. E-step, Internal step II

At this step, the distribution of Y conditioned on X is computed: $\ln \gamma(y) \propto \mathsf{E}_r \ln p(X, T, y|w, \mu, \Sigma, \nu)$, where the current distribution of T is assumed.

Let us write this expression up to a constant that does not depend on Y, given that $\sum_{j=1}^{k} r_{ij} = 1$

$$\ln \gamma(Y) = \mathsf{E}_r \ln p(X, T, y|w, \mu, \Sigma, \nu) \propto$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{k} \mathsf{E}_r T_{ij} \left[\frac{d}{2} \ln y_i - \frac{y_i}{2} (X_i - \mu_j)^T \Sigma_j (X_i - \mu_j) + \left(\frac{\nu}{2} - 1 \right) \ln y_i - \frac{\nu}{2} y_i \right] \propto$$

$$\propto \sum_{i=1}^{n} \sum_{j=1}^{k} r_{ij} \left[\left(\frac{\nu + d}{2} - 1 \right) \ln y_i - \left(\frac{\nu}{2} + \frac{1}{2} (X_i - \mu_j)^T \Sigma_j^{-1} (X_i - \mu_j) \right) y_i \right] =$$

$$\propto \sum_{i=1}^{n} \left[\left(\frac{\nu + d}{2} - 1 \right) \ln y_i - \left(\frac{\nu}{2} + \frac{1}{2} \sum_{j=1}^{k} r_{ij} (X_i - \mu_j)^T \Sigma_j^{-1} (X_i - \mu_j) \right) y_i \right].$$

Thus, we obtain the gamma distribution with parameters

$$a_i = \frac{\nu}{2} + \frac{1}{2} \sum_{j=1}^{k} r_{ij} \, (X_i - \mu_j)^T \Sigma_j^{-1} (X_i - \mu_j), \quad b_i = \frac{\nu + d}{2}.$$

The values of mathematical expectations are updated using relations $\mathsf{E}_\gamma Y_i = \frac{b_i}{a_i}$ and $\mathsf{E}_\gamma \ln Y_i = \psi(b_i) - \ln a_i$, which are stated in Section 2. For brevity, denote $c_i = b_i / a_i$.

4.3. M-step

At this step, the values of the mixture parameters are updated by maximizing $\mathsf{E}_{r,\gamma} \ln p(X, T, Y|w, \mu, \Sigma, \nu)$, where the current distributions of T and Y are assumed.

We leave only those summands that depend on w_j, μ_j, Σ_j

$$F_{X,\nu}(w, \mu, \Sigma) = \mathsf{E}_{r,\gamma} \ln p(X, T, Y|w, \mu, \Sigma, \nu) \propto$$

$$\propto \sum_{i=1}^{n} \sum_{j=1}^{k} \mathsf{E}_r T_{ij} \left[\ln w_j - \frac{1}{2} \ln \det \Sigma_j - \frac{1}{2} (X_i - \mu_j)^T \Sigma_j^{-1} (X_i - \mu_j) \mathsf{E}_\gamma Y_i \right] =$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{k} r_{ij} \left[\ln w_j - \frac{1}{2} \ln \det \Sigma_j - \frac{c_i}{2} (X_i - \mu_j)^T \Sigma_j^{-1} (X_i - \mu_j) \right].$$

Maximization by w_j of $F_{X,\nu}(w, \mu, \Sigma)$ is equivalent to finding a solution of the problem

$$\begin{cases} \sum\limits_{i=1}^{n} \sum\limits_{j=1}^{k} r_{ij} \ln w_j \longrightarrow \max\limits_{w}, \\ \sum\limits_{j=1}^{k} w_j = 1, \\ w_j \geqslant 0. \end{cases}$$

Let us forget about the restrictions of the inequality type for a while, make a Lagrange function, and find its maximum

$$L = \sum_{i=1}^{n} \sum_{j=1}^{k} r_{ij} \ln w_j - \lambda \left(\sum_{j=1}^{k} w_j - 1 \right),$$

$$\frac{\partial L}{\partial w_j} = \frac{1}{w_j} \sum_{i=1}^{n} r_{ij} - \lambda = 0,$$

$$w_j = \frac{1}{\lambda} \sum_{i=1}^{n} r_{ij}.$$

From the condition $\sum_{j=1}^{k} w_j = 1$ we get

$$w_j = \sum_{i=1}^{n} r_{ij} \Bigg/ \sum_{i=1}^{n} \sum_{s=1}^{k} r_{is}.$$

Note that the conditions $w_j \geqslant 0$ are met. The problem is solved due to the convexity.

To maximize by μ_j, we equate the derivative of $F_{X,\nu}(w, \mu, \Sigma)$ with respect to vector μ_j to zero

$$\frac{\partial F_{X,\nu}(w, \mu, \Sigma)}{\partial \mu_j} = \sum_{i=1}^{n} r_{ij} c_i \, \Sigma_j^{-1} (X_i - \mu_j) = 0,$$

$$\sum_{i=1}^{n} r_{ij} c_i X_i = \sum_{i=1}^{n} r_{ij} c_i \mu_j,$$

$$\mu_j = \sum_{i=1}^{n} r_{ij} c_i \, X_i \Bigg/ \sum_{i=1}^{n} r_{ij} c_i.$$

To maximize by Σ_j, we equate the derivative of $F_{X,\nu}(w, \mu, \Sigma)$ with respect to matrix Σ_j to zero. Note that

$$(X_i - \mu_j)^T \Sigma_j^{-1} (X_i - \mu_j) = \mathrm{tr} \left((X_i - \mu_j)^T \Sigma_j^{-1} (X_i - \mu_j) \right) = \mathrm{tr} \left(\Sigma_j^{-1} (X_i - \mu_j)(X_i - \mu_j)^T \right).$$

Using this transformation, as well as matrix derivative formulas for square matrices [14]

$$\frac{\partial}{\partial X} \det X = \det X \cdot X^{-T},$$

$$\frac{\partial}{\partial X} \mathrm{tr} \left(X^{-1} A \right) = - \left(X^{-1} A X^{-1} \right)^T,$$

get

$$\frac{\partial F_{X,\nu}(w, \mu, \Sigma)}{\partial \Sigma_j} = -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{k} r_{ij} \left[\frac{\partial}{\partial \Sigma_j} \ln \det \Sigma_j + c_i \frac{\partial}{\partial \Sigma_j} (X_i - \mu_j)^T \Sigma_j^{-1} (X_i - \mu_j) \right] =$$

$$= -\frac{1}{2} \sum_{i=1}^{n} r_{ij} \left(\Sigma_j^{-1} - c_i \Sigma_j^{-1} (X_i - \mu_j)(X_i - \mu_j)^T \Sigma_j^{-1} \right) = 0.$$

Multiplying both parts of the last equation by Σ_j we get

$$\sum_{i=1}^{n} r_{ij}\Sigma_j = \sum_{i=1}^{n} r_{ij}c_i\,(X_i-\mu_j)(X_i-\mu_j)^T,$$

$$\Sigma_j = \sum_{i=1}^{n} r_{ij}c_i\,(X_i-\mu_j)(X_i-\mu_j)^T \Bigg/ \sum_{i=1}^{n} r_{ij}.$$

4.4. Variational lower bound and convergence of the method

Iterations of the EM algorithm continue until the convergence of the variational lower bound [6]

$$\mathcal{L}(w,\mu,\Sigma,r,a,b) = \mathsf{E}_{r,\gamma}\ln p(X,T,Y|w,\mu,\Sigma,\nu) - \mathsf{E}_r \ln r(T) - \mathsf{E}_\gamma \ln\gamma(Y).$$

Let us write each summand separately

$$\mathsf{E}_{r,\gamma}\ln p(X,T,Y|w,\mu,\Sigma,\nu) =$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{k}\mathsf{E}_r T_{ij}\left[\ln w_j - \frac{d}{2}\ln 2\pi + \frac{d}{2}\mathsf{E}_\gamma \ln Y_i - \frac{1}{2}\ln\det\Sigma_j - \right.$$

$$\left. -\frac{1}{2}(X_i-\mu_j)^T\Sigma_j^{-1}(X_i-\mu_j)\mathsf{E}_\gamma Y_i + \frac{\nu}{2}\ln\frac{\nu}{2} - \Gamma\left(\frac{\nu}{2}\right) + \left(\frac{\nu}{2}-1\right)\mathsf{E}_\gamma \ln Y_i - \frac{\nu}{2}\mathsf{E}_\gamma Y_i\right] =$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{k} r_{ij}\left[\ln w_j - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln\det\Sigma_j - \right.$$

$$\left. -\frac{b_i}{2a_i}\left[\nu + (X_i-\mu_j)^T\Sigma_j^{-1}(X_i-\mu_j)\right] + \frac{\nu}{2}\ln\frac{\nu}{2} - \Gamma\left(\frac{\nu}{2}\right) + \left(\frac{\nu+d}{2}-1\right)(\psi(b_i)-\ln a_i)\right],$$

$$\mathsf{E}_r \ln r(T) = \sum_{i=1}^{n}\sum_{j=1}^{k} r_{ij}\ln r_{ij},$$

$$\mathsf{E}_\gamma \ln\gamma(Y) = \sum_{i=1}^{n}\mathsf{E}_\gamma\ln\gamma(Y_i) = \sum_{i=1}^{n}\left[b_i\ln a_i - \ln\Gamma(b_i) + (b_i-1)\mathsf{E}_\gamma\ln Y_i - a_i\mathsf{E}_\gamma Y_i\right] =$$

$$= \sum_{i=1}^{n}\left[b_i\ln a_i - \ln\Gamma(b_i) + (b_i-1)\left(\psi(b_i)-\ln a_i\right) - b_i\right].$$

Let us study the convergence of the method. The standard EM algorithm is based on the formula (see [6])

$$\log L(w,\mu,\Sigma) = \mathcal{L}(w,\mu,\Sigma,q) + KL(q,p_{w,\mu,\Sigma}(t,y|x)),$$

where q is a distribution of the vector (T,Y), KL is Kullback-Leibler divergence. Here, for the variational lower bound, we use the notation $\mathcal{L}(w,\mu,\Sigma,q)$ since, in general, q does not depend on parameters. In order to maximize the likelihood function, the variational lower bound is maximized at each step, at the E-step by q, at the M-step by the parameters w,μ,Σ.

Thus, its values do not decrease. Maximization at the E-step is equivalent to minimizing the divergence by q, which is equivalent to choosing $q = p_{w,\mu,\Sigma}(t, y|x)$, where the divergence is exactly zero. Thus, the variational lower bound after the E-step is equal to the logarithm of the likelihood function $\log L(w, \mu, \Sigma)$.Therefore the likelihood function also does not decrease and converges to the local maximum.

For the modification of the EM algorithm under consideration, the formula takes the form

$$\log L(w, \mu, \Sigma) = \mathcal{L}(w, \mu, \Sigma, r, a, b) + KL(r \times \gamma, p_{w,\mu,\Sigma}(t, y|x)).$$

Variational output at the E-step minimizes divergence. However in the class of distributions under consideration, where the distributions of T and Y are conditionally independent, the zero divergence may not be achieved. Thus, the convergence of L to the local maximum cannot be guaranteed. In the practical problems we are considering, it is natural to assume that the distributions T and Y are close to be conditionally independent, that should provide a good approximation of the local maximum using the resulting estimation.

5. APPLICATIONS OF PROBABILISTIC MODEL

The model of a mixture distribution allows solving various machine learning problems listed in the introduction and obtaining consistent results. Let us describe each of them in more detail, assuming a mixture model with density

$$p(x) = \sum_{j=1}^{k} w_j p(x|\theta_j),$$

where θ_j is the parameter of the jth component distribution (for example, the mean vector and the covariance matrix).

5.1. Clustering

The components of the mixture can be considered as overlapping clusters. Each object $x \in \mathbb{R}^d$ can be assigned to one of the clusters with some probability. According to Claim 3.2, the conditional probability that an object x corresponds to a cluster j is

$$p_j(x) = \frac{w_j p(x|\theta_j)}{\sum\limits_{s=1}^{k} w_s p(x|\theta_s)}.$$

If $\widehat{w}_j, \widehat{\theta}_j$ are estimations of the parameters w_j, θ_j respectively, then we can estimate $\widehat{p}_j(x) = \widehat{w}_j p(x|\widehat{\theta}_j) \bigg/ \sum\limits_{s=1}^{k} \widehat{w}_s p(x|\widehat{\theta}_s)$. These probability estimation can be considered as the confidence level of the method when assigning an object x to a cluster j, which is sufficient to solve the following problem. Performing "hard" clustering, where an object x must be strictly attributed to one of the clusters, the cluster with the maximum probability $\widehat{p}_j(x)$ is selected

$$j_* = \arg\max_j \widehat{p}_j(x) = \arg\max_j \widehat{w}_j p(x|\widehat{\theta}_j).$$

5.2. Anomalies

Object $x \in \mathbb{R}^d$ is considered abnormal if density value $p(x)$ is less than some threshold value q. The value q is chosen as density value $p(x)$ so that the probability of getting an object with a density value not exceeding q is exactly 0.05. In other words, the value q is the solution of the equation

$$\int\limits_{\mathbb{R}^d} p(x)I\{p(x) \leqslant q\}dx = 0.05.$$

The described procedure for determining anomalous objects is a special case of the procedure for checking statistical hypotheses. In that case, the hypothesis that object x is typical is tested against the alternative one about object abnormality. If the density at the point x is less than the threshold value q, the hypothesis of object typicity is rejected in favor of an alternative one at the significance level 0.05. This rule is a criterion for testing a hypothesis.

A probability of getting objects with a density at least $p(x)$ can be considered as the level of typicity of object x. This value is analogous to p-value and is computed using the integral

$$\int\limits_{\mathbb{R}^d} p(y)I\{p(y) \leqslant p(x)\}dy.$$

5.3. *Missing data*

Both methods discussed above work only if all values of components $x \in \mathbb{R}^d$ are known, i.e. there are no missing values. Otherwise, the density of the object x can be estimated as the density integral over a subspace of omitted values. Formally, let x_k be a vector of known object values, and x_u are all other object values that are omitted. Then

$$\int\limits_{\mathbb{R}^{d_u}} p(x)dx_u,$$

where d_u is a dimension of vector x_u.

For a normal or Student distribution mixture, the density is equal to the mixture of marginal distributions obtained in Claims 2.2.

5.4. *Conditional distribution and probabilistic regression on features*

Let object $x \in \mathbb{R}^d$ be recognized as anomalous. Let us select elements of the vector x to trust in and evaluate others through them. Without loss of generality we assume that $x^T = (x_a^T, x_b^T)$ and the values of x_b are trusted. In addition, due to missing data, some values of x_a may not be known.

According to Claim 3.2, vector x_a under the condition of values x_b has density

$$\widetilde{p}(x) = \sum_{j=1}^{k} \widetilde{w}_j p(x|\widetilde{\theta}_j),$$

where $\widetilde{\theta}_j$ is the parameter of the mixture distribution jth component given x_b. In the case of the normal mixture, the parameters $\widetilde{\theta}_j$ for each cluster are computed in accordance with relations from Claim 2.3. In the case of the Student mixture, the parameters are computed in accordance with relations from the Theorem 2.1.

Replacing the parameters with their estimations, we get an estimation of the conditional distribution of vector x_a, which is sufficiently informative for making various conclusions about the vector x_a in practice. In particular, it can be used to compute

- Expectation value $\mathsf{E}(x_a \mid x_b)$ according to Claim 3.1. This estimation solves the problem of regression of x_b features to x_a features. Note that the regression problem can be solved for different features x_a and x_b using only one model.

- Variance estimation $\mathrm{Var}(x_a \mid x_b)$ according to Claim 3.2.
- Evaluation of conditional cluster distributions for all objects that have X_b attributes fixed and equal to x_b.
- Set of the highest density, i.e. set of values x_a, for which the density of the conditional distribution is greater than for the other values. Such a set is analogous to the confidence domain of the minimum volume. In the one-dimensional case, this set is an interval or set of intervals.

6. MODELING PVT PROPERTIES OF RESERVOIR FLUIDS USING A PROBABILISTIC MODEL

6.1. Data description

There are several of methods for evaluating the representativeness of reservoir fluid samples [15], such as checking the tightness of sampling chambers, comparing the oil saturation pressure with the separation pressure at the separation temperature, etc.; the Hoffman-Kramp-Hockot method, based on the correlation of equilibrium constants; determining the representativeness of samples by the criterion of contamination with process fluids used in drilling, perforation and development of wells. In conditions where only raw data is available the above methods cannot be applied. Thus, it is reasonable to develop algorithms for detecting potentially incorrect values from raw data.

For practical application of research on PVT-properties of fluids, a database containing the results of studies of more than 3,200 samples of reservoir fluids was analyzed. Among the considered features, there are the following values: reservoir pressure, reservoir temperature, surface gas density, surface oil density, gas content, saturation pressure, reservoir oil density, oil volume coefficient, and reservoir oil viscosity.

The problem of predicting PVT properties using machine learning methods was previously considered in a very limited version. For example, in [16], [17], [18] the prediction of saturation pressure through other properties using artificial neural networks (ANN) is considered. In [18], [19] predictions of the oil volume coefficient are made in the same way. In [20] SVM regression is used to predict the features mentioned above.

This paper offers a fundamentally different approach to solving these problems. It is based on the introduction of a probabilistic model in the space of reservoir fluid properties, where it is assumed that the characteristic description of the sample is obtained independently of all other samples from a certain probability distribution.

6.2. Normal mixture model

First we use a mixture of normal distributions of four components to describe data. Since the iterative procedure of the EM algorithm converges to the point of the local maximum of the logarithmic likelihood function, the method was run several times from random initial values of parameters. The final score is obtained in the iteration with the highest value of the variational lower score.

The result of parameter estimation is shown on Figure 6.1. The diagonal shows the densities of features for the estimated mixture. Each non-diagonal cell on Figure corresponds to the projection of the feature space on all possible coordinate planes. Lines of the density level of the resulting mixture of distributions are drawn above the diagonal. Below the diagonal, ellipses of different colors indicate the relative location of clusters. Notice that a gray semi-transparent cluster covers other two clusters, due to the presence of noise objects in the data. The normal distribution has light tails, so the estimations of its parameters are not stable to the presence of noise objects in the data. When constructing a model of a mixture of normal distributions on PVT data, the EM algorithm tries to describe the main part of the data using three clusters, and the other less typical objects using the fourth one.

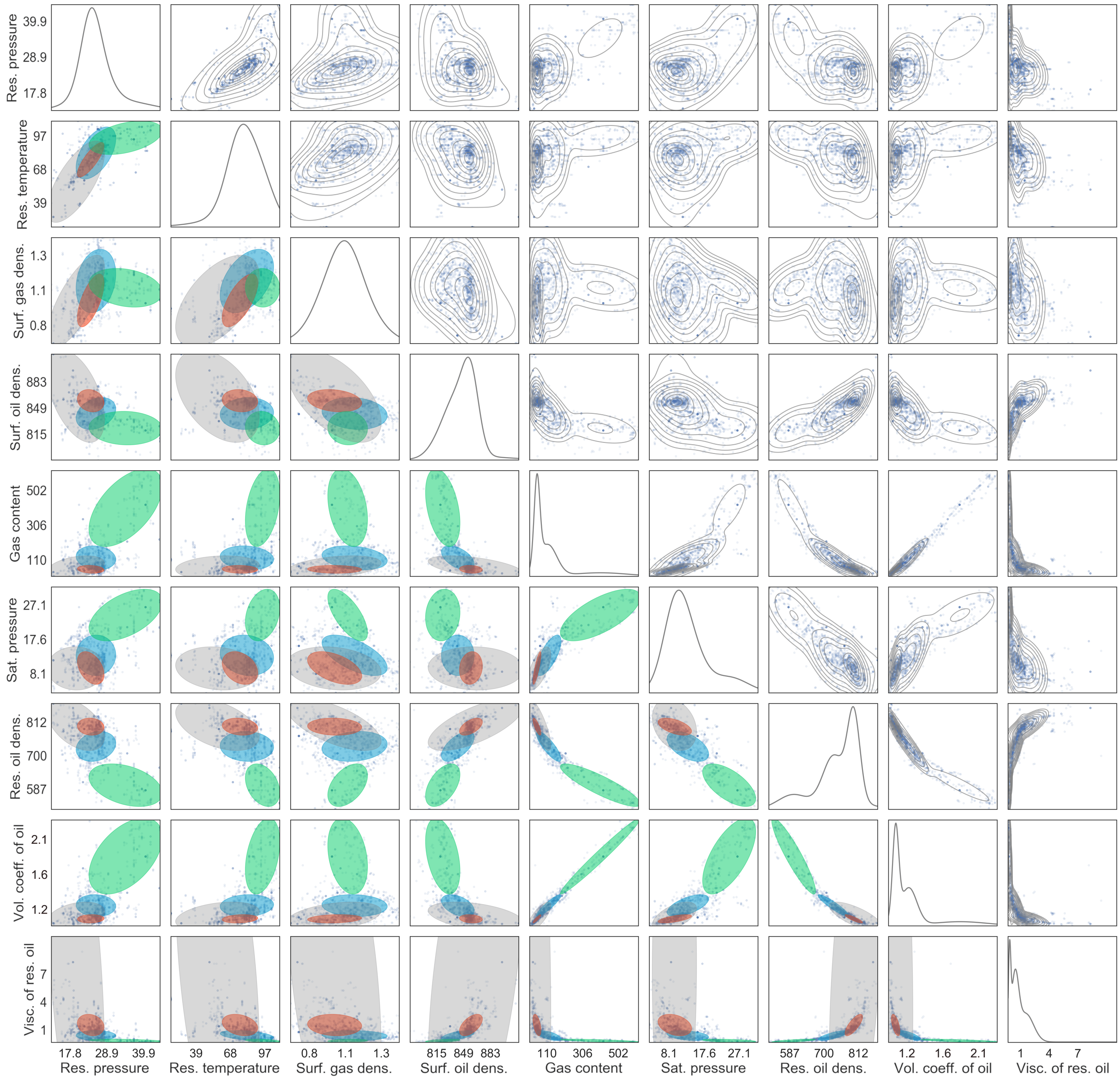

Fig. 6.1. Data visualization and results of application of the normal mixture model. The plots above the diagonal show the density levels of joint distribution of features projected on subspace of every two features. The diagonal plots show the density of every feature. The plots under the diagonal show the clusterization based on the application of the normal mixture model. Every ellipse corresponds to one of four clusters/components of the mixture

6.3. Student mixture model

To eliminate the above disadvantages, the Student distribution mixture model is applied. The number of degrees of freedom nu needs some expertise. On Figure 6.2 the dependence of the variational lower bound on the iteration for the best result among several runs from different initial approximations is shown. The result of evaluating parameters for four clusters is shown on Figure 6.3. The blue and red clusters obtained using the Student multidimensional distribution mixture model correspond to the same clusters in the multi-dimensional normal distribution mixture model. The green cluster is split into two. Note the clusters have no noise. This result is a consequence of the stability of the Student distribution to emissions.

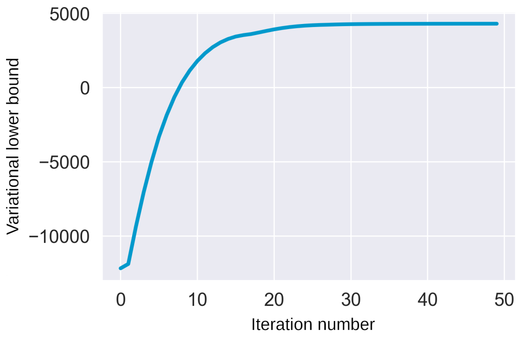

Fig. 6.2. The dependence of the variational lower bound on iteration number

Table 6.1. Centers of clusters obtained using the Student mixture model

	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Sample rate	28.3%	27.7%	29.0%	15.0%
Reservoir pressure, MPa	23.86	25.36	25.63	36.04
Reservoir temperature, ^{o}C	75.33	80.47	83.91	96.12
Surface gas density, kg / m3	1.01	1.07	1.17	1.11
Surface oil density, kg / m3	859.74	855.61	836.75	823.55
Gas content, m3 / t	49.20	69.58	138.74	425.62
Saturation pressure, MPa	8.46	11.25	13.61	24.49
Oil reservoir density, kg / m3	808.33	772.59	716.74	598.29
Volume coefficient. oil, m3 / m3	1.12	1.20	1.36	2.04
Reservoir oil viscosity, MPa * s	2.29	1.24	0.67	0.27
Oil type	Heavy & Medium	Medium	Light	Extra light

Cluster centers are given in Table 6.1. During training, clusters are defined by the model up to permutation so the order of clusters is determined by experts. One can notice that for most clusters their centers are strictly ordered by most attributes. The type of oil corresponding to each cluster is also determined by experts.

7. MODEL RESEARCH

The behavior of the model based on a mixture of four components of Student distributions is tested on artificial data, the results of experiments are given below. In addition, the quality of model predictions is tested by test data that is not involved in the training set.

7.1. Artificial experiments

In our research, the results of the model application to artificial samples are analyzed. Some sample features are fixed in three different versions of the experiments. Their values are given in Table 7.2. In each of three variants of experiments, the gas content values are varied between 0 and 800 m3/t. For each value, the probabilistic density of the sample (i.e., the probability of belonging to each of the three clusters), the expected values of saturation pressure, reservoir oil density, and oil volume coefficient are computed using a mixture of Student distributions.

The sample density depending on the gas content for the three variants under study is given on Figure 7.4. The black dotted line shows the anomaly threshold. If the density is below the threshold, the sample is considered abnormal. The graph shows that in the first variant typical samples correspond to gas content values from 90 to 190 m3/t, in the second variant from 0 to 110 m3/t, in the third from 0 to 90 m3/t.

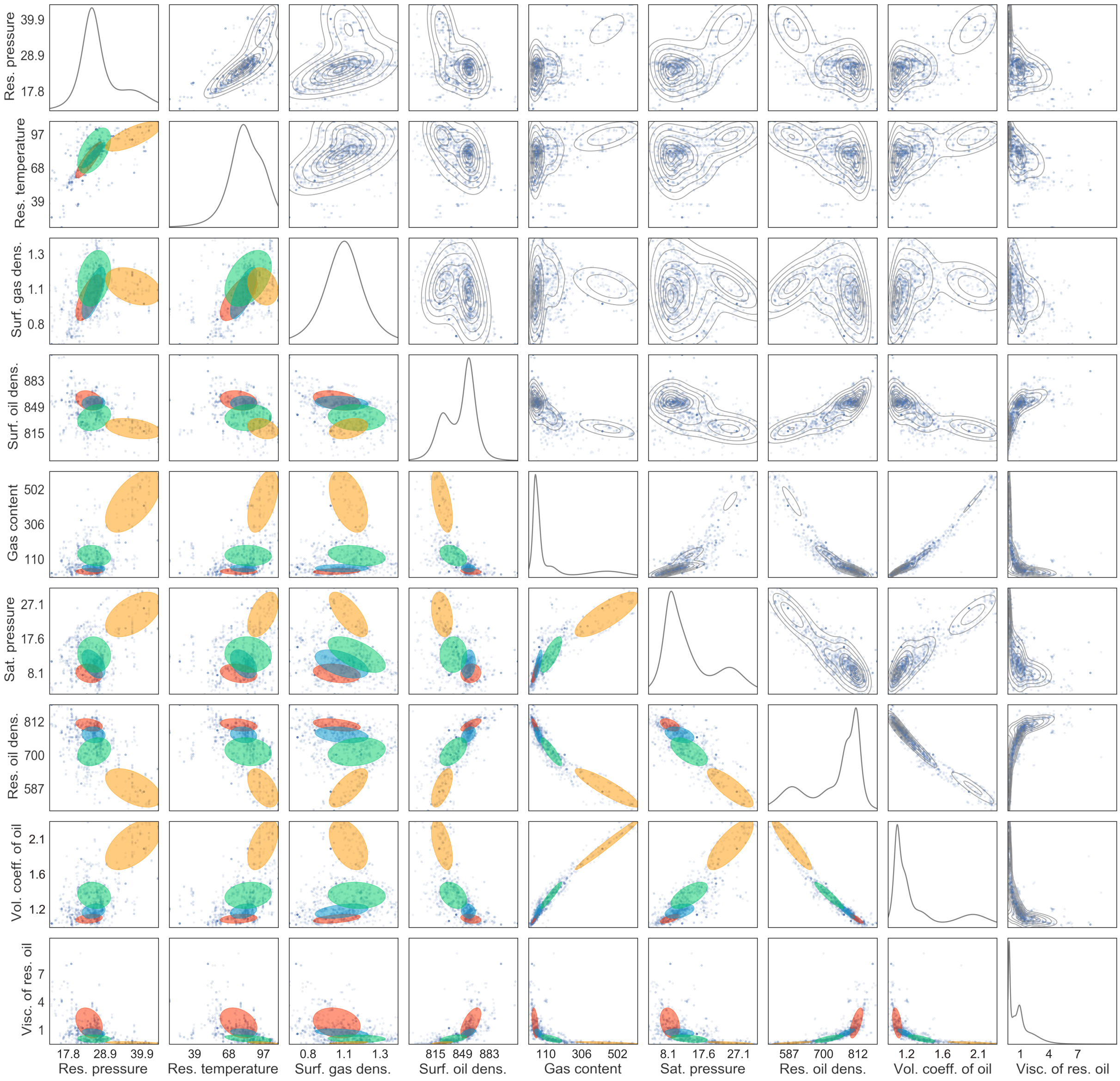

Fig. 6.3. Data visualization and results of application of the Student mixture model. The plots above the diagonal show the density levels of joint distribution of features projected on subspace of every two features. The diagonal plots show the density of every feature. The plots under the diagonal show the clusterization based on the application of the Student mixture model. Every ellipse corresponds to one of four clusters/components of the mixture

The figure 7.5 shows the probability estimations that the sample corresponds to each of four clusters for three variants of experiments. For example, considering the first variant of experiments, we can conclude the sample belongs to a blue or green cluster depending on the value of the gas content .

Figure 7.6 contains the graphs of saturation pressure predictions, reservoir density of oil, and volume coefficient of oil in the three above variants depending on the gas content. Orange dots corresponds the samples used in train data. The shaded area denote the predictive interval. The less typical the sample, the greater the uncertainty of the model and the confident interval is wider. Note the trajectories of the predicted values are smooth that follows from the model construction. It is also worth noting that the observed shift in predictions relative

Table 7.2. Fixed values of features in model experiments

Feature	Variant 1	Variant 2	Variant 3
Reservoir pressure, MPa	35	25	18
Reservoir temperature, ^{o}C	100	75	50
Surface gas density, kg / m3	1.2	0.95	0.68
Surface oil density, kg / m3	800	845	880
Reservoir oil viscosity, MPa * s	0.5	2	5
Gas content, m3 / t	0 . . . 800	0 . . . 800	0 . . . 800
Saturation pressure, MPa	?	?	?
Oil reservoir density, kg / m3	?	?	?
Volume coefficient. oil, m3 / m3	?	?	?

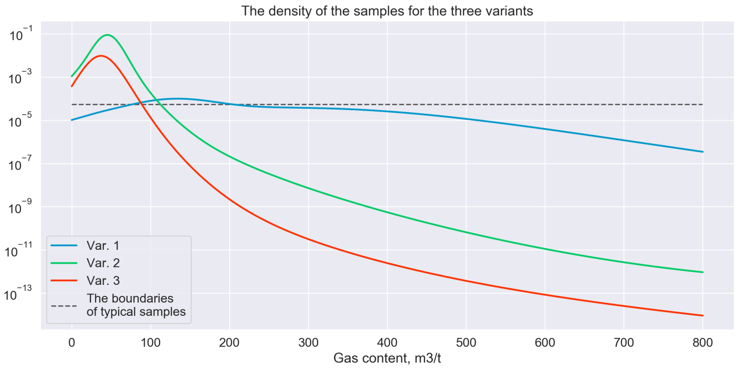

Fig. 7.4. Sample density depending on the gas content for three variants of experiments

to the total mass of points is due to the fact that each variant of experiments has fixed features not presented on the charts. If there weren't fixed features the predictions would look like averages.

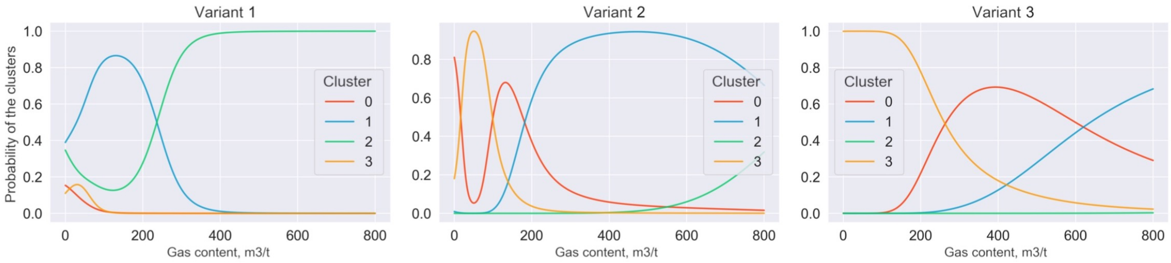

Fig. 7.5. Probability estimation of the sample belonging to each of four clusters depending on the gas content

7.2. *Predictions quality*

The developed model calculates the expected values for unknown feature values based on the entered sample. The degree of deviation of the estimated values from the true values is estimated using quality metrics. Let x_i be a true value of the feature, and $\widehat{x}_i$ be a predicted value under the assumption that x_i is unknown. Then the relative error of the prediction x_i is $e_i = (x_i - \widehat{x}_i)/x_i$. It measures the deviation of the predicted value from the true value in relation to the actual value of the true value. Prediction errors are calculated for a test set containing about 650 samples that are not involved in building the model. When predicting each value of the model passed all the known values of the samples in addition to predicted. The prediction quality is calculated as the average absolute and standard error in percentages,

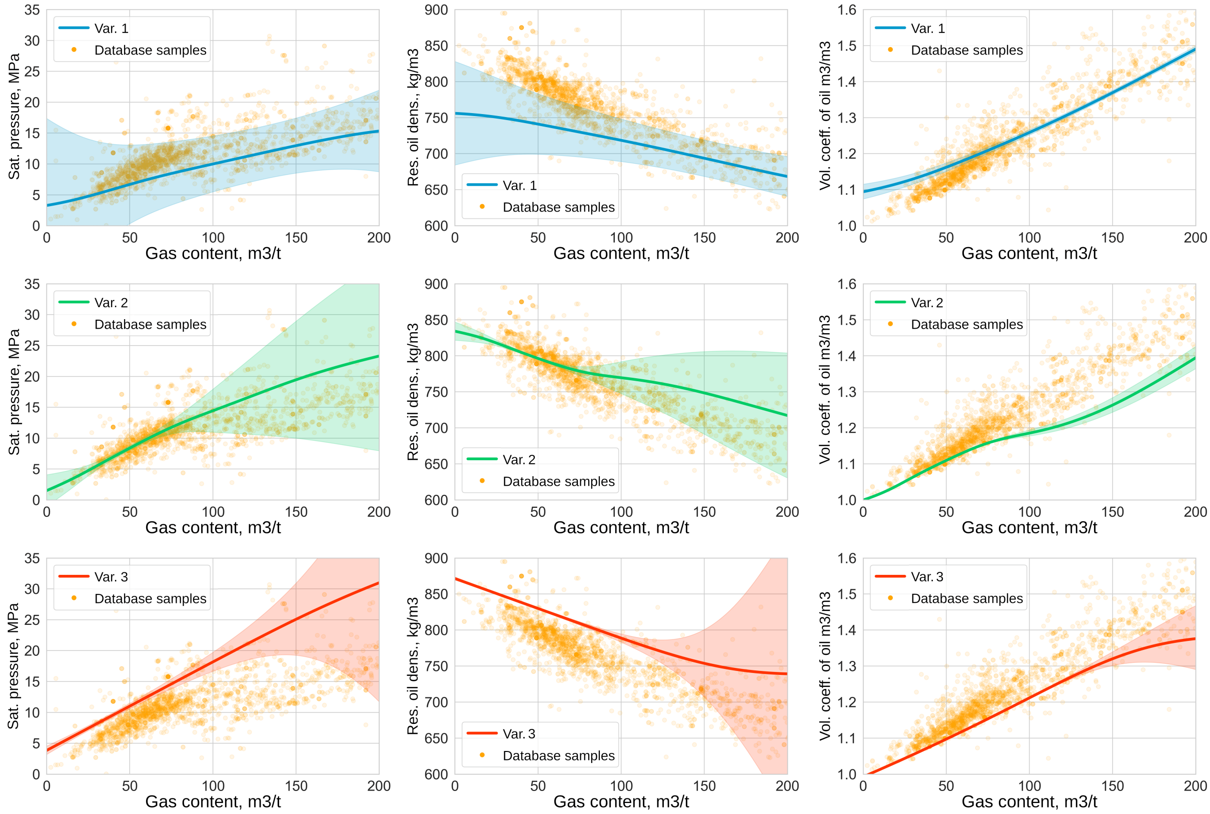

Fig. 7.6. Predictions of three features depending on gas content. Shaded area denotes confidence interval

which are defined as

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}|e_i|, \qquad RMSPE = 100\%\sqrt{\frac{1}{n}\sum_{i=1}^{n}e_i^2}.$$

For metrics evaluation, 2.5% of the highest and 2.5% of the lowest values are excluded from the set of numbers e_i due to outliers and incorrect values in the source data.

For comparison, 9 regression models were built for each feature for the following machine learning methods: gradient boosting (XGBoost[†], LGBM[‡], CatBoost[§]), as well as sklearn implementations[¶] of random forest (RF), SVM regression, neural network (ANN). Each such model is trained to predict one of the features, considering all the other features as a feature description. Optimal hyperparameters for each model are selected using cross-validation on the training set. Tables 7.3, 7.4 contain the values of the MAPE and RMSPE metrics for the test dataset. One can notice in 5 out of 9 cases the model prediction of the Student mixture is more accurate than all the other models, and in the other cases it is not far behind them. Moreover, experts usually trust reservoir pressure and temperature, which are more accurately predicted by gradient boosting, and therefore their prediction is not very significant in this task.

[†] https://xgboost.ai/
[‡] https://lightgbm.readthedocs.io/
[§] https://catboost.ai/
[¶] https://scikit-learn.org/

Table 7.3. Comparison of prediction quality based on the MAPE metric

	t-Mix	XGBoost	LGBM	CatBoost	RF	SVM	ANN
Reservoir pressure, MPa	8.13	6.70	6.56	**6.53**	8.31	6.77	9.63
Reservoir temperature, ^{o}C	5.96	**5.50**	5.67	5.75	5.78	6.58	6.25
Surface gas density, kg / m3	**3.42**	6.21	6.20	7.08	6.23	4.30	7.38
Surface oil density, kg / m3	**0.54**	0.87	0.95	0.88	0.87	0.67	1.02
Gas content, m3 / t	**3.92**	6.51	6.85	7.83	7.33	7.28	9.03
Saturation pressure, MPa	11.69	**9.15**	9.25	9.34	10.33	9.23	18.07
Oil reservoir density, kg / m3	**0.71**	1.35	1.27	2.04	1.58	1.77	1.85
Volume coefficient. oil, m3 / m3	**0.68**	1.43	1.57	2.70	2.13	2.23	2.62
Reservoir oil viscosity, MPa * s	22.32	30.60	26.86	27.30	57.60	123.73	**9.66**

Table 7.4. Comparison of prediction quality based on the RMSPE metric

	t-Mix	XGBoost	LGBM	CatBoost	RF	SVM	ANN
Reservoir pressure, MPa	10.15	9.05	8.96	**8.56**	12.53	8.86	15.85
Reservoir temperature, ^{o}C	7.50	**7.15**	7.26	**7.15**	7.61	8.23	7.93
Surface gas density, kg / m3	**4.75**	7.80	7.79	8.72	7.95	5.44	9.38
Surface oil density, kg / m3	**0.88**	1.11	1.21	1.07	1.12	1.01	1.58
Gas content, m3 / t	**5.64**	8.57	8.57	9.92	11.12	10.15	11.21
Saturation pressure, MPa	16.52	12.79	13.27	12.57	14.89	**12.71**	21.63
Oil reservoir density, kg / m3	**1.47**	2.08	1.78	2.98	2.36	3.95	2.37
Volume coefficient. oil, m3 / m3	**1.49**	2.12	2.36	3.74	3.34	4.03	3.29
Reservoir oil viscosity, MPa * s	30.22	44.11	36.69	36.82	91.35	398.20	**12.68**

8. CONCLUSIONS

The probabilistic model of a mixture of multidimensional Student distributions proposed in the paper for describing the properties of PVT samples has a wide range of practical applications, including checking the samples for abnormality, dividing the samples into four clusters, computing recommended values for missing values in the sample, and in the case of sample abnormality – for all features not selected as trusted. Experiments have shown that the recommended values obtained do not contradict the physical properties of PVT samples, in particular, they have smoothness in terms of arguments.

The division of samples into four clusters corresponding to components of the multidimensional Student mixture is empirically justified by comparison with a mixture of multidimensional normal distributions. In the latter case, the quality is unsatisfactory due to the presence of noise objects that the model adjusts to a separate cluster. The Student distribution has heavier tails, so it adjusts less to emissions.

The probabilistic model has significant advantages over other models since it can be used to solve several problems at once and obtain consistent results. Considering only the regression problem, the probabilistic model, in contrast to the traditional approach, allows getting predictions of set of features for other set of features and vise verse without repeated model training. Furthermore, in the conducted experiments, the quality of the obtained predictions is also superior to other models.

REFERENCES

1. Lagutin, M. B. (2009) *Naglyadnaya matematicheskaya statistika* [Visual mathematical statistics]. Moscow, Russia: BINOM. Laboratoriya znanij, [in Russian].
2. Kozlov M.V., Prohorov, A. V. (1987) *Vvedenie v matematicheskuyu statistiku* [Introduction to mathematical statistics]. Moscow, USSR: MSU, [in Russian].
3. Shiryaev, A. N. (2004) *Veroyatnost* [Probability]. Moscow, Russia: MCNMO, [in Russian].

4. Kotz, S., Nadarajah, S. (2004). *Multivariate T-Distributions and Their Applications.* Cambridge: Cambridge University Press. doi:10.1017/CBO9780511550683
5. Kibria, B. M. G., Joarder, A. H. (2006). A short review of multivariate t-distribution. *Journal of Statistical Research ISSN.* 40. 256-422.
6. Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer. ISBN 978-0-387-31073-2.
7. Peel, D., Mclachlan, G.. (2000). Robust Mixture Modelling Using the t Distribution. *Stat Comput.* 10. 10.1023/A:1008981510081.
8. Shoham S., Fellows M., Normann R. (2003). Robust, automatic spike sorting using mixtures of Multivariate T-Distributions. *Journal of neuroscience methods.* 127. 111-22. 10.1016/S0165-0270(03)00120-1.
9. Bishop C. M., Svensen M.. (2004). Robust Bayesian Mixture Modelling. *Neurocomputing*. 64. 235-252. 10.1016/j.neucom.2004.11.018.
10. Eaton M. L. (1983). Multivariate Statistics: a Vector Space Approach. *John Wiley and Sons.* pp. 116–117. ISBN 978-0-471-02776-8.
11. Gantmaher F. R. (2010) *Teoriya matric* [Matrix theory]. Moscow, Russia: Fizmatlit, [in Russian].
12. Fruhwirth-Schnatter S. (2006) Finite Mixture and Markov Switching Models. *Psychometrika*. 74. 559-560. 10.1007/s11336-009-9121-4.
13. Smith W. B., Hocking R. R. (1972) Algorithm AS 53: Wishart Variate Generator. *Applied Statistics*, 21, pp. 341-345.
14. Thomas P. M. (1997). Old and New Matrix Algebra Useful for Statistics. *MIT Media Lab note*.
15. Brusilovskij A. I. (2002) *Fazovye prevrashcheniya pri razrabotke mestorozhdenij nefti i gaza* [Phase transformations in the development of oil and gas fields]. Moscow, Russia: Graal, [in Russian].
16. Alakbari F., Elkatatny S., Baarimah S.. (2016). Prediction of Bubble Point Pressure Using Artificial Intelligence AI Techniques. *Proc. of the SPE Middle East Artificial Lift Conference and Exhibition*, 10.2118/184208-MS.
17. Numbere, O. G., Azuibuike, I. I., Ikiensikimama, S. S. (2013). Bubble Point Pressure Prediction Model for Niger Delta Crude using Artificial Neural Network Approach. *Society of Petroleum Engineers*. doi:10.2118/167586-MS
18. Alcocer Y., Patricia R.. (2001). Neural Networks Models for Estimation of Fluid Properties. *Proc. of the SPE Latin American and Caribbean Petroleum Engineering Conference*, 10.2523/69624-MS.
19. Osman, E. A., Abdel-Wahhab, O. A., Al-Marhoun, M. A. (2001). Prediction of Oil PVT Properties Using Neural Networks. *Society of Petroleum Engineers*. doi:10.2118/68233-MS
20. El-Sebakhy, E. A., Sheltami, T., Al-Bokhitan, S. Y., Shaaban, Y., Raharja et. al. (2007). Support Vector Machines Framework for Predicting the PVT Properties of Crude Oil Systems. *Society of Petroleum Engineers.* doi:10.2118/105698-MS