Covariance approximation of nonlinear regression

Article *in* Journal of Communications Technology and Electronics · January 2016

DOI: 10.1134/S106422691607007X

1 author:

Leonid Labunets
Bauman Moscow State Technical University

ISSN 1064-2269, Journal of Communications Technology and Electronics, 2016, Vol. 61, No. 7, pp. 789–806. © Pleiades Publishing, Inc., 2016.
Original Russian Text © L.V. Labunets, E.L. Labunets, N.L. Lebedeva, 2016, published in Radiotekhnika i Elektronika, 2016, Vol. 61, No. 7, pp. 652–670.

STATISTICAL RADIOPHYSICS

Covariance Approximation of Nonlinear Regression

L. V. Labunets[a, b], E. L. Labunets[c], and N. L. Lebedeva[d]

[a] Bauman Moscow State Technical University, Vtoraya Baumanskaya ul. 5, Moscow, 105005 Russia
[b] Russian New University, ul. Radio 22, Moscow, 105005 Russia
[c] JSB ROSEVROBANK (JSC), ul. Vavilova 24, Moscow, 119991 Russia
[d] VTB Bank, ul. Plyushchikha 37, Moscow, 119121 Russia
e-mail: labunets@bmstu.ru

Received May 20, 2015

Abstract—A nonlinear regression model on the basis of the covariance approximation of a multidimensional probability distribution is constructed. The model is represented by an expansion in the basis functions in the form of partial derivatives of the logarithm of the joint factor probability distribution. The weight coefficients of the expansion are the covariances of the resulting and explanatory variables. On particular examples, the efficiency of the Bayesian approximation of the proposed regression model in which the factor distribution is described by a finite mixture of ellipsoidally symmetric densities is demonstrated.

DOI: 10.1134/S106422691607007X

INTRODUCTION

Nowadays, the kernel methods of estimation are widely used in algorithms of intelligent data analysis (IDA) [1, 2]. The class of IDA problems includes finding hidden regularities by the classical Nadaraya—Watson kernel regression [3—6], the local weighted polynomial regression [7—10], the simulation and prediction of time series [9, 11], and the clusterization and recognition by means of radial basis functions [12, 13]. The kernel interpolation is an efficient instrument for the empirical mode decomposition of images [14].

An important class of kernel models for the IDA is comprised of problems of nonparametric and semiparametric estimation of one-dimensional and multidimensional densities, among which it is worth mentioning the average shifted histogram (ASH) [8], the adaptive kernels density estimation [15—17], and models of finite mixture of standard distributions [1, 18, 19].

A characteristic feature of the Nadaraya—Watson regression is the necessity to choose—subjectively to a certain extent—the parameters of kernel interpolation of data or apply rather computationally expensive methods of optimization of these parameters [13].

In our opinion, the subjectiveness of a kernel model of regression can be reduced by using an alternative approach based on the cumulant description of an N-dimensional probability distribution for explanatory, $\vec{x} = (x_1,\ldots,x_{N-1})^T$, and resulting, x_N, variables.

In practice, the main and sometimes the only information about a system of random variables (RV) X_1, ..., X_N that can be reliably estimated from experimental data is their covariance matrix $\mathbf{B}_N = \{b_{n,m}\}_{n=1,(N-1)}^{m=(n+1),N}$ and the one-dimensional probability density distributions (PDD) $\varphi_1^{X_1}(x_1)$, ..., $\varphi_1^{X_N}(x_N)$. In the framework of these statistics, a reasonable method for constructing a nonlinear regression model

$$\tilde{x}_N(\vec{x}) = \frac{\varepsilon_{N-1}(\vec{x})}{\varphi_{N-1}(\vec{x})},$$
$$\varepsilon_{N-1}(\vec{x}) = \int_{-\infty}^{+\infty} x_N \varphi_N(\vec{x}, x_N)\, dx_N \tag{1}$$

is the covariance approximation of a multidimensional PDD [20]:

$$\varphi_N(\vec{x}, x_N) = \sum_{k_{1,2}} \ldots \sum_{k_{(N-1),N}} \frac{b_{1,2}^{k_{1,2}} \cdot \ldots \cdot b_{(N-1),N}^{k_{(N-1),N}}}{k_{1,2}! \cdot \ldots \cdot k_{(N-1),N}!} \times \prod_{m=1}^{N} \frac{d^{k_m}}{dx_m^{k_m}} \{\varphi_1^{X_m}(x_m)\}. \tag{2}$$

Here, $\varphi_{N-1}(\vec{x})$ is the marginal factor distribution. The indices of summation of power series (2) are the elements of the upper triangular matrix $\{k_{n,m}\}_{n=1,(N-1)}^{m=(n+1),N}$, taking integer nonnegative values in the region

$$0 \le \sum_{n=1}^{N-1} \sum_{m=n+1}^{N} k_{n,m} \le K;\ k_m = \sum_{n=1}^{m-1} k_{n,m} + \sum_{n=m+1}^{N} k_{m,n}.$$

1. THE STRUCTURE OF THE COVARIANCE APPROXIMATION OF A NONLINEAR REGRESSION

The additive-multiplicative form of series (2) enables one to obtain in the explicit form the dependence of the conditional mathematical expectation of the resulting variable on the statistical characteristics of factors. We will assume that the PDD of the objective variable x_N has the Kth order of contact with the abscissa axis. In this case, the corresponding integrals take the form

$$\int_{-\infty}^{+\infty} x_N \frac{d^{k_N}}{dx_N^{k_N}} \left\{ \varphi_1^{X_N}(x_N) \right\} dx_N = \begin{cases} a_N, & k_N = 0; \\ -1, & k_N = 1; \\ 0, & k_N \ge 2. \end{cases} \quad (3)$$

Here, a_N is the unconditional expectation of the RV X_N. The subsequent calculation of integral (1) with allowance for (2) and (3) yields

$$\varepsilon_{N-1}(\vec{x}) = a_N \varphi_{N-1}(\vec{x}) - \sum_{n=1}^{N-1} b_{n,N} \theta_n (\vec{x} | \mathbf{B}_{N-1}), \quad (4)$$

where $\mathbf{B}_{N-1}$ is the covariance matrix of factors and

$$\theta_n(\vec{x}|\mathbf{B}_{N-1}) = \sum_{k_{1,2}} \dots \sum_{k_{(N-2),(N-1)}} \frac{b_{1,2}^{k_{1,2}} \cdot \ldots \cdot b_{(N-2),(N-1)}^{k_{(N-2),(N-1)}}}{k_{1,2}! \cdot \ldots \cdot k_{(N-2),(N-1)}!} \times \prod_{m=1}^{n-1} \frac{d^{k_m}}{dx_m^{k_m}} \left\{ \varphi_1^{X_m}(x_m) \right\} \times \frac{d^{k_n+1}}{dx_n^{k_n+1}} \left\{ \varphi_1^{X_m}(x_m) \right\} \prod_{m=n+1}^{N-1} \frac{d^{k_m}}{dx_m^{k_m}} \left\{ \varphi_1^{X_m}(x_m) \right\}.$$

It is easy to see that the latter equality is a partial derivative of the joint PDD of explanatory variables with respect to the factor x_n, i.e.,

$$\theta_n(\vec{x}|\mathbf{B}_{N-1}) = \frac{\partial}{\partial x_n} \left\{ \varphi_{N-1}(\vec{x}) \right\}.$$

As a result of the aggregation of (1) and (4), we obtain the covariance approximation of the model of nonlinear regression:

$$\begin{aligned} \tilde{x}_N(\vec{x}) &= a_N + \sum_{n=1}^{N-1} b_{n,N} \frac{\partial}{\partial x_n} \left(-\ln \left\{ \varphi_{N-1}(\vec{x}) \right\} \right) \\ &= a_N + \vec{b}_N^T \nabla_{\vec{x}} \left(-\ln \left\{ \varphi_{N-1}(\vec{x}) \right\} \right), \end{aligned} \quad (5)$$

where $\vec{b}_N = \left(b_{1,N}, \ldots, b_{(N-1),N} \right)^T$ is the Nth column vector of the covariances of factors $\vec{x}$ with the objective variable x_N in the matrix $\mathbf{B}_N$.

As expected, the regression model obtained has a pronounced grid architecture, similar to a neural network of radial basis functions [13]. The inputs of the network receive a standard bias signal (+1) and the factor variables $x_1, \ldots, x_{N-1}$. The role of the "radial" neurons of the hidden layer of the network, which provide the nonlinear transformation and mapping onto the space of informative features, is played by partial derivatives of the logarithm of the joint PDD of the explanatory variables. The linear neuron of the output layer of the network sums the reactions of hidden neurons, the weights of which are a_N and $b_{1,N}, \ldots, b_{(N-1),N}$.

2. THE BAYESIAN MODEL OF NONLINEAR REGRESSION

The further concretization of the model of nonlinear regression (5) is possible on the basis of approximating the joint PDD of the factors $\varphi_{N-1}(\vec{x})$ by a finite mixture

$$\Phi_{N-1}(\vec{x}) = \sum_{k=1}^{K} p_k \Phi_{(N-1),k} \left(\vec{x} \middle| \vec{A}_k, \mathbf{C}_k \right), \quad \sum_{k=1}^{K} p_k = 1 \quad (6)$$

of elliptically symmetric partial distributions

$$\Phi_{(N-1),k} \left(\vec{x} | \vec{A}_k, \mathbf{C}_k \right) = \frac{f \left\{ r^2 \left(\vec{x} \middle| \vec{A}_k, \mathbf{C}_k \right) \right\}}{S_1 \nu_{N-2} \sqrt{\det(\mathbf{C}_k)}}. \quad (7)$$

Here, $r^2 \left(\vec{x} | \vec{A}_k, \mathbf{C}_k \right) = \left(\vec{x} - \vec{A}_k \right)^T \mathbf{C}_k^{-1} \left(\vec{x} - \vec{A}_k \right)$ is the squared Mahalanobis distance; p_k, $\vec{A}_k = \left(a_{k,1}, \ldots, a_{k,N-1} \right)^T$, and $\mathbf{C}_k = \left\{ c_{k,n,m} \right\}_{n=1,(N-1)}^{m=1,(N-1)}$ are the a priori probability, the vector of expectation, and the covariance matrix of the kth partial distribution; $S_1 = 2\sqrt{\pi^{N-1}} / \Gamma\{(N-1)/2\}$ is the surface area of a unit sphere in the $(N-1)$-dimensional factor space; $S_1^{-1} \nu_{n-2}^{-1} f(r^2)$ is a radial distribution monotonically decreasing with $r \to \infty$ with the finite moment $\nu_{N-2} = \int_0^\infty r^{n-2} f(r^2) dr$.

In practice, ellipsoidally symmetric distributions (7) are usually simulated with using the following three radial basis functions [12, 19, 21, 22]:

the Gaussian distribution:

$$f\left(r^2\right) = \exp\left(-\frac{r^2}{2}\right), \; \nu_{N-2} = \frac{1}{2} \sqrt{2^{N-1}} \Gamma\left(\frac{N-1}{2}\right);$$

the Pearson distribution with a parameter $\gamma > 1$:

$$f\left(r^2\right) = \begin{cases} \left(1 - \dfrac{r^2}{2\gamma + N - 1} \right)^{1-\gamma}, & r^2 \le 2\gamma + N - 1, \\ 0, & r^2 > 2\gamma + N - 1, \end{cases}$$

$$\nu_{N-2} = \frac{1}{2} \sqrt{2\gamma + N - 1} \, \mathrm{B}\left(\gamma, \frac{N-1}{2}\right);$$

the Student's distribution with a parameter $\gamma > N-2$:

$$f\left(r^2\right) = \left(1 + \frac{r^2}{\gamma - N}\right)^{-\frac{\gamma+1}{2}},$$

$$\nu_{N-2} = \frac{1}{2}\sqrt{\gamma - N}\,\mathrm{B}\left(\frac{\gamma + 2 - N}{2}, \frac{N-1}{2}\right),$$

where $\mathrm{B}(\alpha,\beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$ is the beta function.

The analysis of the partial derivatives of the squared Mahalanobis distance of the kth partial factor distribution is conveniently performed in terms of the accuracy matrix $\mathbf{C}_k^{-1} = \left\{c_k^{(n,m)}\right\}_{n=1,(N-1)}^{m=1,(N-1)}$, i.e., the inverse of the covariance matrix $\mathbf{C}_k$, namely,

$$r^2\left(\vec{x}|\,\vec{A}_k, \mathbf{C}_k\right) = \sum_{n=1}^{N-1} c_k^{(n,n)}\left(x_n - a_{k,n}\right)^2 + 2\sum_{n=1}^{N-2}\sum_{m=n+1}^{N-1} c_k^{(n,m)}\left(x_n - a_{k,n}\right)\left(x_m - a_{k,m}\right).$$

(Below, if not necessary, we do not present the dependence of the Mahalanobis distance $r\left(\vec{x}|\,\vec{A}_k, \mathbf{C}_k\right)$ on the statistics of the position $\vec{A}_k$ and scale $\mathbf{C}_k$.) The corresponding partial derivatives take the form

$$\frac{\partial}{\partial x_n}\left\{r^2\left(\vec{x}\right)\right\} = \sum_{m=1}^{N-1} c_k^{(m,n)}\left(x_m - a_{k,m}\right), \quad n = 1, 2, \ldots, (N-1). \tag{8}$$

The differentiation of the logarithm of finite mixture (6) with allowance for the radial basis functions presented above and formula (8) yield the Bayesian approximation for covariance model (5) of nonlinear regression:

$$\tilde{x}_N\left(\vec{x}\right) = a_N + \sum_{k=1}^{K} \Pr\left(k|\,\vec{x}\right) R_{k,N}\left(\vec{x}\right), \tag{9}$$

where $\Pr\left(k|\,\vec{x}\right) = p_k \Phi_{(N-1),k}\left(\vec{x}|\,\vec{A}_k, \mathbf{C}_k\right)/\Phi_{N-1}\left(\vec{x}\right)$ is the a posterior probability of the association of the current values of factors with the kth partial kernel $R_{k,N}\left(\vec{x}\right)$. The type of the kernel is determined by the choice of radial basis functions. In the case of the Gaussian model, the kernel has the form

$$R_{k,N}\left(\vec{x}\right) = S_{k,N}\left(\vec{x}\right); \tag{10}$$

in the case of the Pearson model,

$$R_{k,N}\left(\vec{x}\right) = \begin{cases} \dfrac{2(1-\gamma) S_{k,N}\left(\vec{x}\right)}{2\gamma + N - 1} \times \left\{1 - \dfrac{r^2\left(\vec{x}\right)}{2\gamma + N - 1}\right\}^{-1}, & r^2\left(\vec{x}\right) \le 2\gamma + N - 1, \\ 0, & r^2\left(\vec{x}\right) > 2\gamma + N - 1; \end{cases}$$

in the case of the Student's model,

$$R_{k,N}\left(\vec{x}\right) = \frac{(\gamma + 1) S_{k,N}\left(\vec{x}\right)}{\gamma - N}\left\{1 + \frac{r^2\left(\vec{x}\right)}{\gamma - N}\right\}^{-1}.$$

Here, the local approximating hyperplane

$$S_{k,N}\left(\vec{x}\right) = \sum_{m=1}^{N-1} W_{k,N}^{(m)}\left(x_m - a_{k,m}\right),$$
$$W_{k,N}^{(m)} = \sum_{n=1}^{N-1} c_k^{(m,n)} b_{n,N}$$

is associated with the kth partial kernel. The expression for this plane is conveniently written in the vector form:

$$S_k\left(\vec{x}|\,\vec{A}_k, \mathbf{C}_k\right) = \vec{b}_N^T \mathbf{C}_k^{-1}\left(\vec{x} - \vec{A}_k\right). \tag{11}$$

The factorization of the accuracy matrix,

$$\mathbf{C}_k^{-1} = \mathbf{U}_k \mathbf{\Lambda}_k^{-1/2} \mathbf{\Lambda}_k^{-1/2} \mathbf{U}_k^T = \left(\mathbf{\Lambda}_k^{-1/2} \mathbf{U}_k^T\right)\left(\mathbf{U}_k \mathbf{\Lambda}_k^{-1/2}\right),$$

makes it possible to represent the quadratic form in equality (11) in terms of a transformation analogous to the decorrelating transformation [23]:

$$\begin{aligned} S_k\left(\vec{x}|\,\vec{A}_k, \mathbf{C}_k\right) &= \left(\mathbf{U}_k \mathbf{\Lambda}_k^{-1/2} \vec{b}_N\right)^T \left\{\mathbf{U}_k \mathbf{\Lambda}_k^{-1/2}\left(\vec{x} - \vec{A}_k\right)\right\} \\ &= \left(\mathbf{\Lambda}_k^{-1/2} \vec{b}_N\right)^T \mathbf{U}_k^T \mathbf{U}_k \left\{\mathbf{\Lambda}_k^{-1/2}\left(\vec{x} - \vec{A}_k\right)\right\} = \vec{w}_{k,N}^T \vec{y}_k. \end{aligned} \tag{12}$$

Here, $\mathbf{\Lambda}_k^{1/2}$ and $\mathbf{U}_k$ are the matrices of singular numbers and orthonormal eigenvectors of the covariance matrix of the kth partial distribution; $\vec{w}_{k,N} = \mathbf{\Lambda}_k^{-1/2}\vec{b}_N$ is the normalized column vector of covariances of the objective variable x_N with the factors $\vec{x}$; and $\vec{y}_k = \mathbf{\Lambda}_k^{-1/2}\left(\vec{x} - \vec{A}_k\right)$ is the column vector of centered and normalized factors.

Now, let us consider in more detail the covariance approximation of nonlinear regression for the poly-Gaussian description of joint PDD of factors, which is popular in practical applications.

3. THE POLY-GAUSSIAN COVARIANCE MODEL OF NONLINEAR REGRESSION

In the case of the poly-Gaussian approximation of the PDD of factors, the Bayesian approximation of the model of nonlinear regression is obtained by the subsequent substitution of formulas (12) and (10) into expression (9):

$$\tilde{x}_N(\vec{x}) = a_N + \sum_{k=1}^{K} \Pr(k|\vec{x})\,\vec{w}_{k,N}^{T}\vec{y}_k. \qquad (13)$$

The geometrical meaning of this model is conveniently analyzed in the N-dimensional informative space of factors and the resulting variable. At a current point $(\vec{x}, x_N)$, the regression surface $\tilde{x}_N(\vec{x})$ is a local hyperplane. The bias of the hyperplane, i.e., the distance from the origin of coordinates of the informative space to the plane along the normal to it is

$$d_N(\vec{x}) = \frac{a_N - \vec{b}_N^{T}\sum_{k=1}^{K}\Pr(k|\vec{x})\,\mathbf{\Lambda}_k^{-1}\vec{A}_k}{\left\|\vec{b}_N^{T}\sum_{k=1}^{K}\Pr(k|\vec{x})\,\mathbf{\Lambda}_k^{-1}\right\|}.$$

The rotation of the hyperplane is determined by the direction of the normal to it:

$$\vec{u}_N(\vec{x}) = \frac{\vec{b}_N^{T}\sum_{k=1}^{K}\Pr(k|\vec{x})\,\mathbf{\Lambda}_k^{-1}}{\left\|\vec{b}_N^{T}\sum_{k=1}^{K}\Pr(k|\vec{x})\,\mathbf{\Lambda}_k^{-1}\right\|}.$$

In other words the weighted mean position and orientation of the local approximating hyperplane are determined by the confidence levels $\Pr(k|\vec{x})$ with which the current values of the factors $\vec{x}$ are associated with the kth partial distribution. Thus, the Bayesian weighting provides the adaptation of the parameters of linear regression to the local regions of the factor subspace.

It should be noted that poly-Gaussian approximation (13) of the covariance model of nonlinear regression conforms to the nonlinear principal component analysis (NLPCA) [24, 25]. In our opinion, beside the hierarchical NLPCA model of data analysis presented in [25, Section 2.3], a reasonable neural network implementation of Bayesian model (13) may be an associative computer with the mixture-of-experts (ME) network architecture [13] (Fig. 1). The associative computer involves K expert networks and an aggregating unit in the form of a gateway network. Each expert suggests a model of linear regression $\vec{w}_{k,N}^{T}\vec{y}_k$, $k = 1, 2, \ldots, K$, the best in the kth local region of the space of factors $\vec{x}$. The gateway network involves K nonlinear neurons estimating the confidence levels $\Pr(k|\vec{x})$ of experts' opinions. The output linear neuron of the computer aggregates the experts' opinions into the mean weighted estimate (13) of the regression surface. As is known [13], the associative computer is a universal approximator.

In the following numerical experiments, we will analyze the results of application of regression model (13) to particular data.

4. ONE-DIMENSIONAL REGRESSION

For two-dimensional data ($N = 2$), the poly-Gaussian approximation (13) of the covariance model of nonlinear regression takes the form

$$\tilde{x}_2(x_1) = a_2 + \sum_{k=1}^{K}\Pr(k|x_1)\frac{b_{1,2}}{\sigma_{k,1}}\frac{x_1 - a_{k,1}}{\sigma_{k,1}}, \qquad (14)$$

where $\sigma_{k,1} = \sqrt{\lambda_{k,1}}$ is the mean squared deviation of the kth partial Gaussian PDD of the factor x_1. It is obvious that, for $K = 1$, equality (14) is equivalent to the equation of one-dimensional linear regression.

Nonlinear regression (14) was tested on a problem of simulation and prediction of a time series (TS) $v(m), m = 1, 2, \ldots, 144$, of the monthly air traffic volume in thousands of air-passengers [26] during 12 years. Approximation (14) was used for forming a nonlinear regression dependence $u(m) = \theta\{u(m-12)\} + e(m)$. Here, $e(m)$ is the data simulation error and

$$u(m) = \{v(m) - a(m)\}/s(m)$$

is the centered and normalized TS obtained by polynomial estimates of the trend, $a(m) = 116.97 + 1.268m + 0.0134m^2$, and volatility, $s(m) = 11.411 + 0.2226m + 0.0017m^2$. In other words, the factor x_1 and the objective variable x_2 were associated with the lag, $u(m-12)$, and current, $u(m)$, samples of the TS, respectively. Below we present the sample estimates of the unconditional statistics of these variables, necessary for the implementation of model (14):

Statistics	a_2	$b_{1,2}$
Estimate	–0.003751	0.909439

The sample estimate of the PDD of the x_1 in the form of a histogram smoothed by the bias of the Gaussian kernel is shown in Fig. 2 (curve *1*). The histogram was obtained from five sampling intervals and five narrowed subintervals. Then the ASH-estimate was approximated by a mixture of two Gaussian distributions (curve *2*) by a modified EM-algorithm [19, 27, 28]. The parameters of the given poly-Gaussian mixture (6) are presented in Table 1. These parameters are characterized by an overestimated value $\sigma = 0.051$ of

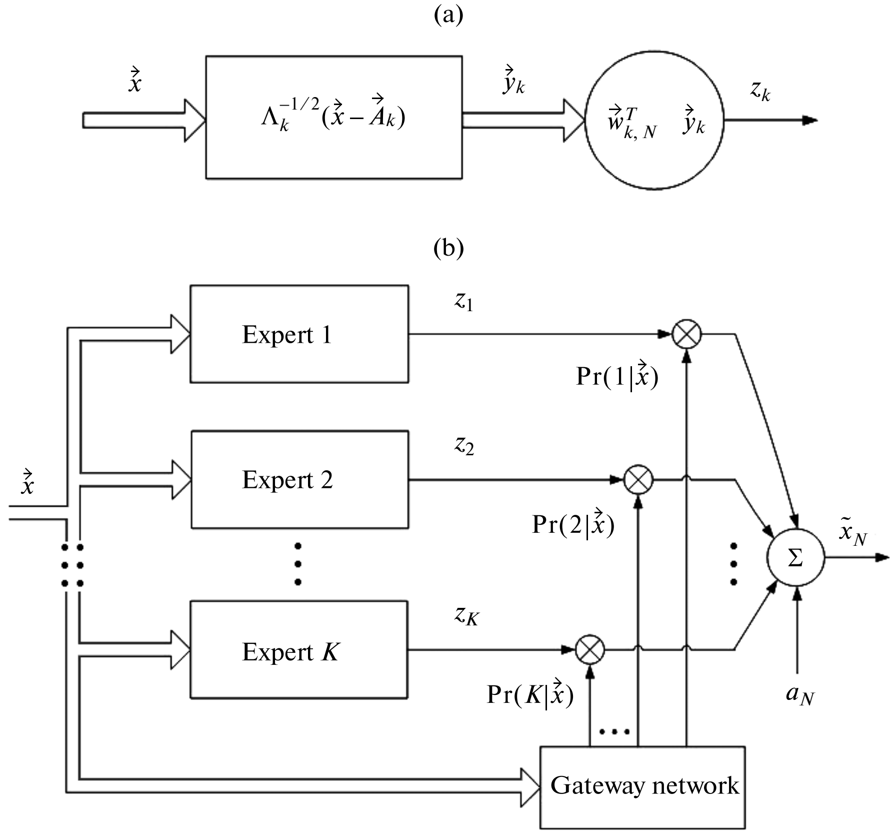


Fig. 1. Experts' bias network: (a) the kth expert's network and (b) ME-network.

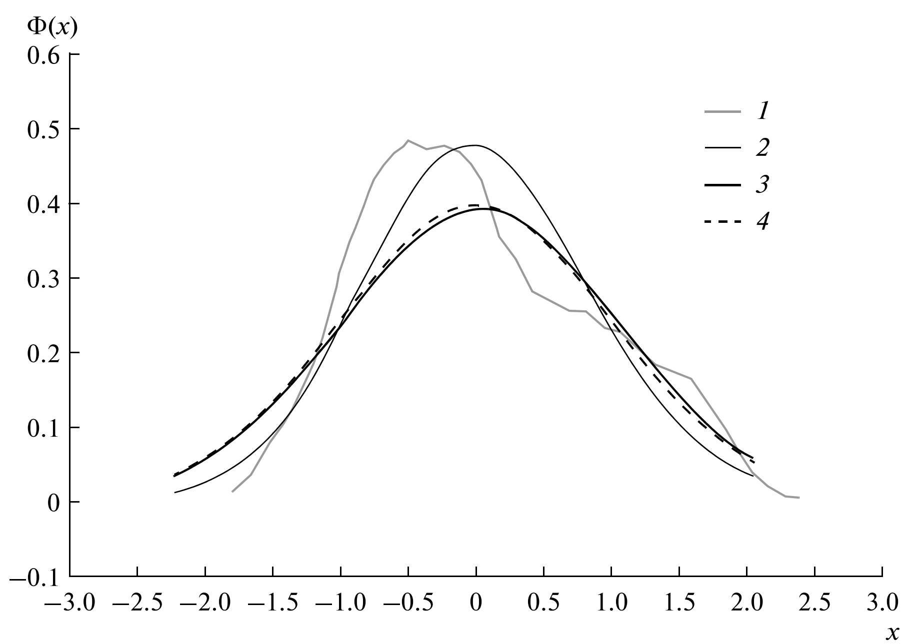


Fig. 2. Estimates of the PDD factor: (*1*) the ASH-estimate of the distribution; (*2*) poly-Gaussian approximation; (*3*) optimal poly-Gaussian model; (*4*) Gaussian curve of the corresponding linear regression.

Table 1. Parameters of nonlinear regression of air transport

k	p_k	$a_{k,1}$	$\sigma_{k,1}$	$\sigma_k^{(\mathrm{opt})}$
1	0.65	–0.1432	0.7625	1.01
2	0.35	0.3563	0.9156	0.97

the mean squared deviation (MSD) of the data simulation error $e(m)$.

The further adjustment of the scales $\sigma_{1,1}$ and $\sigma_{2,1}$ of regression model (14) provides a smaller MSD of $\sigma = 0.0375$. The optimal values $\sigma_k^{(\mathrm{opt})}$ of these parameters are also presented in Table 1. The two-component poly-Gaussian model for the PDD of the factor x_1, corresponding to the optimal values of the parameters of regression dependence (14), is presented in Fig. 2 (curve *3*). This PDD perfectly agrees with the Gaussian distribution $N(-0.002;1.008)$ (see Fig. 2, curve *4*) corresponding to the model of linear regression. The dependence of the a posterior probabilities $\Pr(k|x_1)$ of classes $k = 1$, 2 on the factor x_1 for the optimal two-component poly-Gaussian mixture (6) is presented in Fig. 3. The scatter diagram of the factor, $u(m-12)$, and objective, $u(m)$, variables (Fig. 4) demonstrates the consistency between the model of linear regression and the poly-Gaussian approximation of nonlinear regression. The results of simulation of the centered and normalized TS $u(m)$ of the passenger air traffic and its prediction to 12 months ahead by the optimal regression dependence (14) are presented in Fig. 5.

5. DYNAMIC ONE-DIMENSIONAL REGRESSION

The poly-Gaussian approximation (13) of the covariance model of nonlinear regression should be used for the simulation of the dynamic dependence of the objective variable $x_N(t)$ on the factors $\vec{x}(t)$:

$$\tilde{x}_N(t) = \sum_{j=1}^{J} g_j(t)\,\tilde{x}_N^{(j)}\{\vec{x}(t)\}, \quad \sum_{j=1}^{J} g_j(t) = 1, \qquad (15)$$

where $t_0 \le t \le t_J$ is time; $0 \le g_j(t) \le 1$ are the subjective probabilities with which the local nonlinear regression models

$$\tilde{x}_N^{(m)}\{\vec{x}(t)\} = a_N^{(m)} + \sum_{k=1}^{K} \Pr_m\{k|\vec{x}(t)\}\,\vec{w}_{m,k,N}^{T}\vec{y}_{m,k}(t),$$
$$m = 1, 2, \ldots, J$$

adequately approximate the objective variable $x_N(t)$ in the current time interval $t_{j-1} \le t \le t_j$; and $\vec{y}_{m,k}(t) = \Lambda_{m,k}^{-1/2}\{\vec{x}(t) - \vec{A}_{m,k}\}$ is the centered and normalized column vector of factors in the mth time interval. In our opinion, the dynamic dependence (15) of the objective variable on the factors should be called the Gaussian garland.

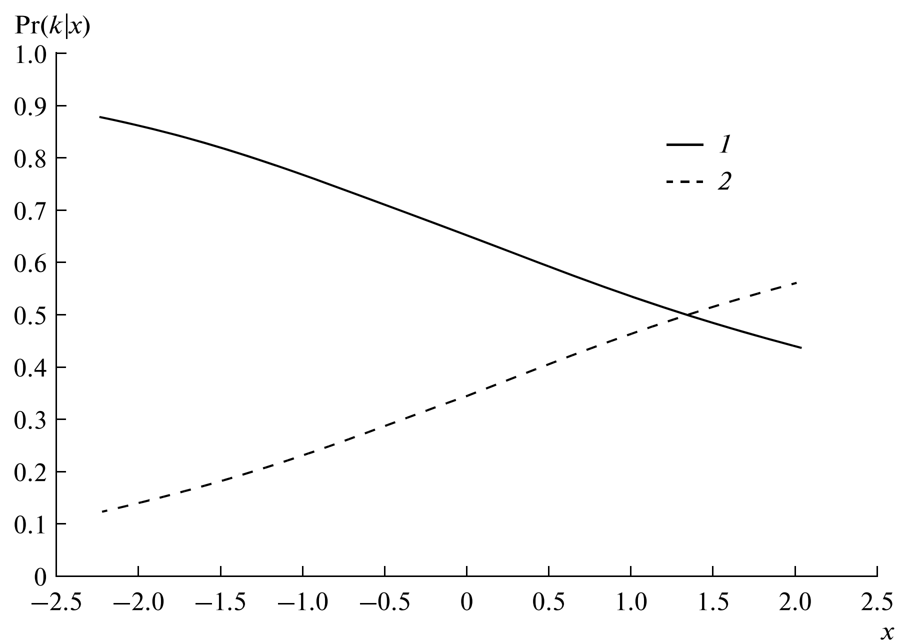


Fig. 3. A posteriori probabilities of the classes of optimal poly-Gaussian model of the PDD of the factor: $k = (1)$ 1 and (2) 2.

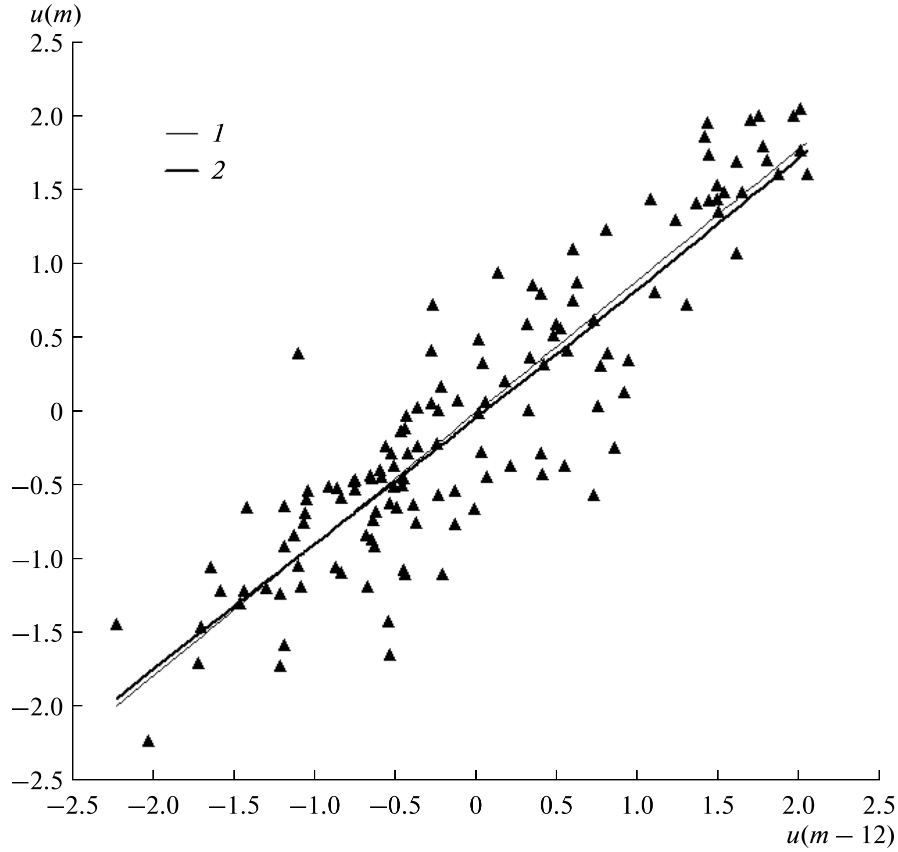


Fig. 4. Scatter diagram of the factor and objective variables: (*1*) linear regression and (*2*) poly-Gaussian model of nonlinear regression.

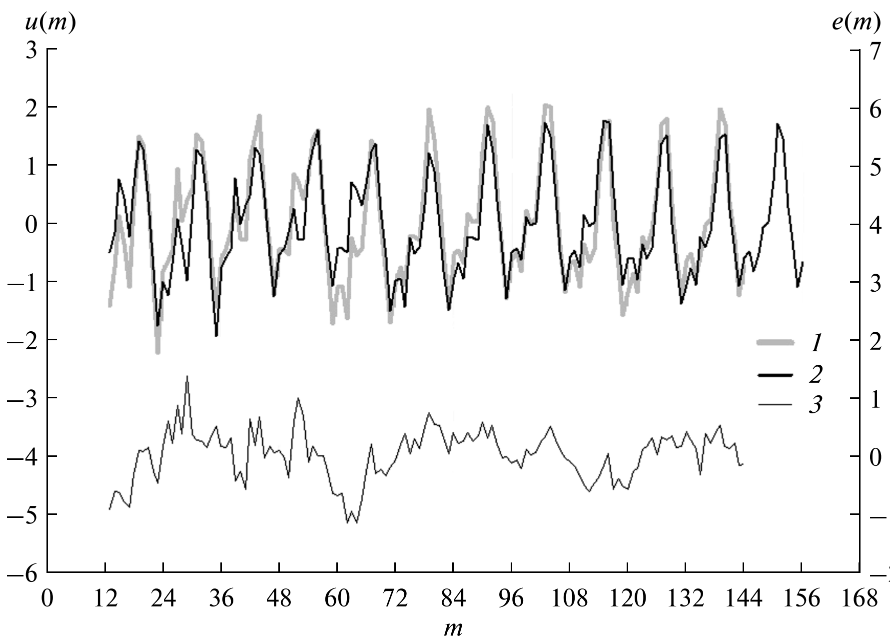


Fig. 5. Poly-Gaussian covariance model of nonlinear regression: (*1*) centered and normalized TS $u(m)$; (*2*) prediction to 12 months; (*3*) error $e(m)$ of simulation of data.

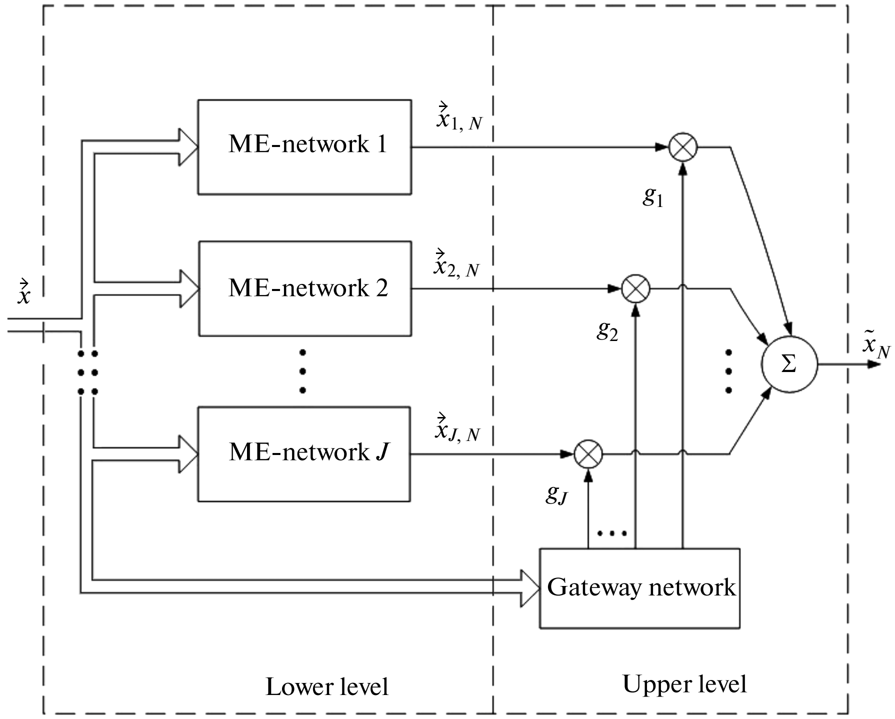


Fig. 6. Hierarchical experts' bias network.

This kind of a regression model conforms to the hierarchical-mixture-of-experts (HME) architecture [13] (Fig. 6). The dynamic HME network consists of two hierarchical levels. The lower level involves J ME expert networks forming nonlinear regression hyperspaces $\tilde{x}_{j,N}\{\vec{x}(t)\}$, $j = 1, 2, \ldots, J$, in the space of factors $\vec{x}$. Each ME network adequately approximates the data in a fixed time interval $[t_{j-1}, t_j]$. The gateway network of the upper level aggregates the output reactions of the ME networks, using a soft transition from a model to model by weighting them with subjective probabilities $g_j(t)$ over time intervals. Here, it also worth referring to the analogy with the auto-associative models [25, Ch. 8] of the NPLCA approach to the analysis of data.

The approach considered above was applied for the simulation of the dynamic regression dependence of the rate of the USA dollar established by the Central Bank of the Russian Federation (the Bank of Russia) on the price of Brent crude oil. The scattering diagram for these assets since April 4, 2012 to January 16, 2013 is shown in Fig. 7. An obviously practically important question is how the volatility of the rate of dollar established by the Bank of Russia conforms to the dynamics of the oil price.

The answer to this question is given by the two-dimensional ($N = 2$) Gaussian garland

$$\tilde{x}_2(t) = \sum_{j=1}^{3} g_j(t)\tilde{x}_2^{(j)}\{x_1(t)\}, \quad \sum_{j=1}^{3} g_j(t) = 1,$$
$$\tilde{x}_2^{(j)}\{x_1(t)\} = a_2^{(j)} + \sum_{k=1}^{K_j} \Pr_j\{k|x_1(t)\}\frac{b_{1,2}^{(j)}}{\sigma_{k,1}^{(j)}}\frac{x_1(t) - a_{k,1}^{(j)}}{\sigma_{k,1}^{(j)}}. \tag{16}$$

This garland was formed for three time intervals ($J = 3$): the first interval, from April 4, 2012 to June 22, 2012; the second interval, from June 22, 2012 to August 23, 2012; and the third interval, from August 23, 2012 to January 16, 2013. The subjective probabilities $g_j(t)$ were trapezoid models with seven-day long linear sections of transition from zero to unity, arranged symmetrically with respect to the dates of June 22, 2012 and August 23, 2012. Sample estimates of the descriptive statistics of assets for the above time intervals are presented in Table 2.

The estimates of the PDD of the factor x_1 for three time intervals were histograms smoothed by the bias of thrice weighted Epanechnikov kernel. The histograms were formed over five sample intervals and two narrowed subintervals. Figure 8 shows the ASH-estimate (curve *1*) for the first time interval. Then the ASH-estimates were approximated by a mixture of three Gaussian distributions (curve *2*) with the help of a

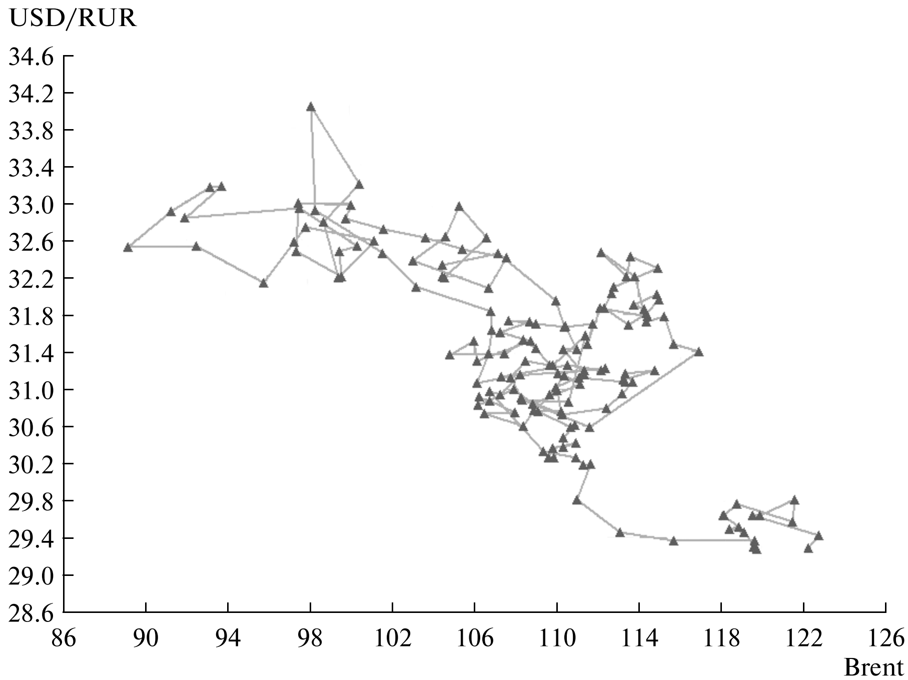


Fig. 7. Scatter diagrams of the prices of Brent crude oil and the USD/RUR currency pair.

modified EM-algorithm. The parameters of these poly-Gaussian descriptions lead to overestimated MSDs of the data approximation error: $\sigma^{(1)} = 0.6678$, $\sigma^{(2)} = 0.1951$, and $\sigma^{(3)} = 0.0675$ for the three time intervals, respectively.

The further adjustment of the parameters of regression models (16) for the three time intervals provided smaller values of the MSDs: $\sigma^{(1)} = 0.0538$, $\sigma^{(2)} = 0.0412$, and $\sigma^{(3)} = 0.0549$, respectively. The quasi-optimal values of the parameters of these models are presented in Table 3, and the quasi-optimal poly-Gaussian model for the PDD of factor x_1 for the first time interval is presented in Fig. 8 (curve *3*). This PDD agrees well with the Gaussian distribution $N(109.62; 9.42)$ (curve *4*), corresponding to the linear regression model. Figure 9 illustrates the dependence of the a posteriori probabilities $\Pr_1\{k|x_1(t)\}$ $(k = 1, 2, 3)$ on the factor x_1 in the first time interval for the quasi-optimal three-component poly-Gaussian mixture (16). The models of linear regression and covariance approximation of the nonlinear regression for the first time interval are presented in Fig. 10. The resulting Gaussian garland approximating on the average the dynamics of the behavior of the scattering matrix of the Brent–USD/RUR assets is illustrated by Fig. 11. The rela-

Table 2. Descriptive characteristics of the prices of Brent crude oil and the USD/RUR currency pair for three time intervals

Interval	j	$a_1^{(j)}$	$a_2^{(j)}$	$b_{1,2}^{(j)}$
April 4, 2012– June 22, 2012	1	109.62	30.50	–12.3572
June 22, 2012– August 23, 2012	2	104.46	32.39	–2.6444
August 23, 2012– January 16, 2013	3	110.52	31.21	2.0627

Table 3. Optimal parameters of nonlinear regression of the prices of Brent crude oil to the USD/RUR currency pair

k	$p_k^{(j)}$	$a_{k,1}^{(j)}$	$\sigma_{k,1}^{(j)}$
$j = 1, \sigma^{(1)} = 0.0538$			
1	0.1657	100.84	9.0
2	0.3616	108.60	5.0
3	0.4727	117.60	10.0
$j = 2, \sigma^{(2)} = 0.0412$			
1	0.2689	99.27	6.0
2	0.3633	105.16	4.4
3	0,3678	108.0	8.0
$j = 3, \sigma^{(3)} = 0.0549$			
1	0.2135	108.11	1.300
2	0.4225	110.24	1.061
3	0.3640	112.50	1.467

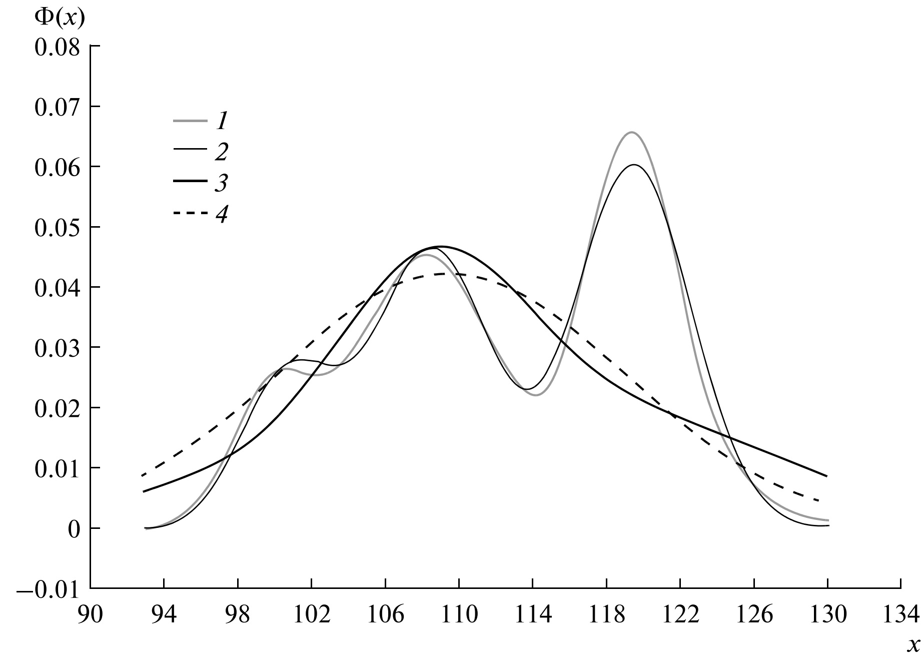


Fig. 8. Estimates of the PDD of the prices of Brent crude oil: (*1*) ASH-estimate of the distribution; (*2*) poly-Gaussian approximation; (*3*) optimal poly-Gaussian model; (*4*) Gaussian curve of the corresponding linear regression.

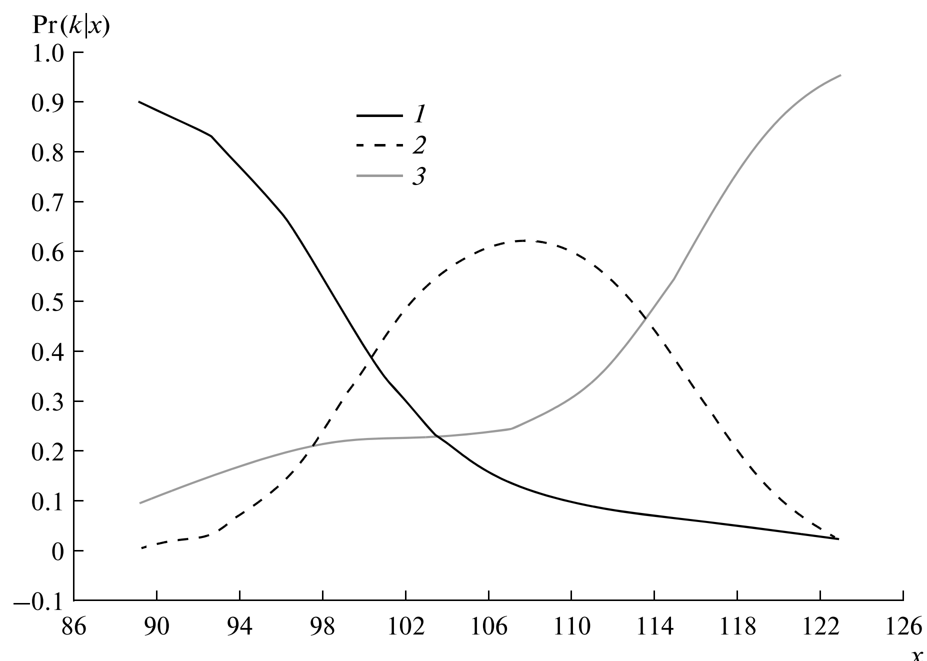


Fig. 9. A posteriori probabilities of the classes of optimal poly-Gaussian model of the PDD of the prices of Brent crude oil: $k = (1)$ 1, (*2*) 2, and (*3*) 3.

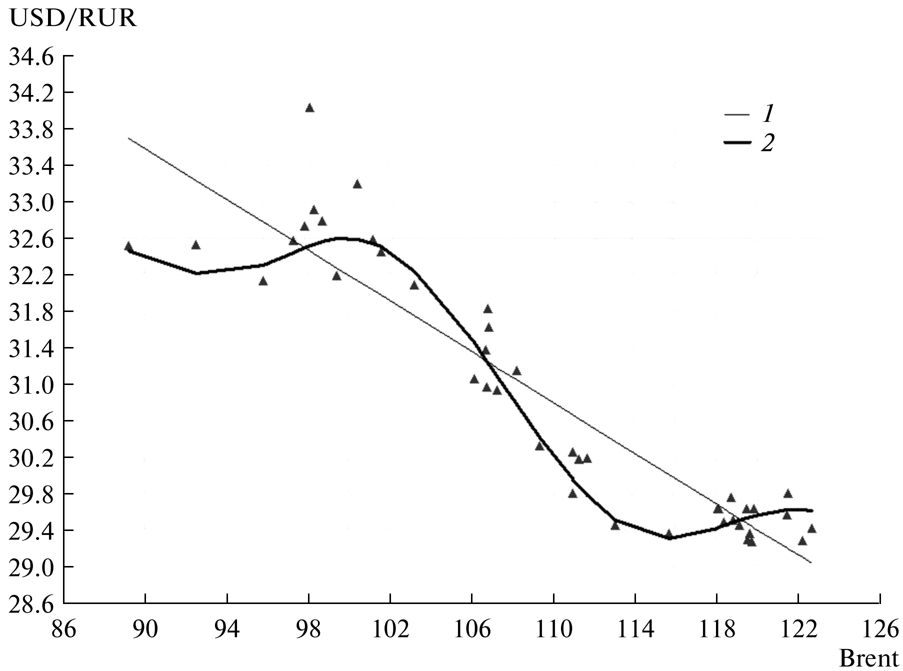


Fig. 10. Scatter diagrams of Brent–USD/RUR since April 4, 2012 to June 22, 2012: (*1*) linear regression and (*2*) poly-Gaussian model of nonlinear regression.

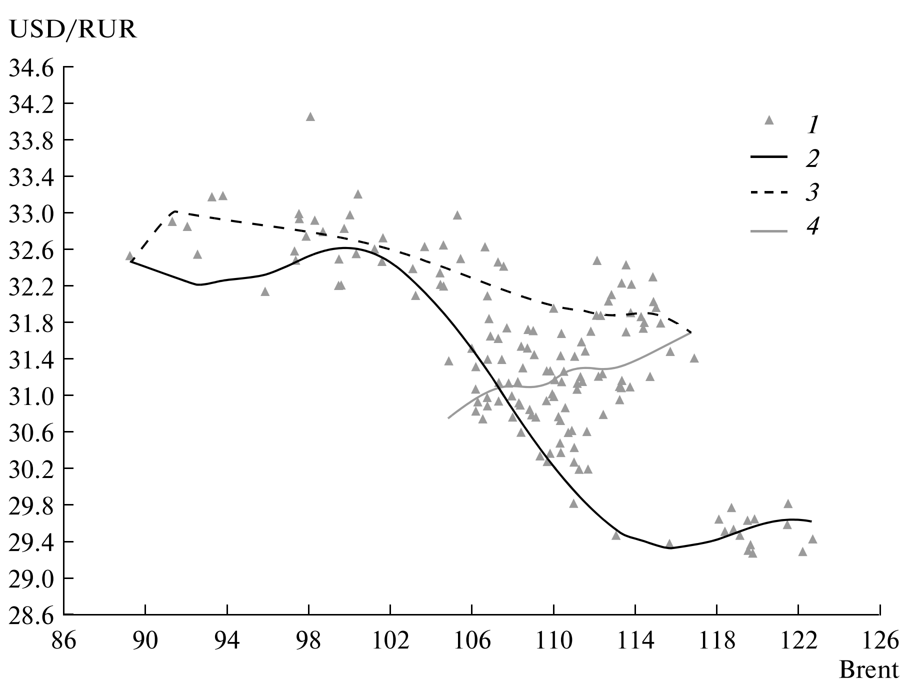


Fig. 11. Gaussian garland: (*1*) scatter diagrams of Brent–USD/RUR; (*2*–*4*) poly-Gaussian model of nonlinear regression in three time intervals, respectively.

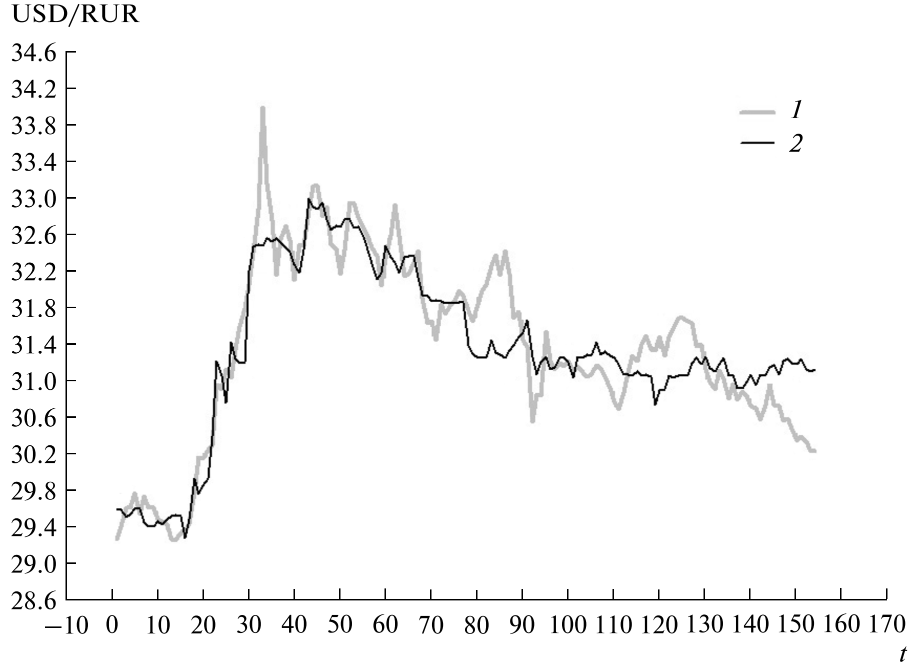


Fig. 12. Poly-Gaussian covariance model of nonlinear regression: (*1*) USD/RUR exchange rate established by the Bank of Russia since April 4, 2012 to January 16, 2013; (*2*) the model of nonlinear regression.

tionship between the time series in the form of the USA dollar rate established by the Central Bank of Russia and its regression model in relation to daily prices of Brent crude oil is demonstrated by Fig. 12. It should be noted that this models provide lower volatility of the rate of this pair of currencies than that of the rate established by the Central Bank of Russia.

6. TWO-DIMENSIONAL REGRESSION

For three-dimensional data (N = 3), the poly-Gaussian approximation (13) of the covariance model of nonlinear regression takes the form

$$\tilde{x}_3(x_1,x_2) = a_3 + \sum_{k=1}^{K}\frac{\Pr(k|x_1,x_2)}{1-\rho_{k,1,2}^2}\left\{\left(\frac{b_{1,3}}{\sigma_{k,1}} - \rho_{k,1,2}\frac{b_{2,3}}{\sigma_{k,2}}\right)\times\frac{x_1-a_{k,1}}{\sigma_{k,1}} + \left(\frac{b_{2,3}}{\sigma_{k,2}} - \rho_{k,1,2}\frac{b_{1,3}}{\sigma_{k,1}}\right)\frac{x_2-a_{k,2}}{\sigma_{k,2}}\right\}, \quad (17)$$

where $\sigma_{k,1}$, $\sigma_{k,2}$, and $\rho_{k,1,2}$ are the mean squared deviations and the correlation coefficient of the kth partial Gaussian PDD of the factors x_1 and x_2. It is obvious that, for $K = 1$, equality (17) is equivalent to the equation of two-dimensional linear regression.

Bayesian approximation (17) was applied for constructing a nonlinear regression model of the objective TS $v(m)$, $m = 1,2,\ldots,252$, of the monthly sales volume of transport vehicles (Fig. 13, curve *3*) on two factor TSs: sales volume of chemical industry (curve *1*) and oil products (curve *2*) since January 1971 to December 1991 [29]. These are sample estimates of the unconditional statistics of the TSs, necessary for the implementation of model (17):

Statistics	a_3	$b_{1,3}$	$b_{2,3}$
Estimate	11.743	13.390418	28.437772

The scatter diagram of the explanatory TS with time as a parameter demonstrates a pronounced cluster structure in the dynamics of data (Fig. 14). Therefore, in the first state of the covariance model of nonlinear regression (17), we took five ($K = 5$) classes of data, corresponding to five time intervals: 01.01.71–31.12.74, 01.01.75–30.06.79, 01.07.79–31.08.83, 01.09.83–31.12.87, and 01.01.88–31.12.91, with 48, 54, 50, 52, and 48 samples, respectively.

The parameters of the two-dimensional poly-Gaussian approximation (6) of the PDD of factors (Fig. 15a), obtained by the EM-algorithm, are presented in Table 4. The values of the MSD of the error $e(m)$ of approximating the objective TS by regression model (17) evidence unacceptably large errors of simulation of data.

The subsequent adjustment of the mathematical expectation (ME) a_3 of the objective TS and the MEs $a_{k,1}$ and $a_{k,2}$ and the MSDs $\sigma_{k,1}$ and $\sigma_{k,2}$, $k = 1, 2, \ldots, 5$, for the classes of factor TSs reduce the MSD of the error $e(m)$ by an order of magnitude (Table 5). The corresponding poly-Gaussian approximation (6) of the PDD of factors is illustrated by Fig. 15b. Figures 16a–16e

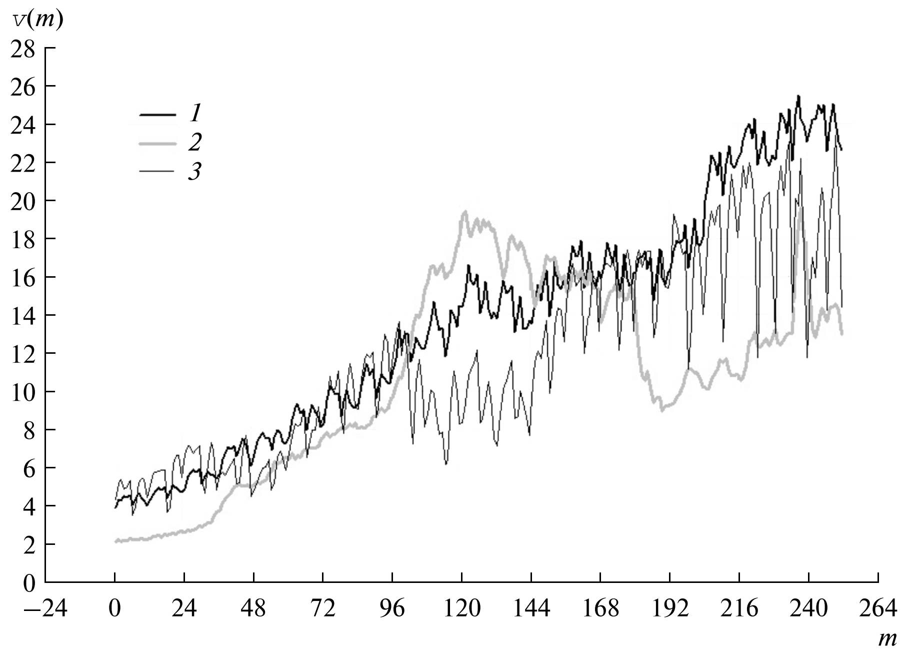


Fig. 13. TS $v(m)$ of monthly sales of oil products and accompanying goods: (*1*) volume of chemical production; (*2*) oil products; (*3*) transport vehicles.

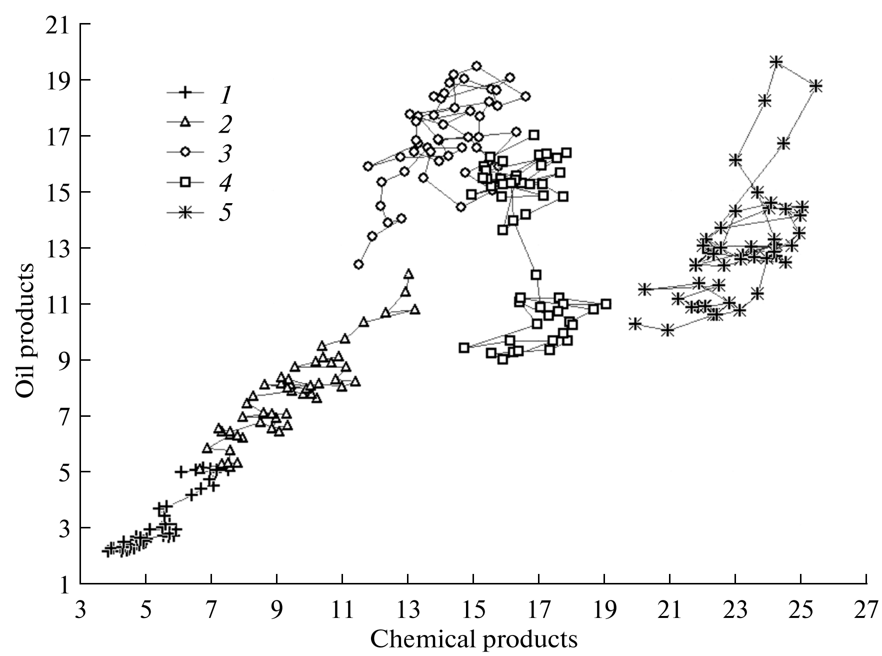


Fig. 14. Cluster structure of the dynamics of factor TSs.

show the dependences of the a posteriori probabilities $\Pr(k|x_1, x_2)$ of classes $k = 1, 2, \ldots, 5$ on the factors x_1 and x_2 for the initial data (left) and for the quasi-optimal regression model (right).

Figure 17a presents the quasi-optimal regression approximation (17) in the case of five classes (curve *1*) for the TS $v(m)$ of monthly sales volume transport vehicles (curve *2*) and the TS of the error $e(m)$ and the

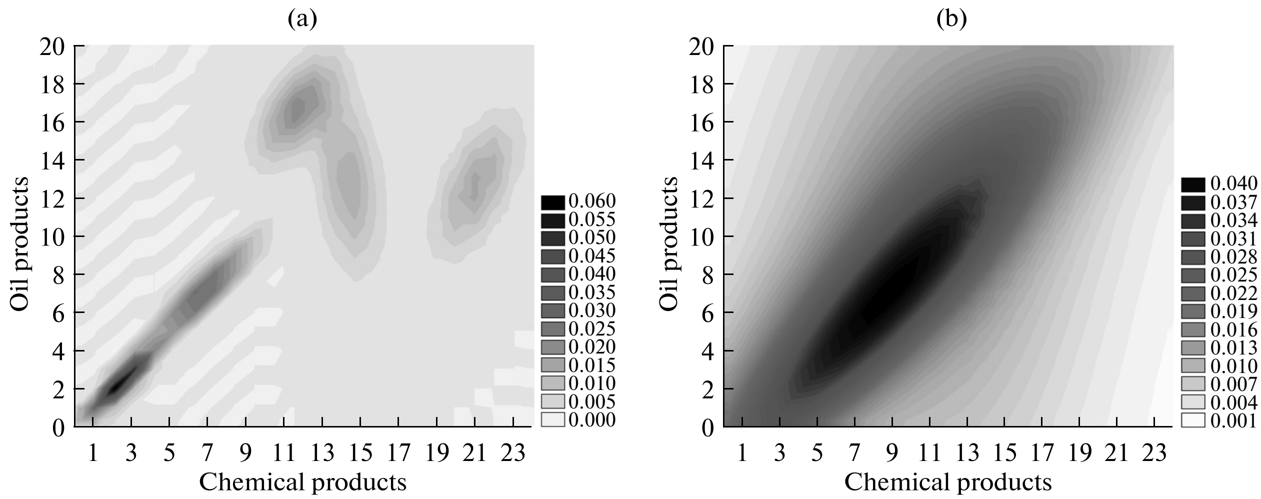


Fig. 15. Poly-Gaussian model of the PDD of factor TSs: (a) initial data; (b) quasi-optimal regression.

systematic error of the regression obtained by recurrent smoothing [30] with the help of the McLain distance weighted least squares (DWLS) model with the weight functions proposed in [31].

The systematic error is successfully reduced by increasing the number of classes in regression approximation (17). In particular, in the second stage of simulation, we took seven classes of data, corresponding to the following time intervals: January 1, 1971–January 31, 1974, February 1, 1974–November 30, 1978, December 1, 1978–June 30, 1979, July 1, 1979–August 31, 1983, September 1, 1983–January 31, 1986,

Table 4. Parameters of poly-Gaussian model of joint PDD of the volume of chemical industry and oil products

k	1	2	3	4	5
p_k	0.1905	0.2143	0.1984	0.2063	0.1905
$a_{k,1}$	3.1153	7.7041	16.8021	13.1280	13.0088
$a_{k,2}$	5.3692	9.41980	14.1693	16.7129	23.0941
$\sigma_{k,1}$	1.0211	1.5837	1.5959	2.7113	2.1229
$\sigma_{k,2}$	0.9989	1.6075	1.2430	0.9731	1.2911
$\rho_{k,1,2}$	0.9217	0.9168	0.5745	–0.2630	0.6244
$\sigma^{(k)}$	20.7316	18.4281	19.9023	19.1368	20.7316

Table 5. Quasi-optimal parameters of poly-Gaussian model of joint PDD of factor TS for $K = 5$ and Me $a_3 = 11.1565$

Parameter	k				
	1	2	3	4	5
$a_{k,1}$	5.7633	8.8597	14.2818	11.1588	8.4557
$a_{k,2}$	9.9330	10.8327	12.0439	14.2059	15.0111
$\sigma_{k,1}$	6.1267	8.7102	9.5754	14.9122	11.6757
$\sigma_{k,2}$	5.9935	8.8410	7.4578	5.3520	7.1009
$\sigma^{(k)}$	1.6552	1.4713	1.5890	1.5278	1.6552

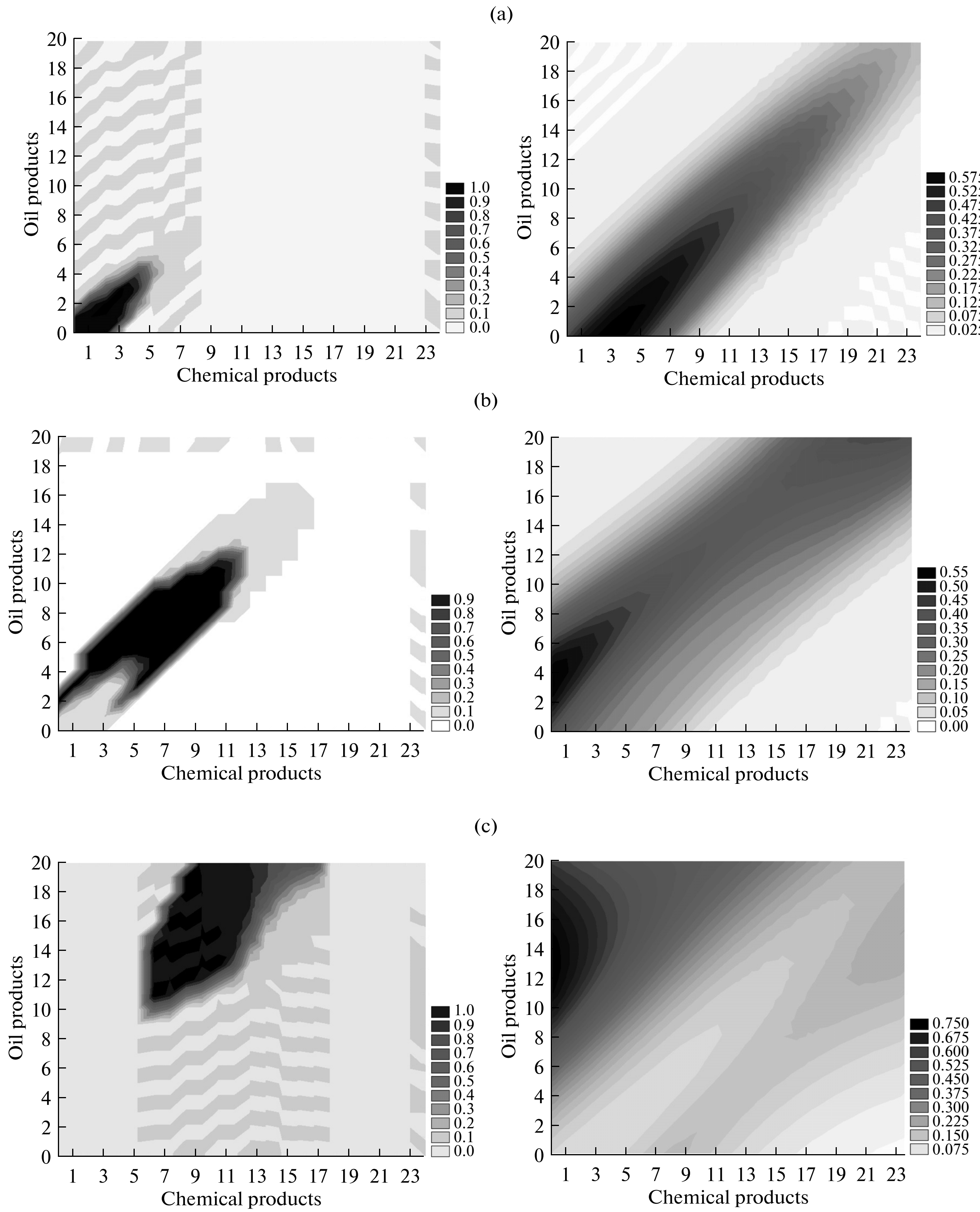
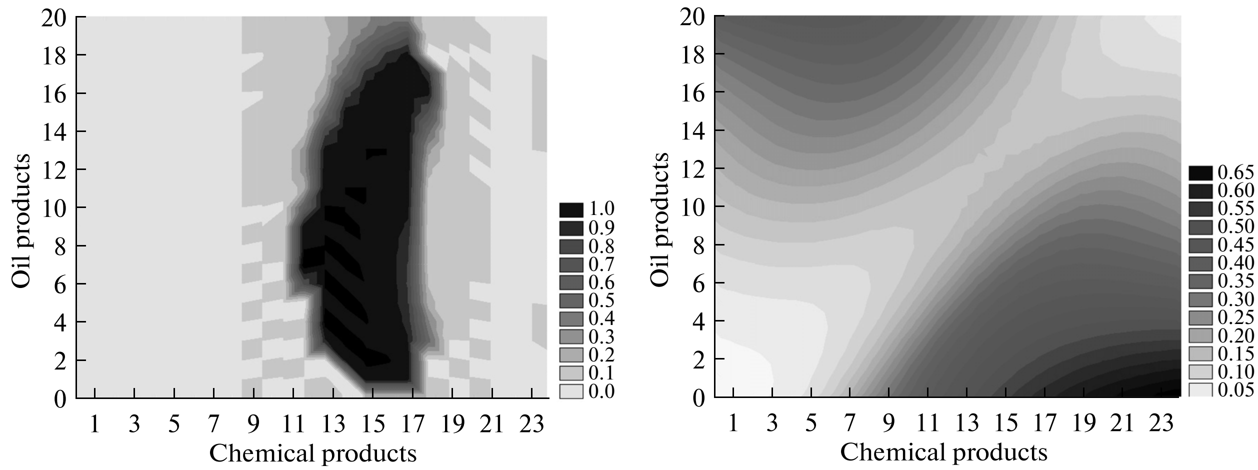


Fig. 16. A posteriori probabilities for (left) initial data and (right) quasi-optimal regression: (a) $\Pr(1|x_1, x_2)$, (b) $\Pr(2|x_1, x_2)$, (c) $\Pr(3|x_1, x_2)$, (d) $\Pr(4|x_1, x_2)$, (e) $\Pr(5|x_1, x_2)$.

(d)



(e)

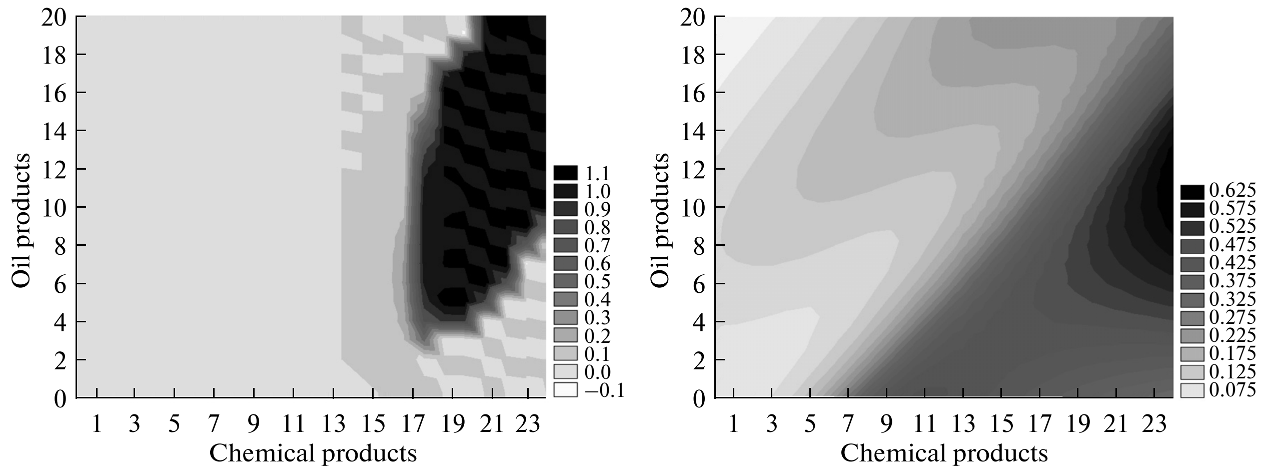


Fig. 16. (Contd.)

February 1, 1986—December 31, 1987, and January 1, 1988—December 31, 1991, with 37, 58, 7, 50, 29, 23, and 48 samples, respectively. The adjustment of the parameters of the two-dimensional poly-Gaussian approximation (6) of the PDD of factors reduces the MSD of the error $e(m)$ of the regression model more than by a factor of 20 (Table 6) from the error of the initial approximation obtained with the help of the

Table 6. Quasi-optimal parameters of poly-Gaussian model of joint PDD of factor TS for $K = 7$ and ME $a_3 = 10.5693$

Parameter	k						
	1	2	3	4	5	6	7
p_k	0.0868	0.0900	0.0278	0.4399	0.1151	0.0913	0.1505
$a_{k,1}$	3.6485	12.9298	10.6327	16.8021	10.0232	6.6546	8.4557
$a_{k,2}$	6.8875	16.3619	12.1274	14.1693	10.6582	11.1220	15.0111
$\sigma_{k,1}$	2.4546	9.4436	3.0634	7.9795	3.3802	3.6667	9.9774
$\sigma_{k,2}$	3.5605	9.4429	3.6747	6.2149	3.71693	4.5668	6.0681
$\rho_{k,1,2}$	0.7853	0.8969	0.8841	0.6345	0.1692	0.4437	0.6244
$\sigma^{(k)}$	2.1050	1.3429	11.1267	1.5577	2.6857	3.3864	1.6226

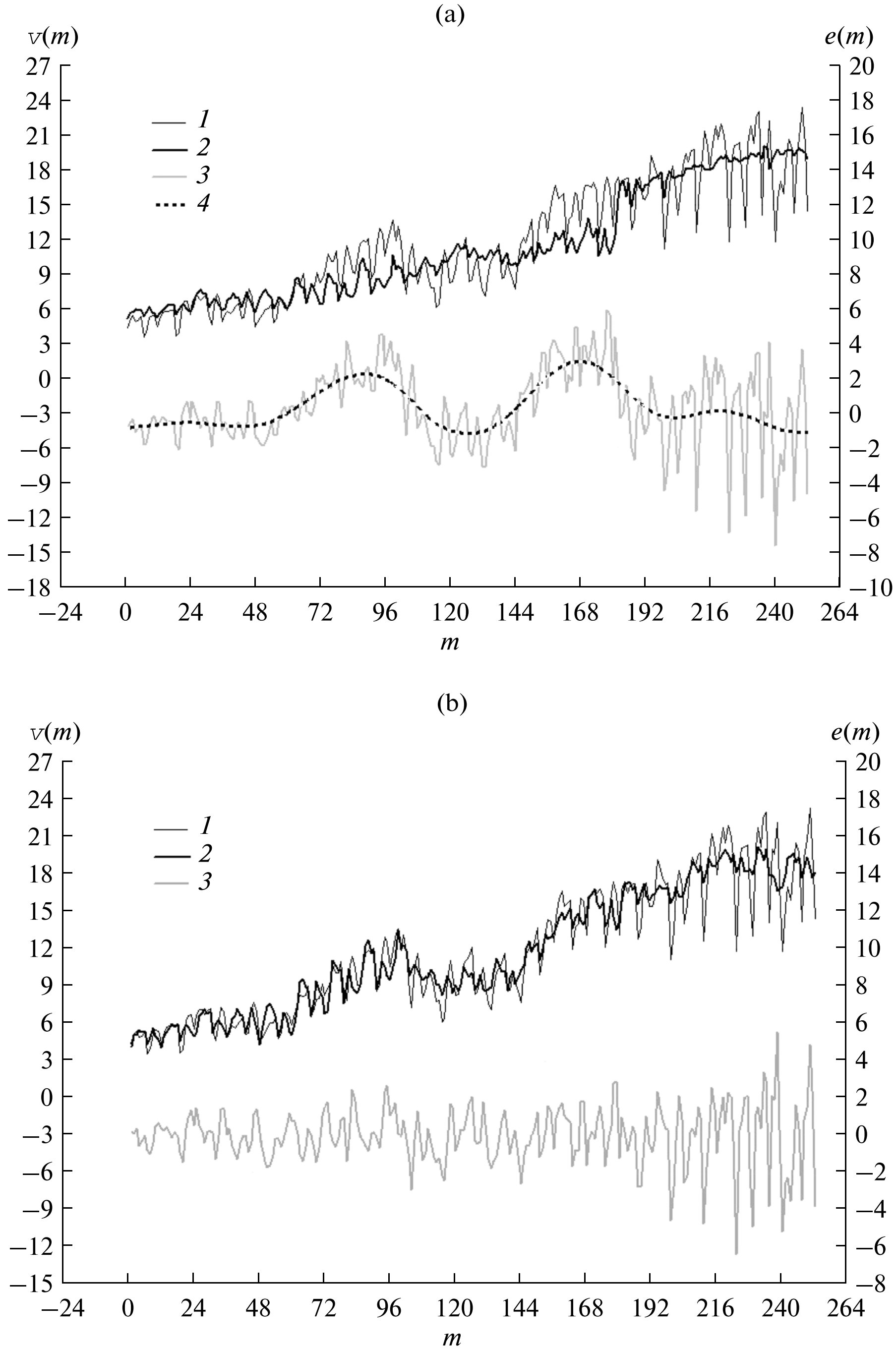


Fig. 17. Poly-Gaussian covariance model of nonlinear regression for the number of classes K = (a) 5 and (b) 7: (*1*) TS $v(m)$ of monthly sales of transport vehicles; (*2*) quasi-optimal regression; (*3*) TS of the error of regression $e(m)$; (*4*) systematic error of regression.

EM algorithm. The corresponding results of simulation of TS $v(m)$ are presented in Fig. 17b.

CONCLUSIONS

In this work, we have obtained the Bayesian approximation of the covariance model of nonlinear regression. The structure of the model is implemented by an associative computer in the form of an experts' bias network. Each expert forms a main surface that is the best in the local factor space. For the current point of this space, the gateway network estimates the confidence of the main surfaces in terms of their a posteriori probabilities. The output linear neuron of the computer aggregates the experts' opinions, forming a weighted averaged a posteriori estimate of the regression surface.

The description of the factor distribution by a finite mixture of ellipsoidally symmetric densities and application of the EM-algorithm for the identification of the parameters of partial distributions provide the possibility of adaptation of the regression model to the varying characteristics of factors. In the framework of the poly-Gaussian approximation of the PDD of factors, the main surfaces are naturally transformed to locally weighted hyperplanes.

The results of computational experiments show that, for minimizing the MSD of the error of approximation of data by the regression model, the parameters of the main surfaces obtained with the help of the EM-algorithm should be corrected to increase the effect of smoothing the factors.

REFERENCES

1. D. W. Scott and R. S. Stephan, "Multi-Dimensional Density Estimation," Handb. Stat. **24**, 229 (2005).
2. J. Racine, Kvantil', No. 4, 7 (2008).
3. E. A. Nadaraya, Teor. Veroyat. Prim. **19** (1), 147 (1964).
4. G. S. Watson, Sankhya. Indian J. Stat. **A26**, 359 (1964).
5. E. A. Nadaraya, *Nonparametric Estimation of Probability Density and Regression Curve* (Tbilis. Univ., Tbilisi, 1983) [in Russian].
6. V. G. Alekseev, Avtom. Telemekh., No. 7, 81 (1988).
7. W. S. Cleveland, J. Am. Stat. Assoc. **74** (368), 829 (1979).
8. D. W. Scott, *Multivariate Density Estimation: Theory, Practice and Visualization* (Wiley, New York, 1992).
9. S. Heiler, *A Course in Time Series Analysis*, Ed. by D. Pena, G. C. Tiao, and R. S. N. Y. Tsay (Wiley, New York, 2001).
10. K. V. Voronzov, *Lectures on Algorithms for Restoration of Regression* (2009). http://www.MachineLearning.ru/wiki/images/a/aa/Voron-ML-Regression.pdf.
11. S. A. Anatol'ev, Kvantil', No. 7, 37 (2009).
12. J. T. Tou and R. C. Gonzalez, *Pattern Recognition Principles* (Addison-Wesley, Reading, Mass., 1974; Mir, Moscow, 1978).
13. S. Haykin, *Neural Networks: A Comprehensive Foundation* (Prentice-Hall, Upper Saddle River, NJ, 1999; ID "Vil'yams", Moscow, 2006).
14. S. M. A. Bhuiyan, N. O. Atton-Okine, K. E. Barner, et al., Adv. Adaptive Data Analys **1** (2), 309 (2009).
15. L. Devroye and L. Gyorfii, *Nonparametric Density Estimation: The L1 View* (Wiley & Sons, New York, 1985; Mir, Moscow, 1988).
16. L. Breiman, W. Meisel, and E. Purcell, Technometrics **19** (2), 135 (1977).
17. D. J. Marchette, Wiley Interdisciplinary Reviews: Comput. Stat. **1** (1), 106 (2009).
18. C. E. Priebe and D. J. Marchette, Pattern Recogn. **26**, 771 (1993).
19. L. V. Labunets, J. Commun. Technol. Electron. **45**, 1093 (2000).
20. L. V. Labunets, N. L. Lebedeva, and M. Yu. Chizhov, J. Commun. Technol. Electron. **57**, 577 (2012).
21. S. A. Aivazyan, V. M. Bukhshtaber, I. S. Enyukov, and L. D. Meshalkin, *Applied Statistics: Classification and Reduction in Dimension. Reference Book* (Finansy i Statistika, Moscow, 1989) [in Russian].
22. M. H. De Groot, *Optimal Statistical Decisions* (Wiley-Interscience, 2004; Mir, Moscow, 1974).
23. K. Fukunaga, *Introduction to Statistical Pattern Recognition* (Academic, New York, 1972; Nauka, Moscow, 1979).
24. M. E. Tipping and C. M. Bishop, Neural Comput. **11**, 443 (1999).
25. A. N. Gorban, B. Kegl, D. C. Wunsch, and A. Y. Zinovyev, *Lecture Notes in Computational Science and Engineering* (Springer Verlag, Berlin, 2008).
26. G. Box and G. Jenkins, *Time Series Analysis. Forecasting and Control* (Freeman, San Francisco, 1970), Part. 2.
27. L. V. Labunets, D. S. Lukin, and A. A. Chervyakov, J. Commun. Technol. Electron. **57**, 1265 (2012).
28. L. V. Labunets, E. L. Labunets, and N. L. Lebedeva, Audit i Finans. Analiz, No. 3, 450 (2014).
29. R. J. Hyndman, Time Series Data Library. http://datamarket.com/data/list/?q=provider:tsdl).
30. L. V. Labunets, N. L. Labunets, and M. Yu. Chizhov, J. Commun. Technol. Electron. **56**, 1440 (2011).
31. L. V. Labunets, N. L. Lebedeva, and M. Yu. Chizhov, Vestn. Ros. Nov. Univ., No. 4, 66 (2013).

Translated by E. Chernokozhin