A Risk Modeling in Heterogeneous Gaussian Systems

Tyrsin A.N. [1,2 a)], Maslennikov D. L. [2, b)]

[1] *Ural Federal University, Yekaterinburg.*
[2] *South-Ural State University, Chelyabinsk*
[a)]*at2001@yandex.ru*
[b)]*asp18mdl319@susu.ru*

Abstract: A novel model of multidimensional risk is proposed. A Stochastic system of model is described as a set of independent Gaussian systems, and a fraction of each component is defined or set as probability of presence in studied population. The case when the sample is formed from a union of two Gaussian systems is considered in details. The results of testing on model and real data are presented.

1. INTRODUCTION.

The systems in real life mostly are multidimensional and their functioning is largely stochastic, and often a dozens of different risk factors can be identified. Solving the problem of risk management leads us to necessity to rely on the risk model. Typically, risk modeling comes down to identifying dangerous outcomes, quantifying the consequences of their occurrence and assessing the probabilities of these outcomes [1]. In this case, each component of the multidimensional system is considered as one-dimensional system defined as random variable and contribution of components is combined [2–4]. But the question of the mutual influence of dangerous situations caused by different elements of a multidimensional system has been little studied, most often it is neglected or significantly simplified, considering different dangerous outcomes to be mutually independent, and the likelihood of their simultaneous occurrence is neglected.

In [5], an approach to modeling the risk of multidimensional stochastic systems is proposed. It was implemented for the common case of Gaussian systems [6]. However, real objects cannot always be adequately described as Gaussian system. This is often caused by the heterogeneity of the studied sample population, which may consist of several pronounced homogeneous subsets and each of subset can be described by a Gaussian random vector. In this case, as shown by the simulation results, considering the entire sample in the form of a homogeneous Gaussian system can lead to significant errors in risk assessment. Such cases of sample population, for example, include multidimensional data in medicine and regional economics. Therefore, it is necessary to take into account the non-Gaussian nature of the multivariate data.

The purpose of the article is to describe a new model of multidimensional risk. A stochastic system of model is described as a set of several mutually independent Gaussian systems and a fraction of each component is define or set as probability of its presence in the studied population. The case when the sample under study is formed from a combination of two Gaussian systems will be considered in details.

2. MULTIDIMENSIONAL RISK MODEL IN HETEROGENEOUS SYSTEM.

Let *S* be some heterogeneous stochastic system. Each of *M* risk factor X_j is considered as a component of random vector $\mathbf{X} = (X_1, X_2, ..., X_M)$ with a certain probability density function $p_{\mathbf{X}}(\mathbf{x})$. Unlike [6], where $\mathbf{X}$ represented as a multidimensional Gaussian random variable, consider a more general case $\mathbf{X} = \bigcup_{k=1}^{K} \mathbf{X}_k$, where all random vectors $\mathbf{X}_k$ are Gaussian vectors with corresponding distributions $F_{\mathbf{X}_k}(\mathbf{x})$. The distribution function of a random vector $\mathbf{X}$ can be represented as

$$F_{\mathbf{X}}(\mathbf{x}) = \sum_{k=1}^{K} \nu_k F_{\mathbf{X}_k}(\mathbf{x}),\ \sum_{k=1}^{K} \nu_k = 1,\ \forall k\ 0 \le \nu_k \le 1. \quad (1)$$

This form (1) allows taking into account the heterogeneity of stochastic systems. Indeed, we actually have here a collection of K subsets (clusters). For example, there can be both healthy and sick people in a population; when considering regions, regions can also be divided into several groups (clusters), etc.

Instead of the common and accepted way of identification of specific dangerous situations, we will define geometric areas of unfavorable outcomes. They may be arbitrary, depending on the specific task, and are determined based on the available information. For clarity, we will describe the proposed approach using the example of the common concept of undesirable events as large and unlikely deviations of a random variable relative to its expectation. Then, dangerous situations are considered as large and unlikely deviations of the sample values x_{ij} of any of the components X_j relative to some best value $\mathbf{z}^* = (z_1^*, z_2^*, ..., z_M^*)$. Then, for a random vector $\mathbf{X}$ the probability of an unfavorable outcome can be given in the form [5]

$$P(D) = P(\mathbf{X} \in D),\ D = \left\{ \mathbf{x} = (x_1, x_2, \ldots, x_m) : \sum_{j=1}^{M} \frac{(x_j - z_j^*)^2}{B_j^2} \geq 1 \right\}. \quad (2)$$

Obviously, when the outcome does not lie on one of the axes, then event D can be realized in the absence of risk deviations in all components, (there are situations when $\mathbf{X} \in D$ and $\forall\ j\ X_j \notin D_j = (-\infty; z_j^* - B_j) \cup (z_j^* + B_j; +\infty)$).

Now we can determine a function of consequences from unfavoured outcomes as $g(\mathbf{x})$, and get a numeric approach to risk analysis

$$r(\mathbf{X}) = \int_{\mathbf{R}^m} \int \ldots \int g(\mathbf{x}) p_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}. \quad (3)$$

If $g(\mathbf{x}) = \begin{cases} 1, \mathbf{x} \in D, \\ 0, \mathbf{x} \notin D, \end{cases}$ then $r(\mathbf{X}) = P(\mathbf{X} \in D)$, and the risk is estimated as the probability of an unfavorable outcome. If it is difficult to accurately describe the function $g(\mathbf{x})$ at an early stage of studying the system, then formula (2) becomes an estimate of $P(D)$ and is a convenient initial approximation of the risk model.

3. AN APPROXIMATION OF PARAMETERS OF DISTRIBUTION OF EACH CLUSTER OF HETEROGENEOUS SYSTEM.

One of the problems of using model (1) is to determine the parameters of the distributions $F_{\mathbf{X}_k}(\mathbf{x})$ and probabilities ν_k. Consider solutions for the example of two clusters. Note that it can be used for three and more clusters.

First we can use discriminant analysis. To implement the parameter search algorithm, it is necessary to know the probability that the next observation belongs to a particular cluster. Discriminant analysis based on logistic regression [7] provides these probabilities. The algorithm for evaluating the parameters of the distributions of clusters is as follows.

Input: Data – an array of coordinates of data samples; probas – probabilities of belonging point to each cluster (a length of probas equal to length of data), N – a number of experiments, K – a number of clusters.

Output: A parameters of distribution for each cluster.

1. Calculate a cumulative sum of probas. Set counter variable equals to 0.
2. Generate a random variable $p \in [0; 1)$.
3. Taking into account value p from step 2 and consider corresponding cumulative probas for each point of data define a cluster number of point in current experiment. It can be defined as index of element of cumulative probas when $\hat{p} \geq p$, where $\hat{p}$ is corresponding value from probas array.
4. Collecting information of step 3 allows us to get clusters of current experiment.
5. Now we can calculate estimate of expectation and covariance for each cluster. Increment counter variable.
6. While counter < N repeat steps 2-5, else calculate estimates of expectations and covariances of whole experiments aggregating result of each separate experiment. These final estimates will be an approximate estimates for expectations and covariances.

Note that this algorithm can be applied with 3 and more cluster, since we will use it in our work with systems of two clusters.

Secondly we can use a Gaussian Mixture Models (GMM) [8]. GMM is a one of implementations of Expectation Maximization algorithm [9].

EM algorithm is a general method for finding estimates of the likelihood function in models with hidden variables. This article discusses the interpretation of a mixture of Gaussian distributions in terms of discrete hidden variables.

In addition to the fact that mixture models allow approximating complex probability distributions, they can also be used to solve the problem of data clustering. Next, we will solve the clustering problem using the EM-algorithm, having previously approximated the solution with the *k*-means algorithm [10].

A comparative analysis on model samples of the estimates of the parameters of the distribution of clusters showed that both algorithms (logistic regression and GMM) have close results.

4. A RISK ESTIMATE IN HETEROGENEOUS STOCHASTIC SYSTEMS WITH TWO CLUSTERS.

Consider 5 different cases when the system contains of $M = 2$ components: 1 – each component has a diagonal covariance matrix; 2 – "stretched" scattering; 3 – each component consists of correlated parameters. 4 – components of each of the set are dependent, but there are significant intersections between clusters; 5 – Clearly separate clusters. The experimental technique was as follows. We generate data for model cases with given parameters - known covariance and expectations.

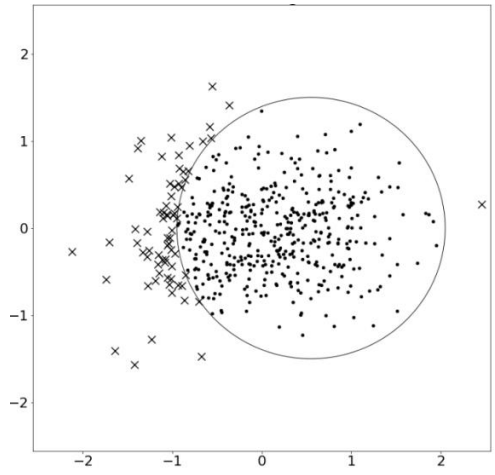
In all of our cases, the proportions of populations are the same, so, the probability of random observation belongs to any cluster is 0.5. For each of the cases, the parameters of the distributions of each population are known in advance - expectations and covariance (due to the fact that we consider each of the populations to be represented as a Gaussian random vector). In each case, one of the populations will be considered as "favorable", and the boundaries of the region will be defined by an ellipse.

We will make the experiment as follows:

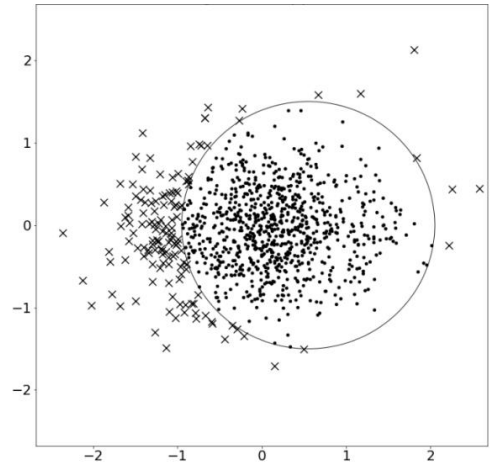
1. For each case, using the known parameters, we will generate 30 pairs of populations.
2. For each case and each of the 30 final samples, we will try to recover the parameters in two ways - using the algorithm proposed in Section 3, and also using GMM.
3. Based on the restored parameters, generate a sufficient number of observations for each population.
4. Using the Monte-Carlo modeling method, calculate the probability of missing a favorable area. This value will be an estimate of the probability of risk in the system.

Since the model examples have real information about the initial parameters of the distributions of each population, it is possible to calculate the real risk with high accuracy. We will also give the calculation of the risk of a given system, presented in the form of a homogeneous Gaussian system [5], and making the calculations with the restored parameters using logistic regression and GMM.

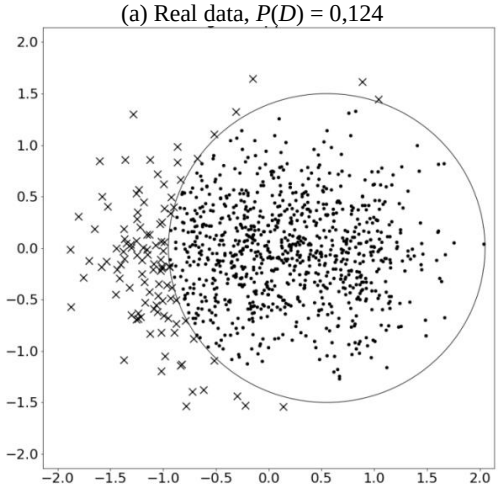
In the figures below, the favorable area is marked with an ellipse, the favorable points are marked with a dots, and the unfavorable points are marked with a cross. Risk probability is the ratio of the number of crosses to the total number of observations.



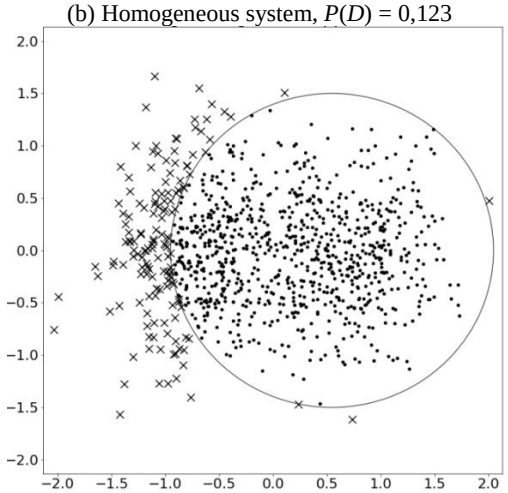
(a) Real data, $P(D) = 0{,}124$



(b) Homogeneous system, $P(D) = 0{,}123$

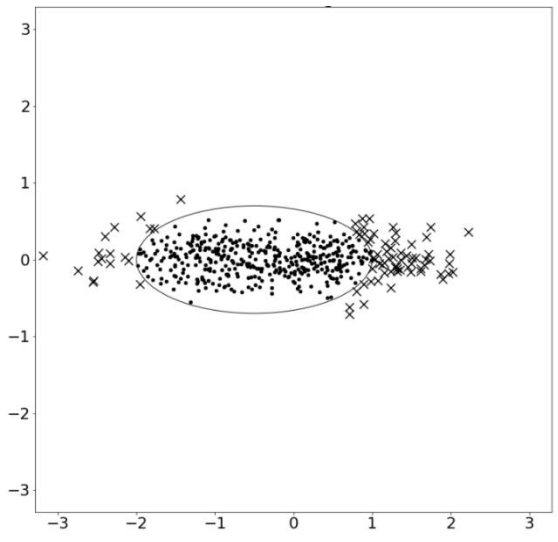


(c) Heterogeneous system. GMM, $P(D) = 0{,}125$

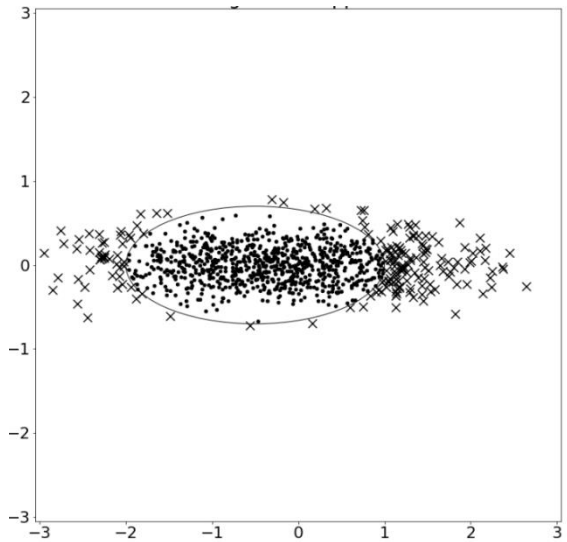


(d) Heterogeneous system. Logistic Regression, $P(D) = 0{,}128$

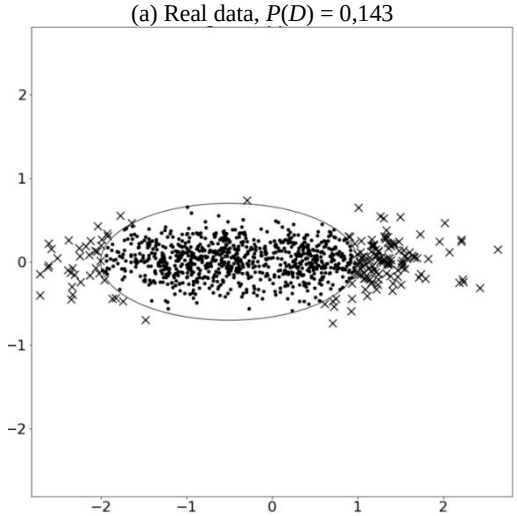
FIGURE 1. Case 1. Each component has a diagonal covariance matrix.



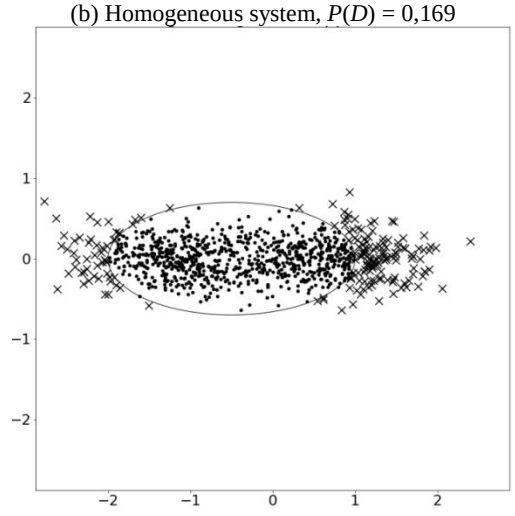
(a) Real data, $P(D) = 0{,}143$



(b) Homogeneous system, $P(D) = 0{,}169$

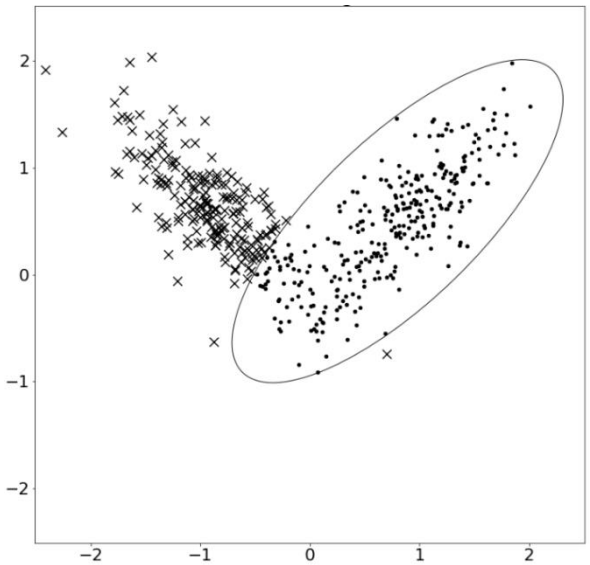


(c) Heterogeneous system. GMM, $P(D) = 0{,}147$

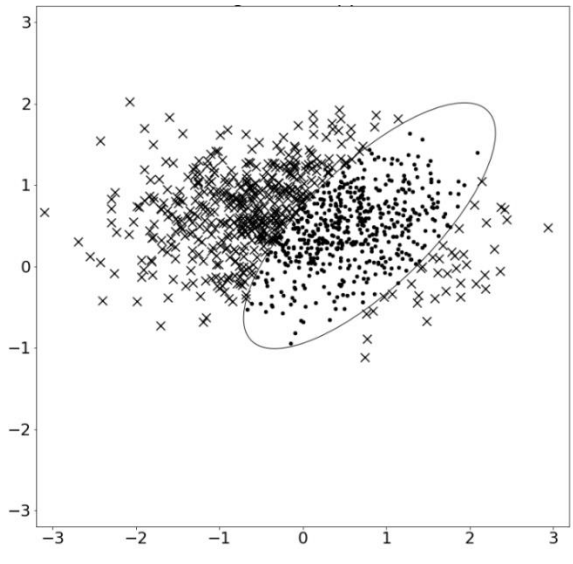


(d) Heterogeneous system. Logistic Regression, $P(D) = 0{,}164$

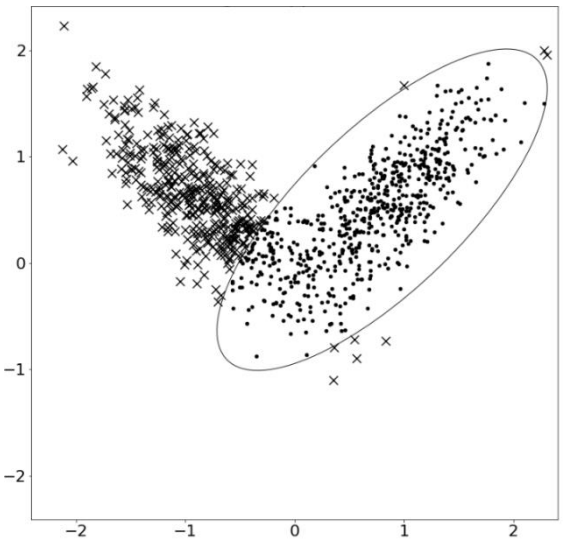
FIGURE 2. Case 2. «Stretched» scatter.



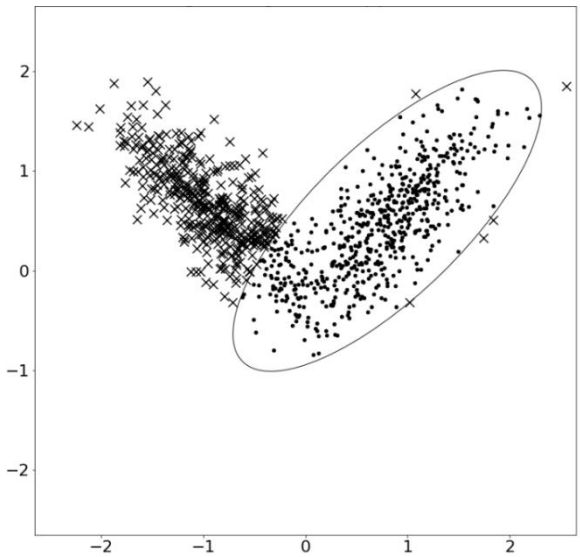
(a) Real data, $P(D) = 0{,}415$



(b) Homogeneous system, $P(D) = 0{,}492$

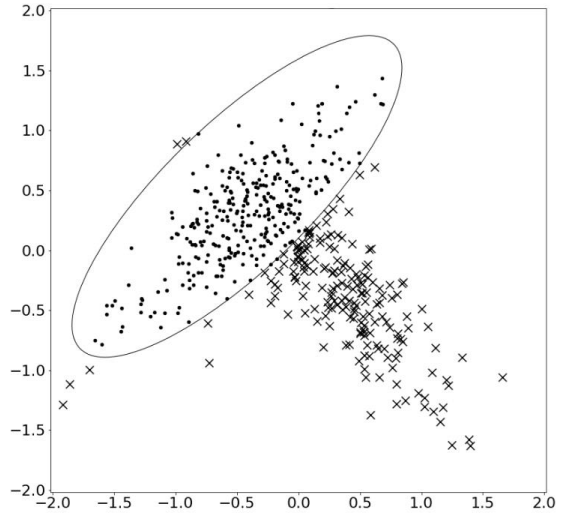


(c) Heterogeneous system. GMM, $P(D) = 0{,}420$

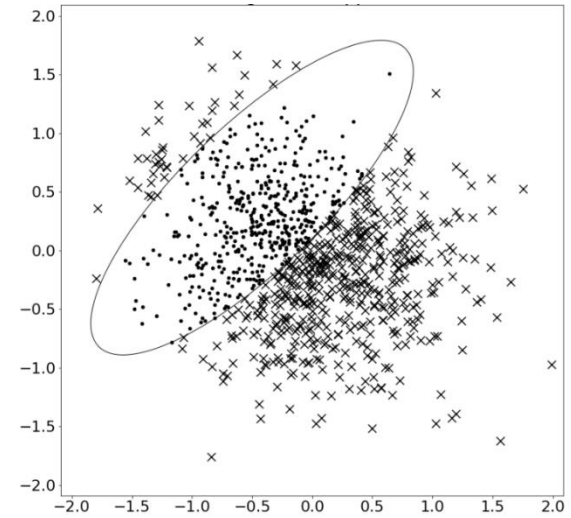


(d) Heterogeneous system. Logistic Regression, $P(D) = 0{,}407$

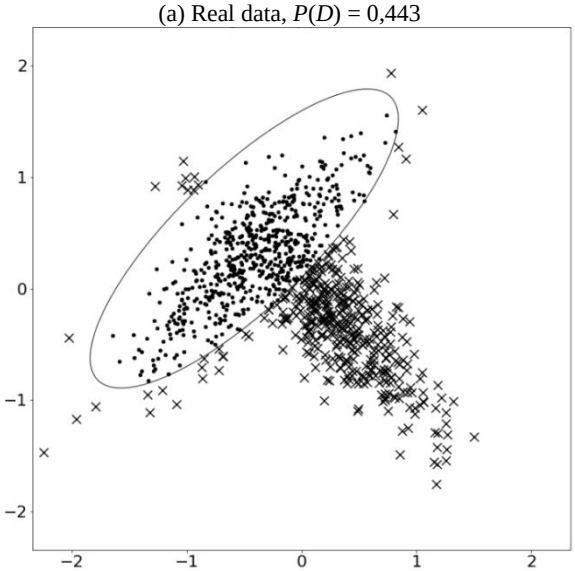
FIGURE 3. Case 3. each component consists of correlated parameters.



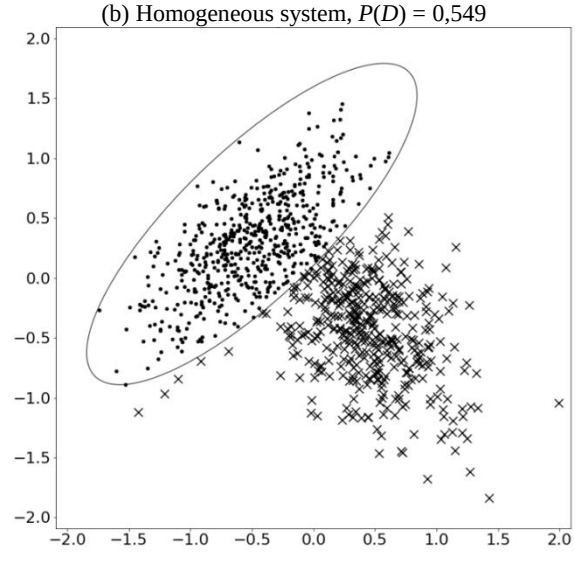
(a) Real data, $P(D) = 0{,}443$



(b) Homogeneous system, $P(D) = 0{,}549$

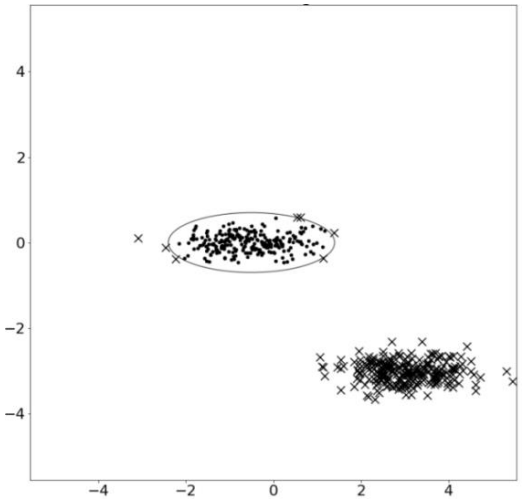


(c) Heterogeneous system. GMM, $P(D) = 0{,}382$

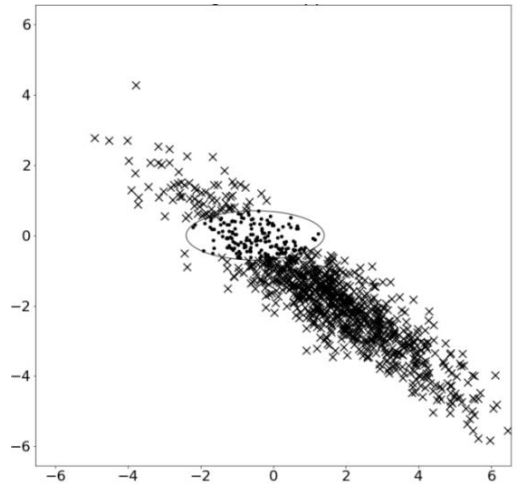


(d) Heterogeneous system. Logistic Regression, $P(D) = 0{,}456$

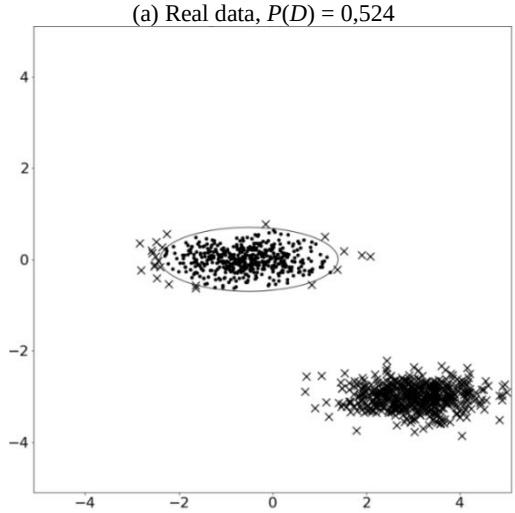
FIGURE 4. Case 4. components of each of the set are dependent, but there are significant intersections between clusters.



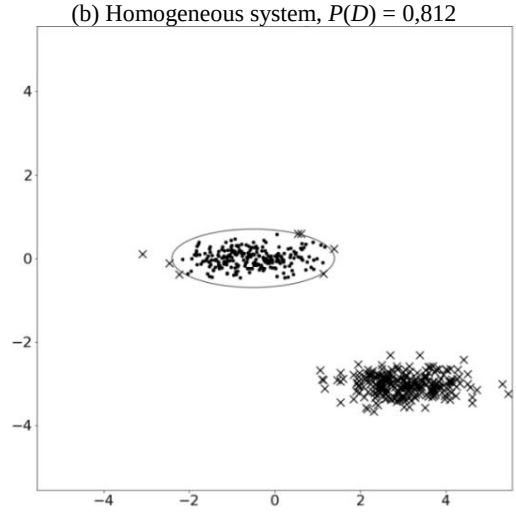
(a) Real data, $P(D) = 0{,}524$



(b) Homogeneous system, $P(D) = 0{,}812$



(c) Heterogeneous system. GMM, $P(D) = 0{,}543$

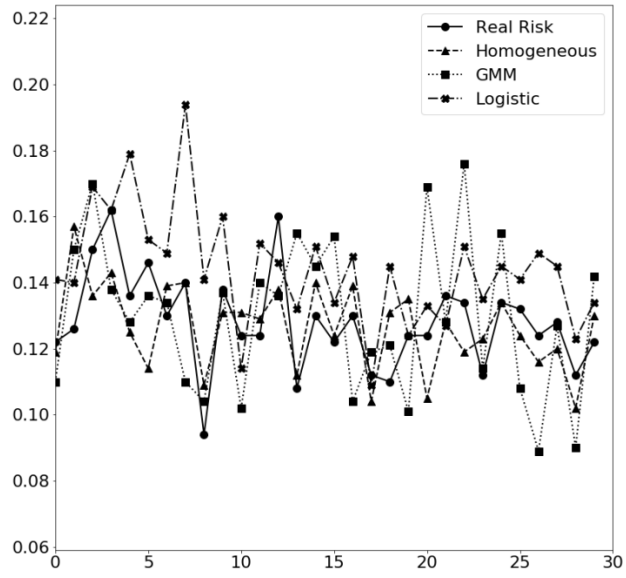


(d) Heterogeneous system. Logistic Regression, $P(D) = 0{,}529$

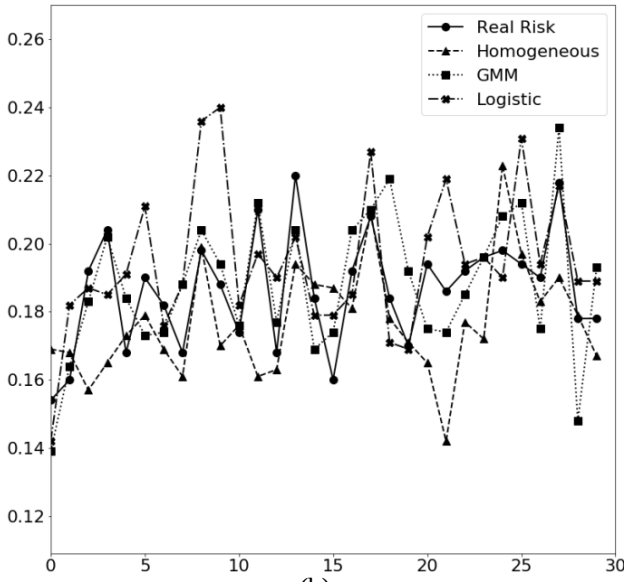
FIGURE 5. Case 5. Clearly separate clusters.

Analysis of these examples shows that in four cases out of five, the use of the model in the form of a homogeneous Gaussian system led to a significant overestimation of the risk estimate.

It makes a sense to consider risk estimates of whole batch of experiments. The results are shown in the figures below. Here, horizontal axis is the indices of the experiments and the vertical axis is the calculated risk probabilities.



(a)



(b)

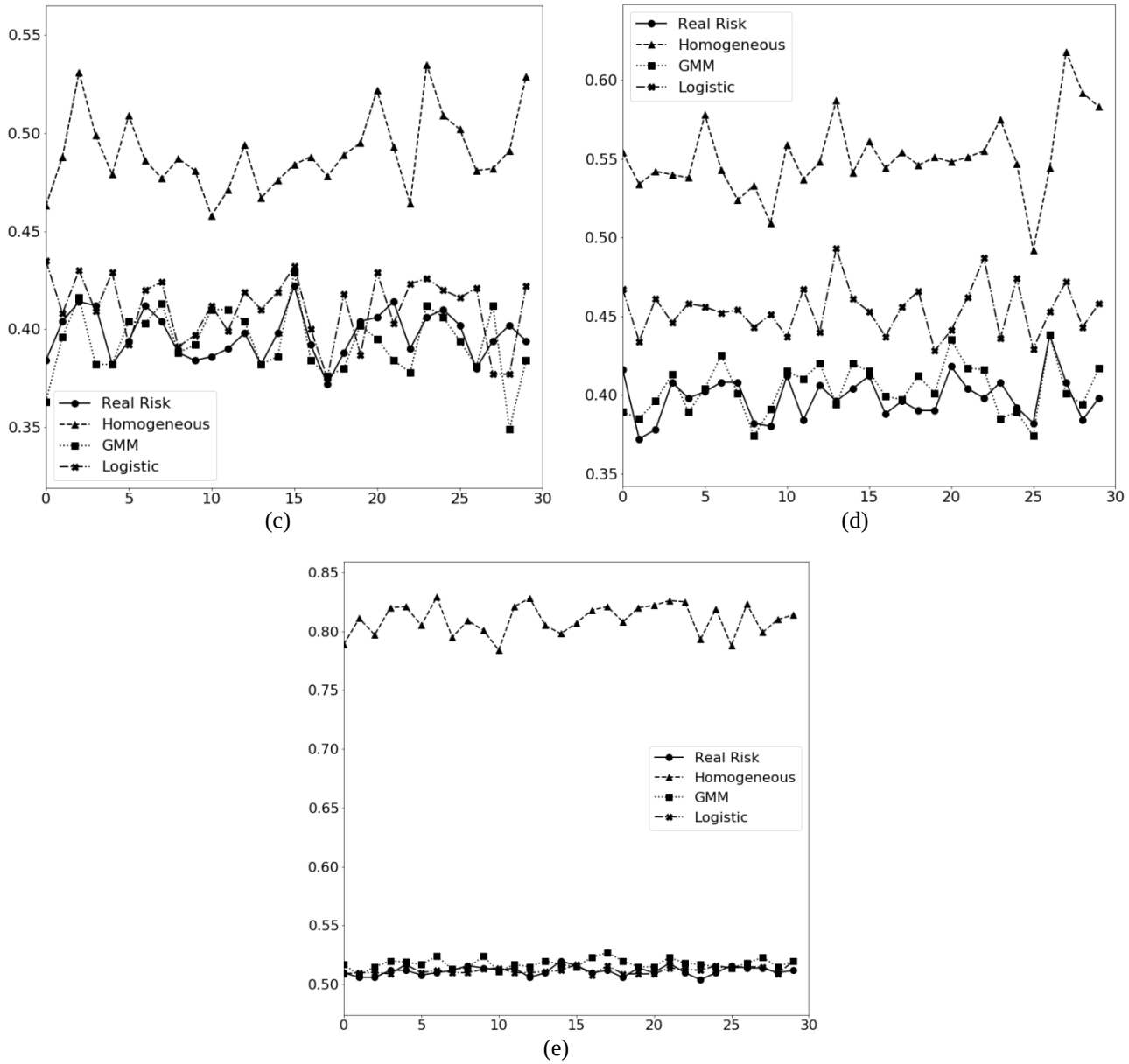


FIGURE 6. The results of risk estimates for all experiments of case 1 (a), case 2 (b), case 3 (c), case 4 (d) and case 5 (e).

Analysis of these figure shows that for the first two cases ((a), (b) of fig. 6) there is no higher accuracy of the risk estimates, but the results are close enough. In other cases ((c)-(e) of fig. 6) a model for heterogeneous system is more preferable, since the risks estimates are close to real, while a homogeneous model significantly overestimates these values.

5. CONCLUSION.

1. A multidimensional risk model for heterogeneous stochastic systems is proposed. In this case, the stochastic system is represented as a set of Gaussian multidimensional random variables. It is shown that in the case of heterogeneous stochastic systems, their modeling in the form of a single Gaussian multidimensional random variable can lead to significant errors in risk assessment.

2. An algorithm for restoring distribution parameters based on the probabilities of membership of each point is proposed and considered. For this article we used logistic regression. However, other classification methods can be used.

3. Comparison of logistic regression with GMM showed the following:

GMM has less stability in calculating risks than logistic regression, so it is preferable to use parameter search using logistic regression. But this way is computationally more expensive and requires some information about the initial populations in the form of a training sample.

The work was supported by the Russian Foundation for Basic Research (project no. 20-51-00001).

REFERENCES

1. J. Mun, *Modeling Risk. 2nd Edition*. (Wiley, 2010).
2. C. Rossi, *Fundamentals of Risk Management*. (Wiley, 2014).
3. Cox L.A., Jr. *Improving Risk Analysis*. (Springer, 2013).
4. V.A. Akimov, V.V. Lesnykh, N. N. Radaev. Risks in nature, technosphere, society and economy. (Moscow, Delovoy Express, 2004). p. 352
5. Tyrsin A. N., Surina A. A., "A risk modeling in multidimensional stochastic systems," Tomsk State University Bulletin. Journal of Control and computer science. № 2(39), 65-72 (2017).
6. Tyrsin A. N., Surina A. A., "A risk management models in Gaussian stochastic systems," Informatics and applications. 12(2), 50-59, (2018).
7. Hosmer D.W., Lemeshow S., Sturdivant R.X. *Applied Logistic Regression. 3rd Edition*. (Wiley, 2013).
8. Dempster A., Laird N., Rubin D. "Maximum likelihood from incomplete data via the EM algorithm," Journal of the Royal Statistical Society, Series B, 39(1). pp. 1–38 (1977).
9. Jordan M.I., Xu L. *Convergence results for the EM algorithm to mixtures of experts architectures* (Tech. Rep. A.I. Memo №1458, MIT, MA, 1993).
10. N. G. Zagoruyko. *An applied methods of analysis data and knowledge*. (Novosibirsk, Institute of mathematics press, 1999). p. 270.