УДК 519.711.3:519.2/.6

МОДЕЛИРОВАНИЕ РИСКА В НЕОДНОРОДНЫХ СТОХАСТИЧЕСКИХ СИСТЕМАХ

[1,3]Тырсин А.Н., [2]Масленников Д.Л.

[1]ФГАОУ ВО «Уральский федеральный университет имени первого Президента России Б.Н. Ельцина», Екатеринбург, e-mail: at2001@yandex.ru;

[2]ФГАОУ ВО «Южно-Уральский государственный университет» (национальный исследовательский университет), Челябинск, e-mail: asp18mdl319@susu.ru;

[3]ФГБУН «Научно-инженерный центр «Надежность и ресурс больших систем и машин» УрО РАН, Екатеринбург, e-mail: at2001@yandex.ru

Реальные объекты не всегда удается адекватно описать в виде гауссовских систем. Это часто вызвано неоднородностью исследуемой выборочной совокупности, в которой могут присутствовать несколько выраженных однородных кластеров, каждый из которых может быть описан гауссовским случайным вектором. В этом случае, как показали результаты моделирования, рассмотрение всей выборки в виде однородной гауссовской системы приводит к значительным ошибкам при оценивании риска. К таким совокупностям, например, можно отнести многомерные данные в медицине и региональной экономике. Поэтому необходимо в модели риска учесть негауссовость рассматриваемых многомерных данных. Рассмотрены два алгоритма для оценивания параметров распределений кластеров неоднородных стохастических систем. Первый алгоритм использует известную модель смеси гауссиан. Второй алгоритм основан на дискриминантном анализе. Предложена новая модель многомерного риска, в которой стохастическая система описывается в виде совокупности нескольких взаимно независимых гауссовских систем, для каждой из которых определяется или задается вероятность (доля) ее присутствия в исследуемой совокупности. Рассмотрен случай, когда исследуемая выборка состоит из совокупности двух гауссовских систем. Приведены результаты апробации на модельных данных.

Ключевые слова: многомерный риск, модель, стохастическая система, случайный вектор, кластер

A RISK MODELING IN HETEROGENEOUS STOCHASTIC SYSTEMS

[1,3]Tyrsin A.N., [2]Maslennikov D.L.

[1]Federal State Autonomous Educational Institution of Higher Education Ural Federal University named after the first President of Russia B.N. Yeltsin, Yekaterinburg, e-mail: at2001@yandex.ru;

[2]Federal State Autonomous Educational Institution of Higher Education South-Ural State University (National Research University), Chelyabinsk, e-mail: asp18mdl319@susu.ru;

[3]Federal State Budgetary Institution of Science Scientific-Engineering Center Reliability and Life of Large Systems and Machines, Ural Branch, Russian Academy of Science, Yekaterinburg, e-mail: at2001@yandex.ru

Real objects cannot always be adequately described as Gaussian systems. 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. Two algorithms for estimating parameters of cluster distributions of heterogeneous stochastic systems are considered. The first algorithm uses the well-known Gaussian mixture model. The second algorithm is based on discriminant analysis. 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.

Keywords: multidimensional risk, model, stochastic system, random vector, cluster

Моделирование риска обычно включает в себя выделение опасных исходов, количественное задание последствий от их наступления и оценку их вероятностей [1]. Вклад различных рисков объединяют и рассматривают одномерную случайную величину [2–4]. Однако у такой модели есть недостатки. Во-первых, не всегда известны заранее все возможные опасные исходы и, соответственно, мы не знаем их вероятности. Во-вторых, реальные системы обычно имеют много различных факторов риска. А поскольку они могут быть взаимосвязанными, то необходимо факторы риска рассматривать совместно, т.е. возникает потребность в многомерном моделировании риска.

В [5] предложен подход к моделированию многомерного риска. Он был реализован для распространенного случая гауссовских систем [6]. Однако реальные объекты

(например, в медицине и региональной экономике) не всегда удается адекватно описать в виде гауссовских систем. Это часто вызвано неоднородностью исследуемой выборочной совокупности, в которой могут присутствовать несколько выраженных однородных подмножеств, каждое из которых может быть описано гауссовским случайным вектором. В этом случае, как показали результаты моделирования, рассмотрение всей выборки в виде однородной гауссовской системы может приводить к значительным ошибкам при оценивании риска. Поэтому необходимо в модели риска учесть негауссовость системы.

Целью статьи является описание новой модели многомерного риска, в которой стохастическая система описывается в виде совокупности нескольких взаимно независимых гауссовских систем, для каждой из которых определяется или задается вероятность (доля) ее присутствия в исследуемой совокупности. Будет подробно рассмотрен случай, когда исследуемая выборка формируется из совокупности двух гауссовских систем.

Материалы и методы исследования

Рассмотрим неоднородную стохастическую систему S. Выделим в ней M факторов риска X_j и будем считать их компонентами случайного вектора $\mathbf{X} = (X_1, X_2, ..., X_M)$. В отличие от [6], когда $\mathbf{X}$ представлял собой гауссовский случайный вектор, рассмотрим более общий случай $\mathbf{X} = \bigcup_{k=1}^{K} \mathbf{X}_k$, где все случайные векторы $\mathbf{X}_k$ являются гауссовскими с соответствующими распределениями $F_{\mathbf{X}_k}(\mathbf{x})$. Функция распределения случайного вектора $\mathbf{X}$ может быть представлена как

$$F_{\mathbf{X}}(\mathbf{x}) = \sum\nolimits_{k=1}^{K} \nu_k F_{\mathbf{X}_k}(\mathbf{x}),$$

$$\sum\nolimits_{k=1}^{K} \nu_k = 1, \ \forall k \ \ 0 \le \nu_k \le 1. \qquad (1)$$

Представление (1) позволяет учесть неоднородность стохастических систем. Действительно, мы фактически здесь имеем совокупность K подмножеств (кластеров). Например, в популяции могут быть как здоровые, так и больные люди, при рассмотрении регионы также могут делиться на несколько групп (кластеров) и т.д.

Неблагоприятные исходы описываем в виде геометрической области, вид которой определяется конкретной системой, решаемой задачей и априорных сведений. Для определенности опишем предлагаемый подход на примере распространенной концепции нежелательных событий как больших и маловероятных значений случайного вектора $\mathbf{X}$. Тогда вероятность неблагоприятного исхода для случайного вектора $\mathbf{X}$ равна [5]

$$P(D) = P(\mathbf{X} \in D),$$

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

где $\mathbf{z}^* = (z_1^*, z_2^*, ..., z_M^*)$ – некоторое наиболее безопасное значение, B_j – пороговые уровни.

Очевидно, когда исход не лежит на одной из осей, то событие D может реализоваться и при отсутствии рисковых отклонений по всем компонентам (например, возможны ситуации, когда $\mathbf{X} \in D$ и $\forall j \ X_j \notin D_j = (-\infty; z_j^* - B_j) \cup (z_j^* + B_j; +\infty)$). Риск оценим как [5]

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

где $p_{\mathbf{X}}(\mathbf{x})$ – плотность вероятности случайного вектора $\mathbf{X}$, $g(\mathbf{x})$ – функция последствий от опасных ситуаций. Задав

$$g(\mathbf{x}) = \begin{cases} 1, \mathbf{x} \in D, \\ 0, \mathbf{x} \notin D, \end{cases} \qquad (3)$$

получим $r(\mathbf{X}) = P(\mathbf{X} \in D)$, т.е. риск будет равен вероятности возникновения неблагоприятного исхода. Оценка (3) хорошо интерпретируется, а также является удобным начальным приближением модели риска на ранней стадии исследования системы, когда сложно достаточно точно описать функцию $g(\mathbf{x})$.

Результаты исследования и их обсуждение

Одной из проблем при использовании модели (1) является определение параметров распределений $F_{\mathbf{X}_k}(\mathbf{x})$ и вероятностей ν_k. Ниже рассмотрим варианты решения на примере двух кластеров. Во-первых, можно воспользоваться дискриминантным анализом. Для реализации алгоритма поиска параметров необходимо знать вероятность принадлежности очередного наблюдения тому или иному кластеру. Дискриминантный анализ на основе логистической регрессии [7] позволяет оценить эти вероятности. Алгоритм разыгрывания для оценки параметров распределений кластеров состоит в следующем.

Вход: *data* – массив координат входных объектов; *probas* – вероятности при-

надлежности к кластерам каждой точки (длина массива равна длине массива *data*), N – количество розыгрышей, K – количество кластеров.

Выход: параметры распределений каждого кластера.

Шаг 1. Вычисляем кумулятивную сумму для векторов вероятностей в кластерах для каждой точки, установить счетчик равным 0.

Шаг 2. Разыгрываем случайную величину p, равномерно распределенную на [0; 1).

Шаг 3. Для всех точек определяем по кумулятивному вектору вероятностей, в какой кластер соотносится точка. Номер кластера будет определяться индексом минимального значения элементов вектора $\hat{p} \geq p$.

Шаг 4. Выделяем из всего набора данных кластеры, полученные на шаге 3.

Шаг 5. Вычисляем оценки математических ожиданий и дисперсий для каждого кластера и записываем результаты, увеличиваем счетчик на 1.

Шаг 6. Если счетчик меньше N повторяем шаги 2–5, иначе вычисляем оценки математических ожиданий и дисперсий по N опытам. Усредненные по всем опытам оценки будут результатом оценки параметров.

Стоит отметить, что описанный алгоритм является частным случаем для двух кластеров и его можно расширить для любого конечного числа кластеров, вместо разыгрывания одной случайной величины p будем разыгрывать вектор вероятностей $\mathbf{p}$, состоящий из $K - 1$ элементов. Тогда попадание на основе этих значений наблюдений в тот или иной кластер будет описываться мультиномиальным распределением. Остальные шаги процедуры останутся неизменными.

Также можно использовать модель смеси гауссиан (GMM) [8]. GMM реализуется с помощью решения оптимизационной задачи максимального правдоподобия (Expectation Maximization) [9]. EM-алгоритм – общий метод нахождения оценок функции правдоподобия в моделях со скрытыми переменными. В данной статье рассматривается интерпретация смеси гауссовых распределений в терминах дискретных скрытых переменных.

Помимо того что смеси распределений позволяют приближать сложные вероятностные распределения, с их помощью можно также решать задачу кластеризации данных. Далее мы будем решать задачу кластеризации с помощью EM-алгоритма, предварительно приблизив решение алгоритмом k-средних [10].

Сравнительный анализ оценок параметров распределения кластеров показал, что оба алгоритма (на основе логистической регрессии и GMM) дали сравнимые по точности результаты.

Проведем расчет риска неоднородных систем с двумя кластерами. Рассмотрим пять модельных случаев систем из двух совокупностей, число компонент $M = 2$: 1 – каждая компонента имеет диагональную ковариационную матрицу; 2 – «вытянутые» рассеяния; 3 – каждая компонента состоит из коррелированных параметров; 4 – компоненты каждой из совокупности зависимы, между ними значимые пересечения; 5 – «непересекающиеся» классы. Методика эксперимента состояла в следующем.

Генерируем данные для модельных случаев с заданными параметрами – известными ковариациями и математическими ожиданиями. На всех модельных примерах доли совокупностей одинаковы, то есть вероятность попадания случайного наблюдения равна 0,5 для одного случая. Посмотрим, как эти доли могут влиять на значения риска. Для каждого из случаев заранее известны параметры распределений каждой совокупности – векторы средних и ковариации (в силу того что считаем каждую из совокупностей представимой в виде гауссовского случайного вектора). В каждом из модельных примеров одна из совокупностей будет считаться «благоприятной», а границы области будут определяться эллипсом. Будем проводить эксперимент следующим образом:

1) для каждого случая по известным параметрам сгенерируем по 30 пар совокупностей;

2) для каждого случая и каждой из 30 итоговых выборок попытаемся восстановить параметры двумя способами – с помощью алгоритма на основе логистической регрессии, а также с помощью EM-алгоритма (GMM-модель);

3) на основе восстановленных параметров сгенерировать достаточное количество наблюдений каждой совокупности;

4) методом Монте-Карло выполнить расчет вероятности непопадания в благоприятную область. Это и будет вероятностью риска в системе.

Поскольку на модельных примерах имеется информация об исходных параметрах распределений каждой совокупности, можно с высокой точностью оценить реальный риск. Также приведем расчет риска заданной системы, представленной в виде однородной гауссовской системы [5], и проведем расчет с восстановленными логистической регрессией и GMM параметрами.

На рис. 1–5 благоприятная область отмечена эллипсом, благоприятные точки – точками, а неблагоприятные точки – крестиком. Вероятность риска – это отношение числа пересечений к общему числу наблюдений.

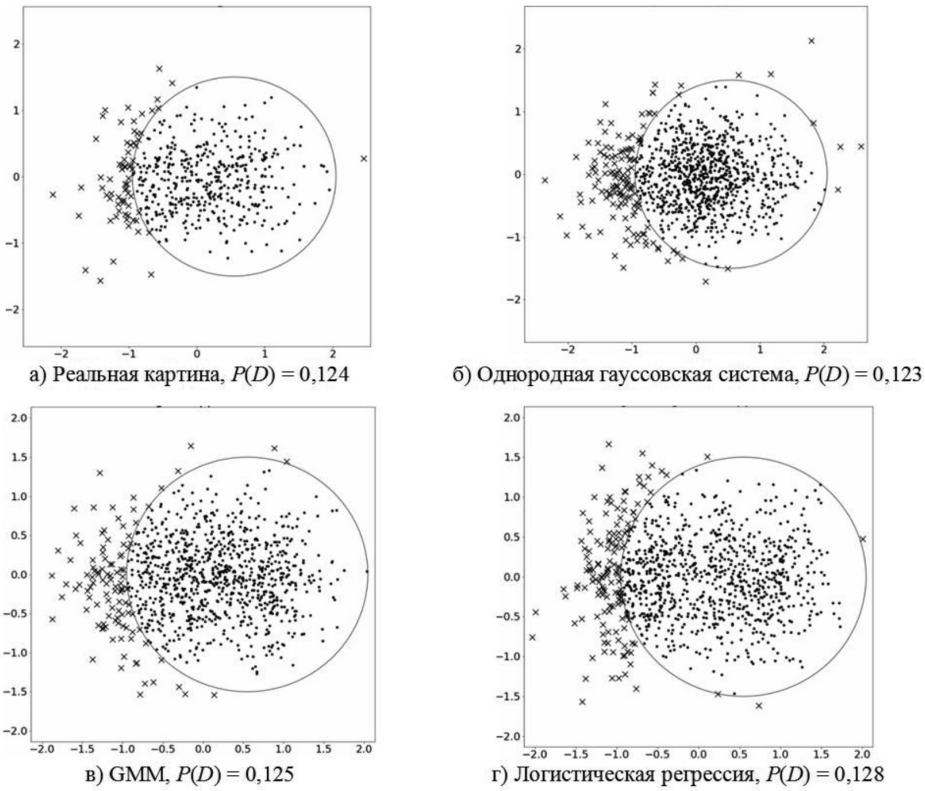

Рис. 1. Каждая компонента имеет диагональную ковариационную матрицу

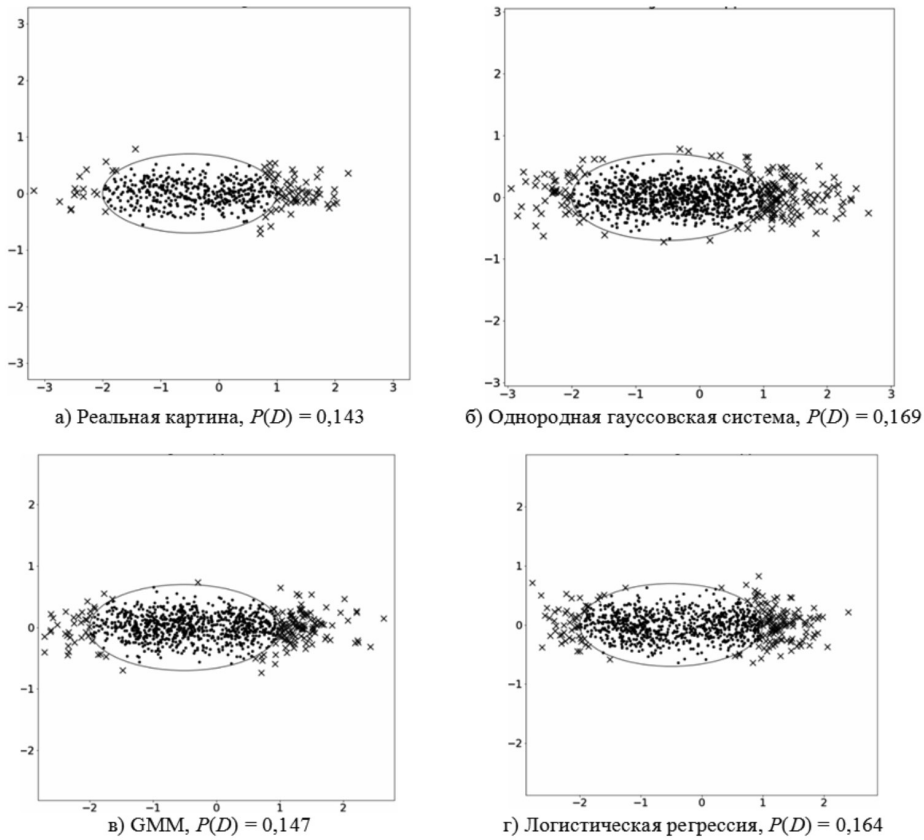

Рис. 2. «Вытянутые» рассеяния

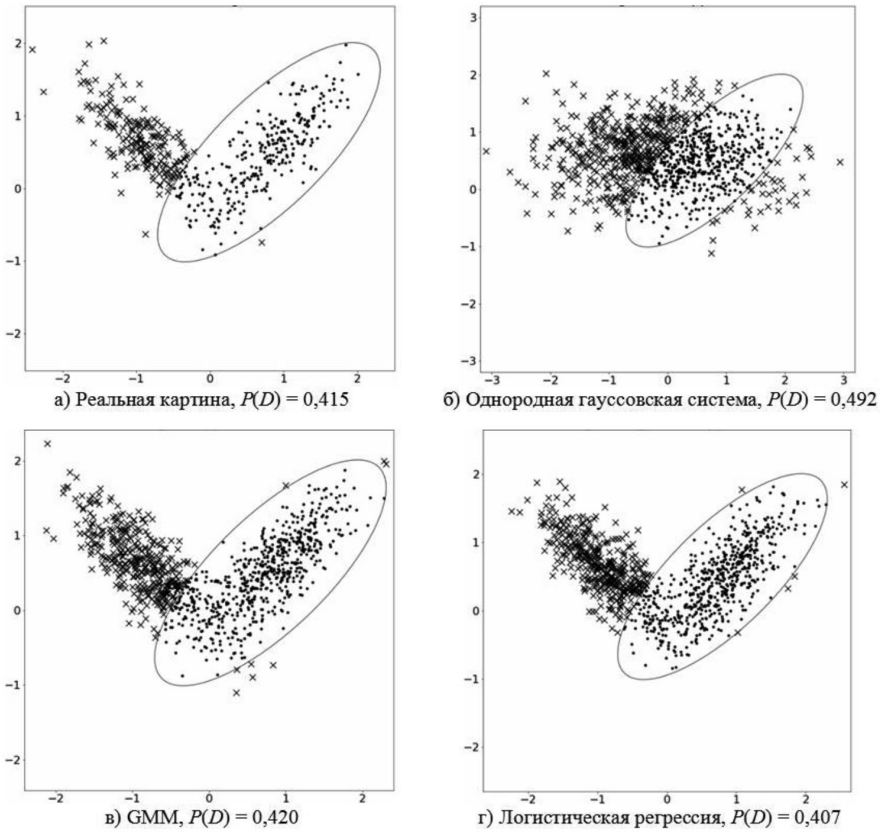

Рис. 3. Каждая компонента состоит из коррелированных параметров

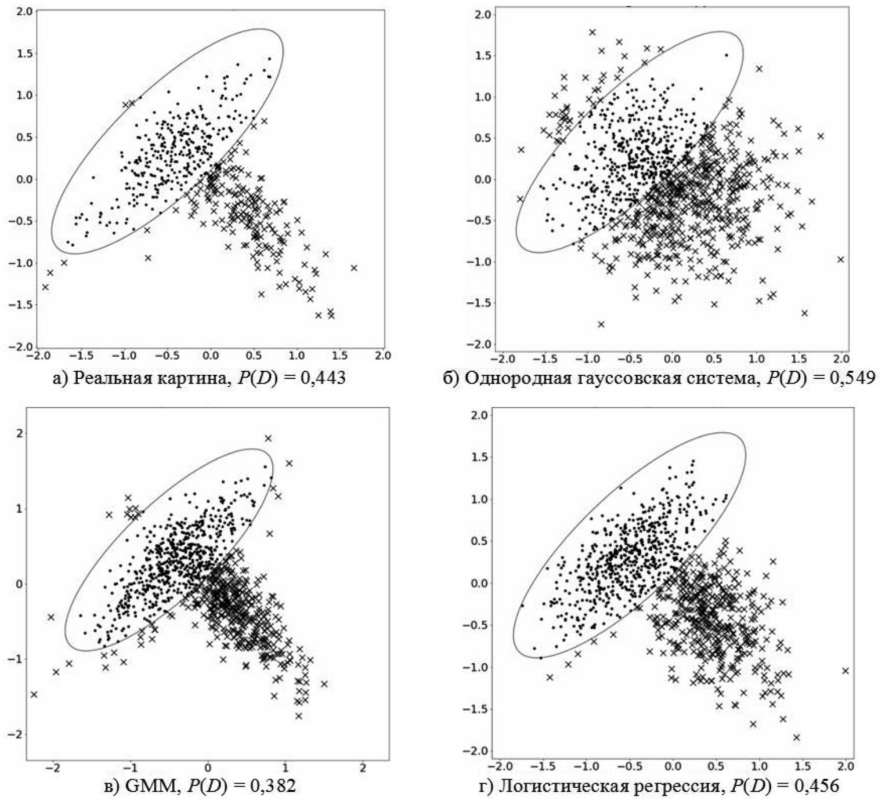

Рис. 4. Компоненты каждой из совокупности зависимы, между ними значимые пересечения

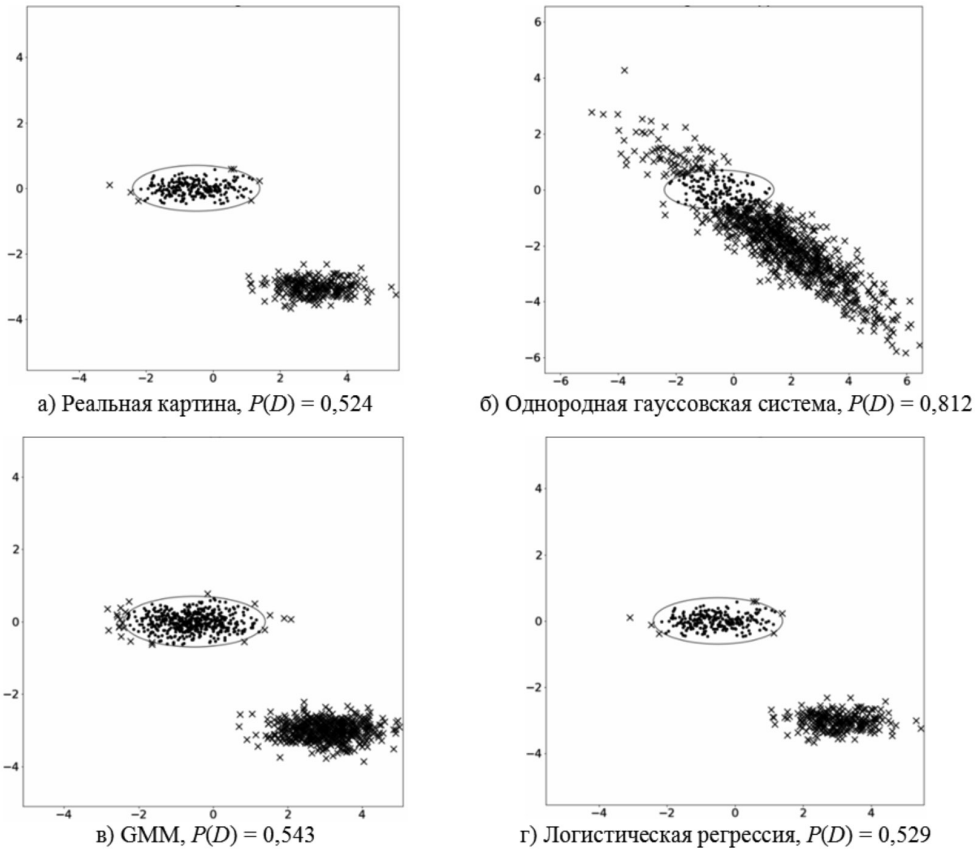

Рис. 5. «Непересекающиеся» классы

Анализ данных примеров показывает, что в четырех случаях из пяти использование модели в виде однородной гауссовской системы привело к значительному завышению оценки риска. Несмотря на то что для первых двух случаев (рис. 1, 2) нет выигрыша в точности оценки риска, снижения качества также не наблюдается. В остальных же случаях (рис. 3–5) использование модели для неоднородной системы предпочтительнее, поскольку результаты расчета рисков более приближены и сравнимы с реальными показаниями, в то время как однородная модель существенно завышает эти показатели.

Выводы

1. Показано, что представление неоднородных систем в виде гауссовского случайного вектора может приводить к значительным ошибкам в оценке риска.

2. Предложена модель многомерного риска для неоднородных стохастических систем. При этом система представляется в виде совокупности гауссовских случайных векторов.

3. Предложен и рассмотрен алгоритм для восстановления параметров распределения на основе вероятностей принадлежности каждой точки. В статье использовалась логистическая регрессии. Однако можно использовать и другие методы классификации.

4. GMM по сравнению с логистической регрессией имеет меньшую стабильность в расчете рисков, поэтому предпочтительнее использовать поиск параметров с помощью логистической регрессии. Но последний алгоритм вычислительно более затратен и требует некоторую информацию об исходных совокупностях в виде обучающей выборки.

Работа выполнена при финансовой поддержке гранта РФФИ, проект № 20-51-00001.

Список литературы

1. Шапкин А.С., Шапкин В.А. Теория риска и моделирование рисковых ситуаций. 5-е изд. М.: Дашков и К°, 2012. 880 с.

2. Акимов В.А., Лесных В.В., Радаев Н.Н. Риски в природе, техносфере, обществе и экономике. М.: Деловой экспресс, 2004. 352 с.

3. Кудрявцев А.А., Радионов А.В. Введение в количественный риск-менеджмент. СПб.: Издательство СПбГУ, 2016. 192 с.

4. Rossi C. Fundamentals of Risk Management. Wiley, 2014. 528 p.

5. Тырсин А.Н., Сурина А.А. Моделирование риска в многомерных стохастических системах // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. 2017. № 2 (39). С. 65–72. DOI: 10.17223/19988605/39/9.

6. Тырсин А.Н., Сурина А.А. Модели управления риском в гауссовских стохастических системах // Информатика и ее применения. 2018. Т. 12. Вып. 2. С. 50–59. DOI: 10.14357/19922264180208.

7. Hosmer D.W., Lemeshow S., Sturdivant R.X. Applied Logistic Regression. Third Edition. Wiley, 2013. 528 p.

8. Dempster A., Laird N., Rubin D. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B. 1977. Vol. 39 (1). P. 1–38.

9. Jordan M.I., Xu L. Convergence results for the EM algorithm to mixtures of experts architectures: Tech. Rep. A.I. Memo. No. 1458: MIT, Cambridge, MA, 1993. 34 p.

10. Тюри А.Г., Зуев И.О. Кластерный анализ, методы и алгоритмы кластеризации // Вестник МГТУ МИРЭА. 2014. № 2. С. 86–97.