Proc. of the European Safety and Reliability Conference (ESREL 2026)
Edited by José C. Matos, Paulo B. Lourenço, Michael Beer, Daniel Oliveira, Edoardo Patelli, Hélder S. Sousa, Rui A. Silva and Monica Santamaria
 Published by Research Publishing, Singapore.
doi: 10.3850/ESREL2026061419_esrel26-p27855-cd

Multidimensional Risk Analysis in Conditions of Stochastic Data Heterogeneity

Alexander Tyrsin

Institute for Risk and Reliability, Leibniz University Hannover, Germany. E-mail: alexander.tyrsin@irz.uni-hannover.de

Michail Gerasimov

Institute for Risk and Reliability, Leibniz University Hannover, Germany. E-mail: michail.gerasimov@irz.uni-hannover.de

Michael Beer

Institute for Risk and Reliability, Leibniz University Hannover, Germany. E-mail: beer@irz.uni-hannover.de

In a broad sense, risk is understood as the possible danger of any adverse outcome. When assessing the reliability of systems, it is often necessary to consider many risk factors that may be correlated. When conducting risk analysis, situations often arise when data samples are stochastically heterogeneous. Such situations include non-Gaussian distributions of risk factors, heterogeneity of data, the presence of outliers and small samples. This imposes requirements on risk analysis methods. In this case, risk analysis is fraught with difficulties. The paper considers the basic model of multidimensional risk, in which set risk factors are represented as a Gaussian vector with correlated components. This model assumes a known range of adverse outcomes. Stochastic heterogeneity of data leads to instability of risk analysis based on this model. Various ways of accounting for data heterogeneity are considered. An approach based on the preliminary identification of outliers for each of the risk factors and subsequent adjustment of the covariance matrix of the data is proposed. The analysis on test cases showed acceptable stability of risk analysis in conditions of contamination data.

Keywords: Risk, random vector, model, system, robustness, Gaussian distribution, heterogeneity of data.

1. Introduction

In a broad sense, "risk" is understood as the possible danger of any adverse outcome (Thompson et al. 2005; Alleman and Quigley 2024). In its simplest form, risk magnitude is the total expected outcomes, the probability, multiplied by the consequences (Ale 2009). Risk modeling, then, boils down to identifying hazardous outcomes, quantifying the consequences of their occurrence, and assessing the probabilities of these outcomes. If the consequences of hazardous outcomes cannot be assessed in advance with the required degree of certainty, then the approach is limited to assessing the probability of their occurrence.

Risk monitoring allows timely detection of malfunctions. It is one of the most important tools in assessing the reliability of systems (Robertson 2016; Abdullina 2025). However, monitoring the risk of real-world systems usually involves several difficulties. Let's list them.

Firstly, many risk factors may be mutually correlated (Engle 2009). This increases the likelihood of simultaneous manifestation of risk factors, which can lead to more significant consequences in the event of malfunctions.

Secondly, the stochastic heterogeneity of the data. This is reflected in their heterogeneity and the presence of emissions. Heterogeneity can be caused by the influence of unaccounted-for factors and limited sampling with multidimensional data (Nunes et al.2020; Wenz et al. 2023). Data outliers can occur for objective reasons because of the development of degradation processes in the system (Abdel-Hameed 2009). It can also be noted that the distributions of the samples analyzed in risk analysis are usually unimodal.

Thirdly, risk monitoring is primarily intended for the rapid detection of changes in the operation of the system, in addition, risk factors can often have trends. This leads to the need for risk assessment in small samples, while the mathematical model used should be fairly simple.

These features of risk monitoring of real systems impose requirements on risk analysis methods: 1) risk analysis should take into account the multidimensionality and stochastic heterogeneity of the data; 2) small sample sizes, the need to take into account the correlation of risk factors and the need for promptness of analysis make it difficult to use non-Gaussian distributions.

The aim of the article is to propose and test an approach to multidimensional risk analysis in conditions of stochastic heterogeneity of data.

2. The multidimensional risk model

Risk monitoring should be based on an adequate mathematical model. In (Tyrsin, Surina 2018; Tyrsin et al. 2025), an approach to risk modeling was proposed, according to which the stochastic system S is represented as a Gaussian vector $\mathbf{X} = (X_1, \dots, X_m)^T$, and its numerical characteristics are used as control variables – the vector of averages $\mathbf{a} = (a_1, \dots a_m)^T$ and the covariance matrix $\mathbf{\Sigma} = (s_{ij})_{m \times m}$. The components of vector $\mathbf{X}$ are risk factors that characterize the state of system S.

Instead of the generally accepted definition of specific dangerous situations, we will define a geometric domain D of adverse outcomes, which is determined based on available a priori information and may look arbitrary depending on the specific task. For the sake of certainty, we describe the proposed approach using the example of the widespread concept of undesirable events as large and unlikely deviations of the Gaussian vector $\mathbf{X}$ relative to the safe state of the system S.

The domain of safe states G can be described in various ways. From a geometric point of view, it seems most justified to define it as the inner domain of an m-axis ellipsoid as

$$G = \left\{\mathbf{x}: \sum_{j=1}^{m} \frac{(x_j - \theta_j)^2}{b_j^2} \le 1\right\}, \tag{1}$$

where $\mathbf{\theta} = (\theta_1, \dots, \theta_m)^T$, b_j are the center and semi–axes of the ellipsoid, the inner domain of which defines the safe states of the system.

Hits in the G and $D = \bar{G} = \Omega \setminus \mathrm{G} = \left\{\boldsymbol{x}: \sum_{j=1}^{m} \frac{(x_j - \theta_j)^2}{b_j^2} > 1\right\}$ domains are opposite events. Then the probability of an unfavorable outcome is defined as the probability of the random vector $\mathbf{X}$ not falling into the safe states of the system

$$P(D) = P(\mathbf{X} \in \bar{G}) = \int \dots \int_{\bar{G}} p_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}, \tag{2}$$

$p_{\mathbf{X}}(\mathbf{x})$ is the probability density of the vector $\mathbf{X}$.

Formula (2) allows us to estimate only the probability of an unfavorable outcome and does not consider the consequences of its occurrence. If we introduce into formula (2) the function of consequences from dangerous situations (the risk function) $g(\boldsymbol{x})$, we get a model for quantifying risk

$$r(\boldsymbol{X}) = \int \dots \int_{\mathbb{R}^m} g(\boldsymbol{x}) p_{\boldsymbol{X}}(\boldsymbol{x}) d\boldsymbol{x}. \tag{3}$$

Note that formula (2) is a special case of expression (3). Indeed, if we assume

$$g(\mathbf{x}) = \begin{cases} 1, \forall \mathbf{x} \in D, \\ 0, \forall \mathbf{x} \in G, \end{cases}$$

then $r(\mathbf{X}) = P(D)$, i.e. the risk is equal to the probability of an unfavourable outcome.

To set the function $g(\boldsymbol{x})$, it is not necessary to know the actual consequences of the occurrence of an unfavourable outcome in the analysed system. It is enough to quantify the dependence of consequences on the values of risk factors in conditional units. This can be done as follows:

1. We assume that the function $g(\boldsymbol{x})$ is a non-negative and continuous function everywhere on $\mathbb{R}^m$, with $g(\mathbf{\Theta}) = 0$.
2. We assume that the function $g(\boldsymbol{x})$ does not decrease in any direction from the point $\mathbf{\Theta}$.
3. $\forall j \; g\left(\theta_1, \dots, \theta_{j-1}, \theta_j \pm b_j, \theta_{j+1}, \dots, \theta_m\right) = 1$.

For example, the above conditions are satisfied by a paraboloid

$$g(\boldsymbol{x}) = \sum_{j=1}^{m} \alpha_j (x_j - \theta_j)^2.$$

At an early stage of the risk analysis of the system, it is difficult to accurately describe the function $g(\boldsymbol{x})$. Therefore, the probability $P(D)$ is a convenient initial approximation of the risk model. Also, the probability $P(D)$ can be used as a risk assessment to predict the occurrence of an adverse outcome. This article discusses the issue of using a multidimensional risk model in the context of stochastic heterogeneity of data. In this case, without losing the generality of the results, it is sufficient to consider the behaviour of the probability estimates $P(D)$.

Practical use of the multidimensional risk model requires knowledge or specification of the probability density of a random vector of $\mathbf{X}$ values of risk factors. We will use a Gaussian distribution with probability density

$$p_{\mathbf{X}}(\mathbf{x}) = \frac{\exp\left\{-\frac{1}{2}(\mathbf{x}-\mathbf{a})^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{a})\right\}}{\sqrt{(2\pi)^m |\mathbf{\Sigma}_{\mathbf{X}}|}},$$

where $\mathbf{a}$ is the vector of mathematical expectations and $\mathbf{\Sigma}_{\mathbf{X}}$ is the covariance matrix of the random vector $\mathbf{X}$.

Using a Gaussian distribution to model a random vector $\mathbf{X}$ is based on the following arguments: 1) the validity of the central limit theorem; 2) the analyzed sample is small, which makes it impossible to estimate the distribution law of a random vector; 3) in this problem, there is no need to identify the actual distribution $p_{\mathbf{X}}(\mathbf{x})$, since we are only interested in the probability $P(D)$.

The risk assessment is carried out as follows: 1) We set the parameters of the safe states of the system in (1); 2) Using a sample of data of size n, we determine the vector of averages a and the covariance matrix $\mathbf{\Sigma}_{\mathbf{X}}$; 3) Using the formula (2), we determine the risk $P(D)$.

Along with the risk, this model allows us to estimate the contribution to the overall risk of its components X_j as $\Delta P(D_j^-) = P(D) - P(D_j^-)$, where $P(D_j^-) = P(\mathbf{X} \backslash X_j \in D_j^-)$, D_j^- is the domain of unfavorable outcomes D after exclusion of risk factor X_j. It is also possible to estimate the contribution of the correlation between risk factors to the overall risk as $\Delta P(\mathbf{\Sigma}_{\mathbf{x}}) = P(\mathbf{x} \in D) - P(\tilde{\mathbf{x}} \in D)$, where all components $\tilde{X}_j$ of the random vector $\tilde{\mathbf{X}}$ are mutually independent and have the same distributions as X_j. Note that the correlation between risk factors X_j and its direction can both increase and decrease the amount of risk.

Calculating $P(D)$ by calculating integral (2) for large dimensions m is very laborious. Instead, the Monte Carlo statistical test method can be used. (Barbu and Zhu 2020). The essence of this procedure is as follows.

Let's have some data sample in the form of a matrix $\mathbf{X}_{n \times m}$, which we will conditionally call the general population. Let's denote its probability density by $p_{\mathbf{X}}(\mathbf{x})$. We must repeatedly generate new observation vectors $\mathbf{z}_i$, $i = 1, \dots, N$, with the distribution law $p_{\mathbf{X}}(\mathbf{x})$.

Then the probability estimate $P(D)$) will be equal to the frequency M/N, where N is the total number of observations generated $\mathbf{z}_i$, M is the number of outcomes when the generated observation $\mathbf{z}_i \in D$.

Example 1. Consider risk monitoring using an example. Let's analyze a two-factor system $\mathbf{X} = (X_1, X_2)$.

The sample numerical characteristics a_j, s_j, and $r_{x_1 x_2}$ and the results of calculating $P(D)$ for periods T_1, T_2, and T_3 are shown in Table 1. Sample size $n = 1000$.

Table 1. Risk values and numerical characteristics of non-emission factors

Parameters	T_1	T_2	T_3
a_1	4.603	4.804	5.305
a_2	4.201	3.901	3.700
s_1	0.356	0.447	0.559
s_2	0.230	0.311	0.404
$r_{x_1 x_2}$	-0.216	-0.376	-0.594
$P(D)$	0.001	0.003	0.077

Figures 1 (a), (b), and (c) show samples of risk factors in three consecutive time periods $T_1 < T_2 < T_3$ and domains G and D.

The increase in risk in the second and third periods is mainly caused by the drift and increase in the variance of factor X_2, as well as the increased closeness of the negative correlation between risk factors.

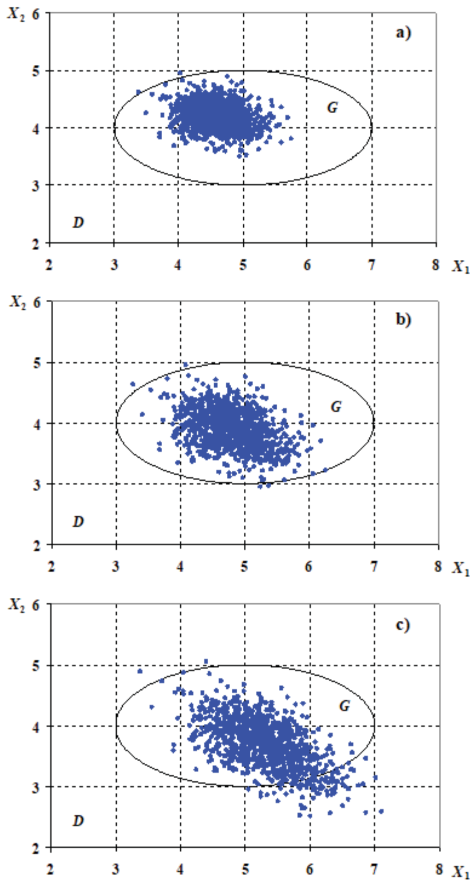

Figure 1. Samples of risk factors during periods T_1 (a), T_2 (b) and T_3 (c).

3. Accounting for stochastic heterogeneity of data

The multidimensional risk model based on (1), (2) has generally proved to be quite adequate in conditions of stochastic uniformity of data, when there are no outliers in the sample, which we mean deviations from the average by more than three standard deviations. The appearance of outliers in the data leads to an increase in covariances and variances. As a result, there is a significant overestimation of the risk assessment. Let's consider ways to account for such emissions.

There is a well–known approach based on the representation of data in the form of Gaussian mixtures (Gaussian Mixture Model – GMM) (Dempster et al. 1977). In the GMM model, the probability density function has the form of a weighted sum of a mixture of K Gaussian distributions

$$p(\mathbf{x}|\boldsymbol{\varepsilon}) = \sum_{k=1}^{K} \tau_k N(\mathbf{x}, \mathbf{a}_k, \boldsymbol{\Sigma}_k),$$

where $\sum_{k=1}^{K} \tau_k = 1, \forall k\ \tau \geq 0$.

The parameters of the distribution $p(\mathbf{x}|\boldsymbol{\varepsilon})$ are estimated using the maximum likelihood method (Jordan and Xu 1993). Since there is no analytical solution for mixtures (due to the presence of hidden variables – component indexes), an iterative EM algorithm (Expectation-Maximization) is used, which monotonically increases likelihood at each step, converging to a local optimum.

However, the use of GMM in risk analysis has shown that this method, with rare outliers and no clustering effect (which is typical in risk analysis), mistakenly determines the number of mixtures and usually overestimates the proportion of outliers (Tyrsin, and Maslennikov 2023). This limits the use of GMM in risk analysis in conditions of stochastic heterogeneity of data.

The second approach is the Minimum Covariance Determinant (MCD) method Hubert (2020). The algorithm searches for a subset of observations whose covariance matrix has a minimal determinant

$$\left(\hat{\mathbf{a}}, \hat{\boldsymbol{\Sigma}}\right) = \arg\min_{\mathbf{a},\boldsymbol{\Sigma}} \det(\boldsymbol{\Sigma}_h). \qquad (3)$$

The determinant of the covariance matrix is equal to the product of the determinant of the correlation matrix $\det(\mathbf{R})$ of risk factors by the product of their variances $s_{X_j}^2$

$$\det(\boldsymbol{\Sigma}) = \left(\prod_{j=1}^{m} s_{X_j}^2\right) \cdot \det(\mathbf{R}). \qquad (4)$$

From (4) we see that a decrease in $\det(\boldsymbol{\Sigma})$ occurs with a decrease in $s_{X_j}^2$ and $\det(\mathbf{R})$. However, as the variances $s_{X_j}^2$ decrease, the sample squeezes, which leads to an artificial underestimation of correlations between risk factors and an overestimation of $\det(\mathbf{R})$. Therefore, minimizing (3) has an ambiguous effect, both from the point of view of risk assessment and the definition of its components. Thus, the use of MCD in the risk assessment task does not have a clear interpretation.

There is a known approach in which data is approximated using a mixture of t-distributions

(Student Mixture Model – SMM) (Peel and McLachlan 2000) as

$$p(\mathbf{x}|\boldsymbol{\varepsilon}) = \sum_{k=1}^{K} \tau_k t_{\nu_k}(\mathbf{x}, \mathbf{a}_k, \boldsymbol{\Sigma}_k).$$

Replacing the normal distribution with a t-distribution makes it possible to correctly describe rare events without misclassifying them as anomalies. However, the procedure for selecting the numbers of degrees of freedom ν_k in SMM is insufficiently formalized. In addition, the weighting of the distribution tails does not correspond well to the essence of risk analysis: too large emissions, as a rule, cannot appear physically, because the system will lose its functionality before they appear.

Next, we use the traditional interpretation of outliers for the Gaussian distribution. As an estimate of the average a_j values of each of the factors, we use sample medians. We fix the significance level of α. For each of the risk factors, using known statistical procedures (Sheskin 2020), we determine the number of outliers m_j, $j = 1, \dots, m$. Accordingly, the emission fractions are $w_j = m_j/n$, where n is the number of observations.

For all factors X_j with $w_j > 0$, we recalculate their average square deviations so that $P(a_j - 3s_j \le X_j \le a_j + 3s_j) = 1 - w_j$. After that, we adjust the covariance matrix and assess the risk.

Let's check the efficiency of the proposed method using an example.

Example 2. In Example 1, we add two percent clogging to all three samples using the Tukey–Huber model as a mixture of two overlapping distributions

$$p(\mathbf{x}|\varepsilon) = 0.98N(\mathbf{x}, \mathbf{a}, \boldsymbol{\Sigma}) + 0.02N(\mathbf{x}, \mathbf{a}_\varepsilon, \boldsymbol{\Sigma}_\varepsilon),$$

where $N(\mathbf{x}, \mathbf{a}, \boldsymbol{\Sigma})$ are the initial Gaussian vectors of dimension $m = 2$, the numerical characteristics of which are given in Table 1; $N(\mathbf{x}, \mathbf{a}_\varepsilon, \boldsymbol{\Sigma}_\varepsilon)$ are the Gaussian vector of dimension m, the components of which have the same mathematical expectations, and the doubled values of the mean square deviations.

The risk calculations were performed using the approximation of the clogged data by a Gaussian distribution (3rd row of Table 2), as well as using the proposed robust method for adjusting the covariance matrix (fourth row of Table 2). The second row shows the estimates of $P(D)$ from Table 1.

From table 2, we see that even a small proportion of clogging significantly overestimates the risk.

Table 2. Risk assessment results $P(D)$: for a Gaussian sample, when the data is clogged with outliers, when correcting $\boldsymbol{\Sigma}_\mathbf{X}$ of the clogged sample

Data	T_1	T_2	T_3
Gaussian sampling	0.001	0.003	0.077
2% clogging	0.006	0.015	0.122
Correction of $\boldsymbol{\Sigma}_\mathbf{X}$	0.002	0.006	0.098

4. Conclusion

The known ways of estimating multidimensional risk in conditions of stochastic heterogeneity of data based on GMM, MCD and SMM models are considered. The analysis showed that they give unstable results in the presence of individual clogging in the form of outliers in the data.

An approach based on the preliminary identification of outliers for each of the risk factors and subsequent adjustment of the covariance matrix of the data is proposed. The analysis on test cases showed acceptable stability of risk analysis in conditions of clogging of Gaussian samples

Funding: This research was funded by the German Research Foundation (DFG), Project Entropy-probabilistic Methods and Models of Risk Analysis to Increase the Reliability of Complex Adaptive Systems (BE 2570/20-1).

References

1. Abdel-Hameed, M. (2009). Degradation Processes: An Overview. In: M.S. Nikulin, N. Limnios, N. Balakrishnan et al. (Eds.), *Advances in Degradation Modeling*. Birkhäuser Boston, pp. 17–25; [https://doi.org/10.1007/978-0-8176-4924-1_2]
2. Abdullina, L. (2025). The Impact of Digital Monitoring Technologies on Industrial Risk Management and ESG-oriented Enterprise Sustainability. *International Journal of Scientific and Management Research*, 8, 9, 19–25; [http://doi.org/10.37502/IJSMR.2025.8903]
3. Ale, B. (2009). *Risk: An Introduction: The Concepts of Risk, Danger and Chance*. Routledge, London; https://doi.org/10.4324/9780203879122]

4. Alleman, G.B., and Quigley, J.M. (2024). *Risk Management*. CRC Press; [https://doi.org/10.1201/9781003425465]
5. Barbu, A., Zhu S.C. (2020). *Monte Carlo Methods*. Springer; [https://doi.org/10.1007/978-981-13-2971-5]
6. Dempster A., Laird N., and Rubin D. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B*, *(Methodological)*. 39, 1, 1–38.
7. Engle, R.F. (2009). *Anticipating Correlations. A New Paradigm for Risk Management*. Princeton University Press; [http://doi.org/10.1007/978-0-387-89014-2]
8. Hubert, M. (2020). Robust Multivariate Statistical Methods. In S. Brown, R. Tauler, and B. Walczak (Eds.), Comprehensive Chemometrics (Second Edition), pp. 107-122, Elsevier; [https://doi.org/10.1016/B978-0-12-409547-2.14879-6
9. Jordan M.I., and Xu L. (1993). Convergence results for the EM algorithm to Mixtures of Experts Architectures: Tech. Rep. A.I. Memo. No. 1458: MIT, Cambridge, MA, 34 p.; [http://hdl.handle.net/1721.1/6620]
10. Nunes, A., Trappenberg, Th., and Alda, M. (2020). The Definition and Measurement of Heterogeneity. *Translational Psychiatry*, 10:299; [https://doi.org/10.1038/s41398-020-00986-0]
11. Peel, D., and McLachlan, G.J. (2000). Robust Mixture Modelling Using the *t*-distribution. *Statistics and Computing*, 10, 339–348; [https://doi.org/10.1023/A:1008981510081]
12. Robertson, D. (2016). *Managing Operational Risk*. Palgrave Macmillan, New York; [https://doi.org/10.1007/978-1-137-44217-8]
13. Sheskin, D.J. (2020). *Handbook of Parametric and Nonparametric Statistical Procedures*, 3d ed. Chapman and Hall, New York; [https://doi.org/10.1201/9780429186196]
14. Thompson, K.M., Deisler, P.F. Jr., and Schwing, R. C. (2005). Interdisciplinary Vision: The First 25 Years of the Society for Risk Analysis (SRA), 1980–2005. *Risk Analysis*, 25, 6, 1333–1386; [https://doi.org/10.1111/j.1539-6924.2005.00702.x]
15. Tyrsin, A.N.; Surina, A.A. Monitoring of Risk of Multidimensional Stochastic System as Tools for a Research of Sustainable Development of Regions. *IOP Conference Series: Earth and Environmental Science*. 2018, 177, 012005; [https://doi.org/10.1088/1755-1315/177/1/012005].
16. Tyrsin, A.N., and Maslennikov, D.L. (2020). A Risk Modeling in Heterogeneous Gaussian. *AIP Conference Proceedings*, 2312, 050025; [https://doi.org/10.1063/5.0035505]
17. Tyrsin, A., Gerasimov, M., and Beer, M. (2025). Monitoring Multidimensional Risk in the Economy. *Computer Sciences Mathematics Forum*, 11, 10. [https://doi.org/10.3390/cmsf2025011010]
18. Wenz, V., Kesper, A., and Taentzer, G. (2023). Clustering Heterogeneous Data Values for Data Quality Analysis. *ACM Journal of Data and Information Quality*, 15, 3, Article 28 (August 2023), 33 p; [https://doi.org/10.1145/3603710]