Robust data analysis in innovation project portfolio management

Boris Titarenko^{1,*}, Amir Hasnaoui², Roman Titarenko³ and Liliya Buzuk⁴

¹Moscow State University of Civil Engineering, Yaroslavskoe shosse, 26, Moscow, 129337, Russia

²La Rochelle Business School, France

³Stockholm School of Economics, Russia

⁴Russian State Social University, Moscow

Abstract. The paper states the mathematical model of portfolio management that allows to create an effective portfolio of innovation projects. Within the framework of this model the robust approach to data analysis is applied and expanded for the tasks of regression analysis of project data. The approach of robust estimation of regression parameters based on the maximum likelihood method in case of arbitrary contamination is suggested. A number of heuristic algorithms for estimating regression parameters in the case of symmetric data contamination is reviewed and modified.

1 Introduction

There are many definitions of *projects portfolio* in the recent scientific literature. For instance, Turner and Müller [1] define a portfolio as an organization where projects are managed together to coordinate interfaces, prioritize resources between projects, and thereby reduce uncertainty. According to [2] a portfolio is a group or set of projects with varying characteristics. Artto et al. [3] define projects portfolio as a collection of projects that are carried out in the same business unit sharing the same strategic objectives and the same resource pool.

In this article we will consider innovation projects portfolio as a set of innovation projects that are grouped in order to increase management efficiency and to achieve the strategic goals of the organization. In a rapidly changing environment and high competition, efficient management of the projects portfolio is an important tool for the success of any company. Project portfolio management (PPM) involves activities aimed at achieving the strategic goals of the organization by forming, optimization, monitoring and control as well as management of any changes of the projects portfolio under certain restrictions [4, 5]. According to [6] PPM solves key problems of project oriented

* Corresponding author: boristitarenko@mail.ru

organizations: overcomes the gap between operating and project management and becomes a core of all organizational activities.

Synergistic effect of a projects portfolio, in particular, is a simultaneous achievement of the best economic, financial, social and other final results. Synergies of a projects portfolio means the situation when obtained usefulness from the implementation of the projects portfolio exceeds the usefulness of each projects portfolio separately [6, 7]. Project portfolio management requires the processing of a large amount of information.

To make competent and effective decisions it's necessary to analyze carefully the available data, to study dependencies between factors that influence on the decision making. In situations of great uncertainty, in which projects are implemented in the modern world, processing methods of not enough robust data are necessary to provide *sufficiently reliable conclusions*. Within the robust estimation [8] that appeared in mathematical statistics in the 1960s and 1970s the ways and methods of obtaining robust estimates of statistical models parameters were identified. Over the last 50 years the scope of application of robust methods has been expanded.

Most of the modern works on this subject are devoted to the problems of multivariate statistical analysis and estimation of their parameters in case of the presence of gross errors in data or their contamination by extraneous data [8, 9].

However, the problems of parameters estimation for regression dependences are the main interest for researchers [10-13]. New approaches to robust procedures are also offered [14, 15]. Thus, the *robustness* concept begins to be interpreted more widely than by J. Tukey and P. Huber [8].

Recently, due to the wide application of project management methods, robust methods have been applied in project risk management [16-18]. This paper shows the robust methods that can be effectively used in innovation project portfolio management.

The purpose of this paper is to develop the model that allows to form an optimal projects portfolio and to suggest robust methods that can be effectively used in innovation project portfolio management.

This paper has the following structure. In Section 2 a model of innovation project portfolio management that allows to maximize the economic effect of the portfolio is shown. In Section 3 a robust approach to regression analysis of project data is suggested. Section 4 concludes the paper.

2 Formation of the innovation project portfolio management model

Formation of an innovation projects portfolio involves the setting priorities of projects within the organization and optimization of the projects components in the portfolio to ensure the best compliance of the portfolio with the strategic goals of the company. To that end, such methods are used that allow to set the priorities of the projects taking into account defined criteria in the organization, given the limited budget and resources [5].

The *effective innovation projects portfolio* means a set of projects that delivers the maximum gains within the existing resource constraints. To solve the problem of the effective innovation projects portfolio formation it is necessary to develop a mathematical model that can be applied to projects of all kinds, types and scope.

Denote needs of the projects in different types of resources by matrix R .

Problem of the optimal portfolio formation is solved under the constraints on the resources: financial, material, labor and other.

Let the number of resources is equal to m and the number of i -type resources is denoted as W_i . Thus, we have vector of resources $W = (W_1, W_2, ..., W_m)$. Projects portfolio can be

formed from a set of n projects $P_1, P_2, \dots, P_n$, i.e. there is a vector of projects $P = (P_1, P_2, \dots, P_n)$.

Let denote needs of the projects in different types of the resources by the matrix R :

$$R = \begin{pmatrix} R_{11} & R_{12} & \dots & R_{1m} \\ R_{21} & R_{22} & \dots & R_{2m} \\ \dots & \dots & \dots & \dots \\ R_{n1} & R_{n2} & \dots & R_{nm} \end{pmatrix},$$

where R_{ij} is a need in j resource of i -type project.

Define some integral indicator of the effectiveness of the i -type project E_i and consider the vector of efficiency of a portfolio $E = (E_1, E_2, \dots, E_n)$.

Let us introduce into consideration an integer binary variable Z_i :

$$Z_i = \begin{cases} 0, & \text{if } i\text{-type project is not included into portfolio;} \\ 1, & \text{if } i\text{-type project is included into portfolio.} \end{cases}$$

In this case, the selection of projects in the portfolio can be hold by solving the resource allocation problems for boolean variables (the so-called problems of integer linear programming). Our goal is to select the combination of projects, on the one hand that they fit within the resource capabilities and on the other hand that they maximize the outcome received by an enterprise.

Mathematical model enabling to form an effective portfolio will be as follows:

$$E = E_1 Z_1 + E_2 Z_2 + \dots + E_n Z_n \rightarrow \max$$

with limitations

$$R_{11} Z_1 + R_{21} Z_2 + \dots + R_{n1} Z_n < W_1;$$

$$R_{21} Z_1 + R_{22} Z_2 + \dots + R_{n2} Z_n < W_2;$$

...

$$R_{1m} Z_1 + R_{2m} Z_2 + \dots + R_{nm} Z_n < W_m.$$

In vector-matrix form it looks as

$$E = (E, Z) \rightarrow \max, \quad (1)$$

$$RTZ < W, \quad (2)$$

where Z is an integer vector,

Using this model, we obtain values Z_i for $i = 1, \dots, n$ that allows on their basis to form the optimal projects portfolio for maximum E economic effect.

To solve the problem of forming the optimal project portfolio the known methods for solving problems of integer linear programming can be applied and they are illustrated in the following approaches.

1st approach. The most natural way is to try to use the traditional methods of linear programming, such as simplex method, just modifying them a bit. So, it is possible to solve the problem of not paying attention to the requirement of integrality of the variables, and then round the coordinates of the obtained solution to the integer numbers. However, it's

possible to give some simple examples of such approach failure, when the solutions are actually far from optimal.

2nd approach. It is based on the effective exhaustive search methods. Their number, of course, is too large, therefore their all exhaustive search is practically impossible or very time and labour-intensive. Effective exhaustive search methods are to review only the most promising options and to represent a rapidly converging iterative procedure.

The problem is to work out modalities for the clipping of the inherently unpromising solutions based on the resource constraints. Here, it seems the most appropriate to apply the Branch and Bound method that consists in the determined exhaustive search of the solution tree branches. This method is often used in solving optimization problems in operations research and allows to obtain the exact solution of the problem for a finite number of steps.

Thus, in one variant of the method, variables are added one by one with a test of their resource endowment and all the sets are rejected for which these conditions stop to implement. The value of the objective function is defined for each possible branch and then it is compared with a maximum reached value.

Uncertain situation in which projects are implemented gives rise to numerous and varied risks. Therefore, the analysis and assessment of risk are very important for the formation of the projects portfolio and, finally, the risk largely defines the projects portfolio efficiency. For the formation of the efficiency criteria it's necessary to take into account both external risks arising from the environment of the enterprise and internal one accompanying a project activity.

Solving problems of innovation projects portfolio formation it is necessary to quantify the key indicators of projects risk. Any experienced specialist can calculate losses on the occurrence of a risk event, whereas the probability of occurrence of a risk event requires the use of special methods based on the proper use of available project information. As a rule, it is either real data of considering similar projects or their probable models.

The unreliability of data and inadequate models in situations of uncertainty are the sources of risk decision-making for managing projects.

For reliable estimates of risk events probabilities, the authors suggest the use the so-called *robust methods* described in Section 3.

3 Robust approach to regression analysis of project data

Classical methods of estimating parameters in mathematical statistics are based on the precise knowledge of the model distributions of random variables. The basic estimation method – *maximum likelihood method* defines the best estimate for each probability distribution. However, a significant disadvantage of this method is that the obtained estimates are sensitive to possible deviations from the assumed model distribution [17].

In practice, the observed distributions match the theoretical models only approximately and classical evaluations in this situation quickly lose their optimal features. This raises the problem of finding the estimates, may be not the most optimal, but resistant to such deviations. These estimates are *robust estimates*. The stability of statistical estimates in conditions of contaminated information is relevant enough in the processing of data for managing projects. While processing data for the purpose of managerial decision-making is often required to establish links between the results of decision-making and a variety of reasons that influence on the results. This problem relates to the field of robust regression analysis.

The dependence between some indicators $z_1, \dots, z_n$ and linked with them by another x indicator in most cases can be expressed in the form of a linear regression equation

$$x = \theta_1 z_1 + \dots + \theta_n z_n + \xi. \quad (3)$$

The estimates of the parameters $\theta_1, \dots, \theta_n$ in (4) usually are obtained by method of least squares that consists in solving the minimization problem

$$\sum_{i=1}^N (x_i - \theta_1 z_{1i} - \dots - \theta_n z_{ni})^2 = \min_{\theta_1, \dots, \theta_n}, \quad (4)$$

i.e. in choosing such $\theta_1, \dots, \theta_n$ that N observed sets $(x, z_1, \dots, z_n)$ provide the least deviation in terms of (4).

Solution of the problem (4) is equivalent to solving a set n of linear equations

$$\sum_{i=1}^N (x_i - \theta_1 z_{1i} - \dots - \theta_n z_{ni}) z_{ji} = 0 \quad (j = 1, \dots, n) \quad (5)$$

Even Tukey [8] suggested that a possible method of obtaining estimates, that are resistant to gross errors, is to replace the quadratic functions in (4) to another, less sensitivity to large fluctuations x_i .

He suggested to describe the presence of gross errors in the observations by the following model. Let $P(y)$ is the theoretical distribution of a random variable ξ in (3), but in the sample there are gross errors with the so called “contaminated” distribution $H(y)$. Then the resulting distribution has the form $P_\varepsilon(y) = (1 - \varepsilon)P(y) + \varepsilon H(y)$ where distribution $P(y)$ and “contaminating” $H(y)$ are symmetrical: $P(y) = 1 - P(-y)$, $H(y) = 1 - H(-y)$.

In the case when $P(y)$ is a function of the normal distribution, this model describes a situation where approximately $(1 - \varepsilon)N$ deviations of ξ_i obey the normal law. The magnitude ε (the intensity of contamination) is considered to be a known number.

For such a case Huber [8] suggested to use developed by him a common approach for the estimation of the location parameter to obtain sustainable estimates $\theta_1, \dots, \theta_n$ and instead of (4) to solve the problem.

$$\sum_{i=1}^N F(x_i - \theta_1 z_{1i} - \dots - \theta_n z_{ni}) = \min_{\theta_1, \dots, \theta_n} \quad (6)$$

with some properly chosen function F . This problem is reduced to solving a set of equations (as a rule, nonlinear already)

$$\sum_{i=1}^N f(x_i - \theta_1 z_{1i} - \dots - \theta_n z_{ni}) z_{ji} = 0 \quad (j = 1, \dots, n), \quad (7)$$

where $f(u) = F'(u)$.

Let's focus on the methods that use the idea of exclusion or modification of certain observations. In fact, they are the result of the transfer in case of a problem of regression estimates of truncated mean and Winsor's mean type. These methods are iterative. At each iteration, exclusion or modification of part of the observations occurs and based on the modified observations estimates of the regression parameters are found using the least squares method.

Consider the linear regression problem (3). Let estimates $\hat{x}_i$ for x_i are found by some method (e.g., least squares). Denote $d_i = x_i - \hat{x}_i$. Put in order deviations $d_i: d_1 \leq d_2 \leq \dots \leq d_N$. By analogy with the estimate of Winsor's mean, let's construct the Winsor's regression line which is calculated using the least squares method on a sample of N points from $(x'_i, z_1, \dots, z_n)$, where x'_i is defined as follows

$$x'_i = \hat{x}_i + d'_i;$$

$$d'_i = \begin{cases} d_{g+1}, & i = 1, \dots, g; \\ d_i, & i = g+1, \dots, N-g; \\ d_{N-g}, & i = N-g+1, \dots, N; \end{cases} \quad (8)$$

where g is a number of extreme deviations (largest and smallest) to be modified by Winsor.

It can be recommended to use this procedure one of the following ways.

The first method is a simple iteration method. The number of g points modifiable in accordance with (8) remains constant at each iteration. Deviations are calculated using observations $(x_i, z_1, \dots, z_n)$ at the first iteration, observations $(x'_i, z_1, \dots, z_n)$ at the second, etc.

The second method is a method of levels. The number of g points (the level of truncation) increases from iteration to iteration, and the procedure of finding Winsor's regression lines is made each time with the initial data $(x_i, z_1, \dots, z_n)$.

The third method is an iterative method with increasing level. This method is a combination of the first two. Deviations are calculated as in the simple iterative method, at first using the initial data $(x_i, z_1, \dots, z_n)$, then using $(x'_i, z_1, \dots, z_n)$, etc., and the level of truncation g increases from iteration to iteration.

4 Discussion

The suggested model of an efficient innovation projects portfolio formation represents theoretical and practical significance due to the following reasons. First, it clearly identifies the goal of portfolio management – maximization of profit of the organization and it shows the way to achieve it. Second, the authors suggest the ways to achieve this goal depending on the quality of management information.

Data analysis when making management decisions plays a big role in ensuring their reliability. In the process of risk analysis in projects of proper regression processing of the data is crucial because robust methods give more reliable estimates of regression parameters.

The paper describes some heuristic algorithms that are implemented by the approach of P. Huber in case of symmetric contamination. It seems the actual problem is development of robust estimation methods of regression parameters in case of arbitrary contamination.

The developed methods can be applied not only in the project management problems, but also in the fields such as cluster analysis, regression models and multivariate analysis, variation analysis, factor analysis, planning of experiments, simulation, statistical estimation of models parameters, estimation of systems reliability, general statistical problems.

Currently the robust approach can be implemented with the information technology, therefore, it is possible to suggest the following approaches to their implementation.

1. Modification of the already developed robust methods for solving specific problems and creating appropriate mathematical software.
2. Adaptation of the ready packages of the robust software to the specific character of the solved problems.
3. “Robustification” of the available mathematical software in order to create quasi-robust procedures.

Thus, the application of the suggested approaches to robust data analysis in innovation project portfolio management in conjunction with the use of information technologies for processing management information will enhance the reliability of management decisions and can be an effective tool for project portfolio managers.

References

1. J.R. Turner, R. Müller, *International Journal of Project Management* **21**, 1–7 (2003)
2. J.R. Meredith, S.J. Mandel, *Project management: a managerial approach* (7th ed.) (Hoboken, John Wiley & Sons, NJ, 2010)
3. K.A. Artto, M. Martinsuo, T. Aalto, *Project Portfolio Management: Strategic Management through Projects* (Project Management Association Finland, Helsinki, 2001)
4. A. Hyväri, *Procedia – Social and Behavioral Sciences* **119**, 229–236 (2014)
5. O. Momčilović, L. Djukic Petromanjanc, S. Doljanica, J. Rajaković, *Annals of the University of Oradea* **3**, 192–196 (2014)
6. H.A. Levine, *Project Portfolio Management* (Jossey-Bass, Wiley Imprint, USA, 2005)
7. B. Maizlish, R. Handler, *IT portfolio management step-by-step: unlocking the business value of technology* (Hoboken, John Wiley & Sons NJ, 2005)
8. P. Huber, E. Ronchetti, *Robust statistics* (J. Wiley, New Jersey, 2009)
9. C. Agostinelli, A. Leung, V.J. Yohai, R.N. Zamar, *TEST* **24**, 441–461 (2015)
- A. Ghosh, A. Basu, *TEST* **25**, 269–290 (2016)
10. L. Feng, C. Zou, Z. Wang, L. Zhu, *TEST* **24**, 185–204 (2015)
11. T. Qingguo, *Stat Papers* **56**, 1137–1161 (2015)
12. S. Hwang, D. Kim, M.K. Jeong, B.-J. Yum, *Journal of the Operational Research Society* **66**, 1385–1398 (2015)
13. J.Á. Višek, *Methodol Comput Appl Probab* **17**, 999–1014 (2015)
14. C. Zeller, C. Cabral, V. Lachos, *TEST* **25**, 375–396 (2016)
15. B. Titarenko, *International Journal of Project management* **15**(1), 11–14 (1997)
16. V. Shulenin, *Robust methods of mathematical statistics* (Russia, Tomsk, 2016)
17. B. Titarenko, S. Titov, R. Titarenko, *Applied Mechanics and Materials* **638–640**, 2338–2341 (2014)