Mean-Variance Models with Missing Data

Marius RĂDULESCU[1], Constanța Zoie RĂDULESCU[2]

[1] Institute of Mathematical Statistics and Applied Mathematics, Casa Academiei Române,
13, Calea 13 Septembrie, 050711 Bucharest 5, ROMANIA,
mradulescu.csmro@yahoo.com

[2] National Institute for Research and Development in Informatics,
8-10, Averescu Avenue, 011455 Bucharest 1, ROMANIA,
radulescucz@yahoo.com

Abstract: A common challenge in the theory of portfolio selection is that certain assets have shorter return histories than others. Consequently, historical data of the returns have missing data. This paper deals with portfolio selection models of mean-variance type in which missing data exist. Two simple methods for constructing a vector and a matrix starting from a matrix of rate of returns are presented. One considers a standard minimum variance model in which the vector and the matrix built replace the vector of means and the matrix of covariance. Several numerical experiments are made and the effect of missing data on the efficient frontiers associated to the minimum variance models is investigated.

Keywords: mean-variance model, minimum variance model, missing data, NaN vector of means, NaN covariance matrix.

1. Introduction

The last decades witnessed a growing amount of attention given to the topic of missing data. Several research papers and books were written on this important subject. Most PhD students in Statistics now claim "missing data" as an area of interest or expertise. Missing data are important to consider, because they may lead to substantial biases in analyses. On the other hand, missing data is often harmless beyond reducing statistical power.

For a complete treatment of the issue of missing data the books written by Little and Rubin [5] and Schafer [15] are excellent choices. A shorter treatment can be found in Allison [1] and a gentle one in McNight et al [8]. Perhaps the nicest treatment of modern approaches can be found in Barladi & Enders [2].

There are some traditional treatments for missing data. The simplest approach is called listwise deletion or complete case analysis**.** It consists in deleting those cases with missing data and continuing analyses on what remains. For example if we want to compute the

arithmetic mean of n numbers $a_1, a_2, ..., a_n$ but only the numbers $a_{i_1}, a_{i_2}, ..., a_{i_k}$ are known then we shall consider that the arithmetic mean of the numbers $a_1, a_2, ..., a_n$ is equal to

$$m = \frac{a_{i_1} + a_{i_2} + ... + a_{i_k}}{k}$$

Another simple approach is to replace all missing data with the arithmetic mean of all known data. If $a_{j_1}, a_{j_2}, ..., a_{j_{n-k}}$ are the unknown terms of the sequence $a_1, a_2, ..., a_n$ we shall put $a_{j_1} = a_{j_2} = ... = a_{j_{n-k}} = m$. Then one can easily note that the arithmetic mean of the numbers $a_1, a_2, ..., a_n$ is equal to m.

Although the listwise deletion approach often is applied for analyses with small sample size, it does have important advantages. In particular, under the assumption that data are missing completely at random, it leads to unbiased parameter estimates.

Other simple methods for treatment are: pairwise deletion, mean substitution, averaging the available variables, regression-based single imputation. Recommended methods for handling missing data fall into two general categories: model-based procedures and data-based procedures. Model-based approaches rewrite the statistical algorithms so as to handle the missing data and estimate parameters all in a single step. Data-based approaches, on the other hand, handle the missing data in one step, and then perform the parameter estimation in a second, distinct, step.

Software modules for handling problems with missing data are included in the following software packages: IBM SPSS, SAS STAT, MATLAB, SOLAS, AMELIA.

The IBM SPSS Missing Values software may be used by survey researchers, social scientists, data miners, market researchers and others to validate data. It uses statistical algorithms and allows the users to examine data, to uncover missing data patterns, then to estimate summary statistics and

to impute missing values. The SPSS Missing Values software allows the imputation of missing data, draw valid conclusions and remove hidden bias. It quickly diagnose missing data imputation problems using diagnostic reports, it replaces missing data values with estimates using a multiple imputation model, display and analyze patterns to gain insight and improve data management.

SAS STAT® software offers the MI and MIANALYZE procedures for creating and analyzing multiply imputed data sets for incomplete multivariate data. Multiple imputation provides a useful strategy for dealing with data sets with missing values. Instead of filling in a single value for each missing value, multiple imputation procedure (Rubin [13]) replaces each missing value with a set of plausible values that represent the uncertainty about the right value to impute. These multiply imputed data sets are then analyzed by using standard procedures for complete data and combining the results from this analysis. No matter which complete-data analysis is used, the process of combining results from different imputed data sets is essentially the same. This results in statistically valid inferences that properly reflect the uncertainty due to missing values.

The representation of missing or unavailable data values in MATLAB® code is made with the special value, NaN which stands for Not-a-Number. When the data is plotted on a time-plot that contains missing values, gaps appear on the plot where missing data exists. In order to estimate missing values one can use the command misdata. This command linearly interpolates missing values to estimate the first model. Then, it uses this model to estimate the missing data as parameters by minimizing the output prediction errors obtained from the reconstructed data. The model structure is specified by the user in the argument of the command misdate. Alternatively a default-order model using the n4sid method will be estimated.

An interesting paper which contains a survey on the software for missing data is Hox [3]. A study of the efficient frontier of portfolio selection models with missing data, using MATLAB can be found in Taylor [14].

In our paper we show that to each matrix **R** containing missing data one corresponds a binary matrix **B** having the same dimension with the matrix **R**, that describes the location of missing data in the matrix **R**. Thus the arrays of matrix **B** are equal to zero in the corresponding location of matrix **R** where there is a missing data and are equal to one where in the corresponding location of matrix **R** there exist an array whose value is known.

Vice-versa if we have a complete matrix **R** (that is all its arrays are known values) and a binary matrix **B** having the same dimension with the matrix **R** then we can treat the matrix **R** as a matrix with missing data, the locations of missing data in matrix **R** being those corresponding to the locations where the arrays

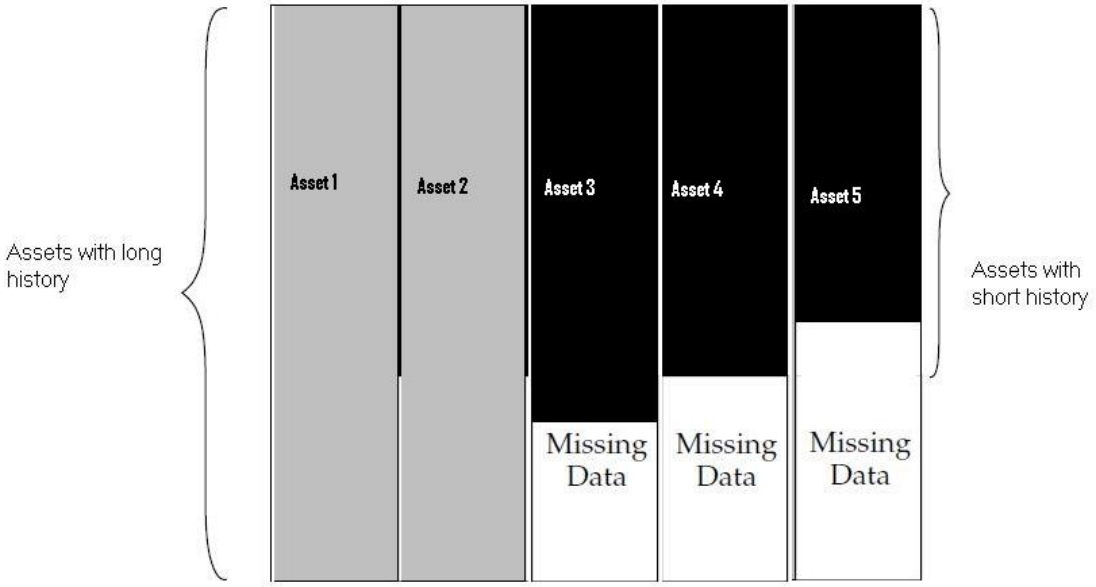

Figure 1. Assets with various rate of return history

of matrix **B** are equal to zero. Two simple methods for constructing a vector and a matrix starting from a matrix of rate of returns are presented. The vector and the matrix built will replace the vector of means and the matrix of covariance in a standard minimum variance model. Several experiments are made and the effect of missing data on the efficient frontiers associated to the minimum variance models is investigated. The purpose of the paper is to consider some numerical examples in order to show how diverse, the impact of increasing the proportions of missing data, on efficient frontiers can be.

2. The NaN - Vector of Means and the NaN - Covariance Matrix

The mean-variance portfolio optimization theory of Markowitz [6], [7] is widely regarded as one of the major theories in financial economics. It is a single-period theory on the choice of portfolio weights that provide optimal tradeoff between the mean and the variance of the portfolio return. Mean-variance theory is an important model of investments based on decision theory. It is the simplest model of investments that is sufficiently rich to be directly useful in applied problems. There exist many applications of mean-variance theory to domains that do not imply finance such as agriculture, sire selection, forestry, biodiversity, aquaculture, energy, sustainable production planning etc. For supplementary references regarding applications of portfolio theory to non-financial areas see Radulescu [10]-[12].

In order to use the mean-variance theory it is necessary to estimate the covariance matrix of a random vector.

Estimation of the covariance matrix is a basic problem in multivariate statistics. It arises in various applications such as financial mathematics, pattern recognition, genomics, functional analysis, computational geometry etc. A good reference for the estimation of the covariance matrix of the stock returns with application to portfolio selection is the paper

Ledoit and Wolf [4]. Another approach to portfolio selection models with missing data can be found in Page [9].

A classical and the simplest estimator of the covariance matrix is the sample covariance matrix. Consider a $m \times n$ matrix $\mathbf{R} = (r_{ti})$ where each column represents m observations of a random variable and each row observations at a particular time. That is r_{ti} is the t-th observation of the i-th random variable. Denote by $\boldsymbol{\mu}$ the vector of means, by $\mathbf{C}$ the sample covariance matrix and by $\mathbf{e}_m$ the m-dimensional vector having all components equal to one. Then

$$\boldsymbol{\mu} = \frac{1}{m}\mathbf{R}^T\mathbf{e}_m$$

$$\mathbf{C} = \frac{1}{m}\mathbf{R}^T\left(\mathbf{I}_m - \frac{1}{m}\mathbf{e}_m\mathbf{e}_m{}^T\right)\mathbf{R}$$

Note that the following inequality holds:

$$\text{rank}(\mathbf{C}) \le \text{rank}\left(\mathbf{I}_m - \frac{1}{m}\mathbf{e}_m\mathbf{e}_m{}^T\right) = m\text{-}1$$

Hence when $n \ge m$-1 the matrix $\mathbf{C}$ is rank deficient, that is rank($\mathbf{C}$)=0. Intuitively the data do not contain enough information to estimate the covariance matrix.

Consider n assets and historical data on the rate of return for the assets for m periods. Let r_{ti} be the rate of return of asset i at moment t. Let $\mathbf{R} = (r_{ti})$ be the rate of returns matrix. $\mathbf{R}$ is a $m \times n$ matrix. We consider that t=1 is the present and t=m is the earliest moment taken into account. Some elements of the matrix $\mathbf{R} = (r_{ti})$ may be unknown. These elements will be called missing data and in their places we shall fill with the label NaN (MATLAB convention for Not a Number). Starting from the matrix $\mathbf{R} = (r_{ti})$ we shall define the matrices $\tilde{\mathbf{R}} = (\tilde{r}_{ti})$ and $\mathbf{B} = (b_{ti})$,

$$\tilde{r}_{ti} = \begin{cases} 0, \text{ if } r_{ti} \text{ is not known (NaN)} \\ r_{ti}, \text{ otherwise} \end{cases}$$

$$b_{ti} = \begin{cases} 0, \text{ if } r_{ti} \text{ is not known (NaN)} \\ 1, \text{ otherwise,} \end{cases}$$

Note that $\mathbf{B} = (b_{ti})$ is a binary matrix, that is all its elements are zero and one. If the binary $m \times n$ matrix $\mathbf{B} = (b_{ti})$ has the property that for every $i, j \in \{1,2,\ldots,n\}$ there is at least $t \in \{1,...,m\}$ such that both $b_{ti} = b_{tj} = 1$ we can define the NaN vector of means

$\boldsymbol{\mu} = (\mu_1, \mu_2, ..., \mu_n)$ and the NaN covariance matrix $\mathbf{C} = (c_{ij})_{1 \le i,j \le n}$

$$\mu_i = \frac{\sum_{t=1}^{m} \tilde{r}_{ti} b_{ti}}{\sum_{t=1}^{m} b_{ti}}, \quad (1)$$

$$c_{ij} = \frac{\sum_{t=1}^{m} b_{ti} b_{tj} (\tilde{r}_{ti} - \mu_i)(\tilde{r}_{tj} - \mu_j)}{\sum_{k=1}^{m} b_{ti} b_{tj}}, \quad (2)$$

In case of financial historical data the time series of assets may have various lengths. The columns of the rate of returns matrix **R** correspond to historical data of the assets. One column may contain NaN entries for the period before the moment the asset begin to be publicly traded followed by known values of assets until the most recent moment.

Let $\sigma : \{1,2,...,n\} \to \{1,2,...,m\}$, $\sigma(i) =$ the moment the asset *i* begin to be publicly traded.

Then

$b_{ti} = 0$ if $t > \sigma(i)$, $b_{ti} = 1$ if $1 \le t \le \sigma(i)$

In the following two sections we shall consider two methods for the treatment of missing data from the historical data on the rate of returns. The first method replaces the missing data for the asset with the mean of the known data for the rate of returns of the asset. The second method follows the listwise approach and ignores the missing data. All computations are made only with the data that have known values.

3. The First Method for Treatment of Missing Data

We consider a matrix **R** whose arrays describe historical data on the rate of return of *n* assets for *m* periods. Each column of **R** corresponds to an asset and each row corresponds to the rate of return of the assets. Note that historical data of the assets have variable lengths since the assets started to be publicly traded at various moments in time. Consequently the matrix **R** contains missing data. We shall consider the corresponding binary matrix **B** that describes the location of missing data in the matrix **R**. We shall denote matrix **R** by $\mathbf{R}_0$. That is $\mathbf{R}_0$=**R**.

Let *i* be the asset for which the historical data have minimum length. Denote by *k* be the length of historical data for asset *i*. Delete from matrix **R** the last *s* rows (*s*=0,1,…,*n*-*k*). Denote by $\mathbf{R}_s$ the matrix obtained as a result of the deletion.

Replace the missing data from the matrices $\mathbf{R}_s$ with the mean of the known values of the return of the assets from the column. At this moment all the matrices $\mathbf{R}_s$ become complete matrices, that is, all the values of their arrays are known. One can formulate a minimum variance model starting from the vector of means and the covariance matrix obtained from the matrices $\mathbf{R}_s$. For each of these minimum variance models we plot the efficient frontier.

We consider a numerical example in which the matrix **R** has *m*=112 rows and *n*=14 columns. The historical data for six assets have maximum length, that is 112. The historical data for the asset *i*=11 have minimum length, that is 5. By successive deletion of rows from matrix **R** and by replacing the missing data with the mean of the known values of the rate of return of the assets from the column we obtain matrices $\mathbf{R}_s$ (*s*=0,1,…,107). We consider the minimum variance models starting from the vector of means $\boldsymbol{\mu}_s$ and the covariance matrix $\mathbf{C}_s$ obtained from the matrices $\mathbf{R}_s$.

$$\begin{cases} \min\left[\mathbf{x}^T \mathbf{C}_s \mathbf{x}\right] \\ \boldsymbol{\mu}_s{}^T \mathbf{x} \ge W \\ \mathbf{e}^T \mathbf{x} = 1 \\ \mathbf{x} \ge 0 \end{cases}$$

The efficient frontiers of the minimum variance models are displayed in Figure 2. One of the axes from the horizontal plane represents the percent of the missing data. In fact it corresponds to parameter *s*. If the parameter *s* increases then the percent of the missing data in the matrix $\mathbf{R}_s$ decreases. The other axis from the horizontal plane represents the lower limit for the mean expected return (parameter *W*). The vertical axis represents the optimal value of the minimum variance model (defined for the parameter *W* and parameter *s*).

One can note that if the proportion of missing data is greater than 5% then the efficient frontiers are very distinct from those with the proportion of missing data smaler than 5%.

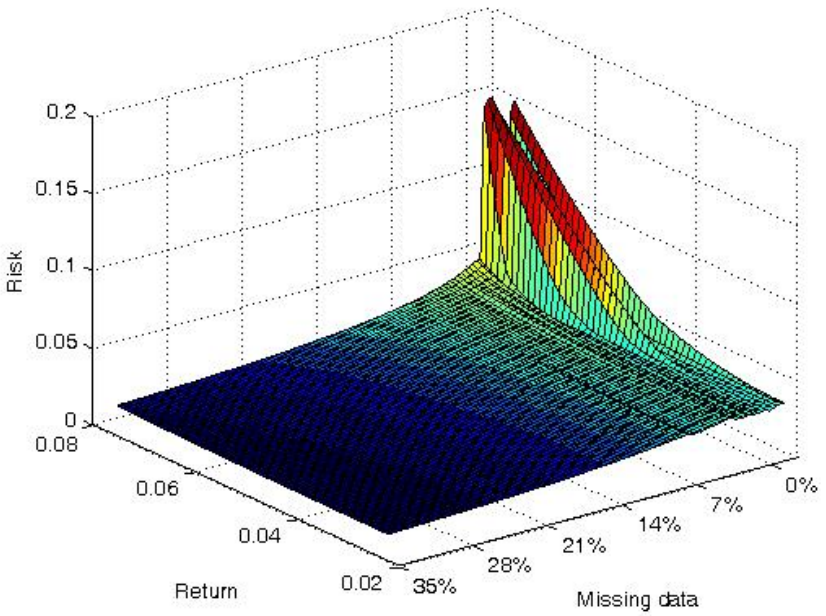

Figure 2. The efficient frontiers of the minimum variance models when the first method for treating missing data is applied

4. The Second Method for Treatment of Missing Data

The second method for treatment of missing data is based on the listwise deletion approach. Consider a matrix **R** whose arrays describe historical data on the rate of return of *n* assets for *m* periods. Each column of **R** corresponds to an asset and each row corresponds to the rate of return of the assets. We consider that matrix **R** is complete, that is, all its arrays are known. We consider a set of binary matrices B that describe fictitious locations of missing data in the matrix **R**. For each binary matrix **B** from B we consider the matrix **R**(**B**) which contain missing data at locations corresponding to arrays that have values equal to zero in matrix **B**. For each matrix **R**(**B**) we build the NaN - vector of means and the NaN-covariance matrix. The efficient frontiers the minimum variance models formulated with the NaN vector of means and the NaN-covariance matrices are displayed in Figure 2.

One can formulate a minimum variance model starting from the NaN vector of means $\boldsymbol{\mu} = (\mu_1, \mu_2, ..., \mu_n)$ and the NaN covariance matrix $\mathbf{C} = (c_{ij})_{1 \le i,j \le n}$

$$\begin{cases} \min\left[\mathbf{x}^T \mathbf{C} \mathbf{x}\right] \\ \boldsymbol{\mu}^T \mathbf{x} \ge W \\ \mathbf{e}^T \mathbf{x} = 1 \\ \mathbf{x} \ge 0 \end{cases}$$

The problem is that the NaN covariance matrix **C** may fail to be semi-positive definite. This will imply that the objective function may fail to be convex.

But the restriction of the objective function to the set of feasible solutions may be convex.

If the volume of missing data is sufficiently large it is possible that even the restriction of the objective function to the set of feasible solutions fail to be convex.

This will imply difficulties in solving the optimization problem since a function which is not convex may attain its minimum at several points. In this situation standard algorithms for convex problems may not apply. Of course if the volume of missing data is sufficiently large the effects can result in significantly different results.

We consider a numerical example in which the matrix $\mathbf{R} = (r_{ti})$ has *m*=112 rows and *n*=14 columns. Each column in the matrix **R** corresponds to historical data of the rate of returns for a specific asset. All historical data of the assets have full length. We consider a set of

binary matrices $\mathbf{B}=\{\mathbf{B}_s : s\in\{0,1,...,m-5\}\}$. Each matrix $\mathbf{B}_s=\left(b_{ti}^{(s)}\right)$ is a $m\times n$ matrix. $b_{ti}^{(s)}=1$ for $(t,i)\in\{1,2,...,m\}\times\{1,2\}$ and for $(t,i)\in\{1,2,...,m-s\}\times\{3,4,...,n\}$. $b_{ti}^{(s)}=0$ for $(t,i)\in\{m-s+1,m-s+2,...,m\}\times\{3,4,...,n\}$.

Denote by $\boldsymbol{\mu}_s$ the NaN vector of means and with $\mathbf{C}_s$ the NaN covariance matrix corresponding to the couple of matrices $(\mathbf{R},\mathbf{B}_s)$ Recall that by the couple $(\mathbf{R},\mathbf{B}_s)$ we understand a matrix $\mathbf{R}_s=\left(r_{ti}^{(s)}\right)$

$$r_{ti}^{(s)}=\begin{cases} r_{ti} & \text{if } b_{ti}=1 \\ \text{NaN} & \text{if } b_{ti}=0 \end{cases}$$

Consider the minimum variance models starting from the vector of means $\boldsymbol{\mu}_s$ and from the covariance matrix $\mathbf{C}_s$.

We define the missing data percent of a matrix as being the ratio between the number of NaN arrays and the total number of arrays.

The missing data percent of the matrix $\mathbf{R}_s=\left(r_{ti}^{(s)}\right)$ is equal to the number of zero elements from the matrix $\mathbf{B}_s=\left(b_{ti}^{(s)}\right)$ divided by mn. In Figure 3 are displayed three efficient frontiers of the minimum-variance model with missing data. One can easily see that

- All three graphs are very close each other for W in the range [0.01;0.038].
- The efficient frontier graph for the missing data percent 75% starts to move away from the two other graphs starting from W=0.038.

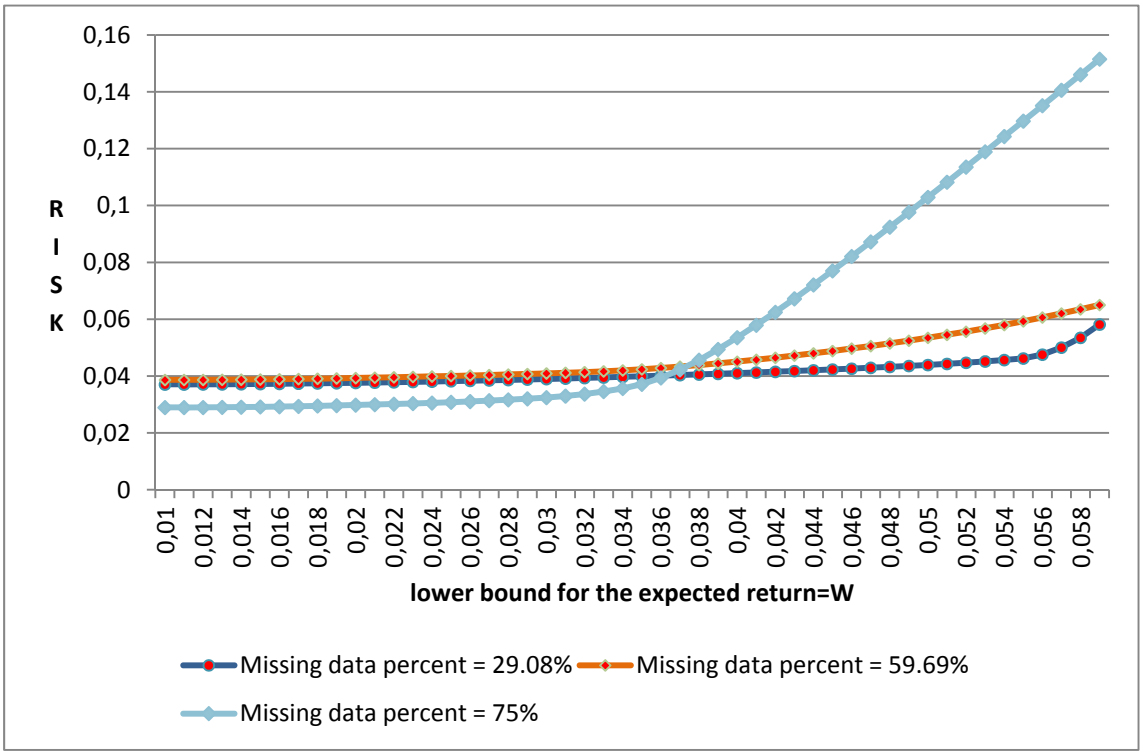

Figure 3. Efficient frontiers of the minimum variance models for various missing data percents

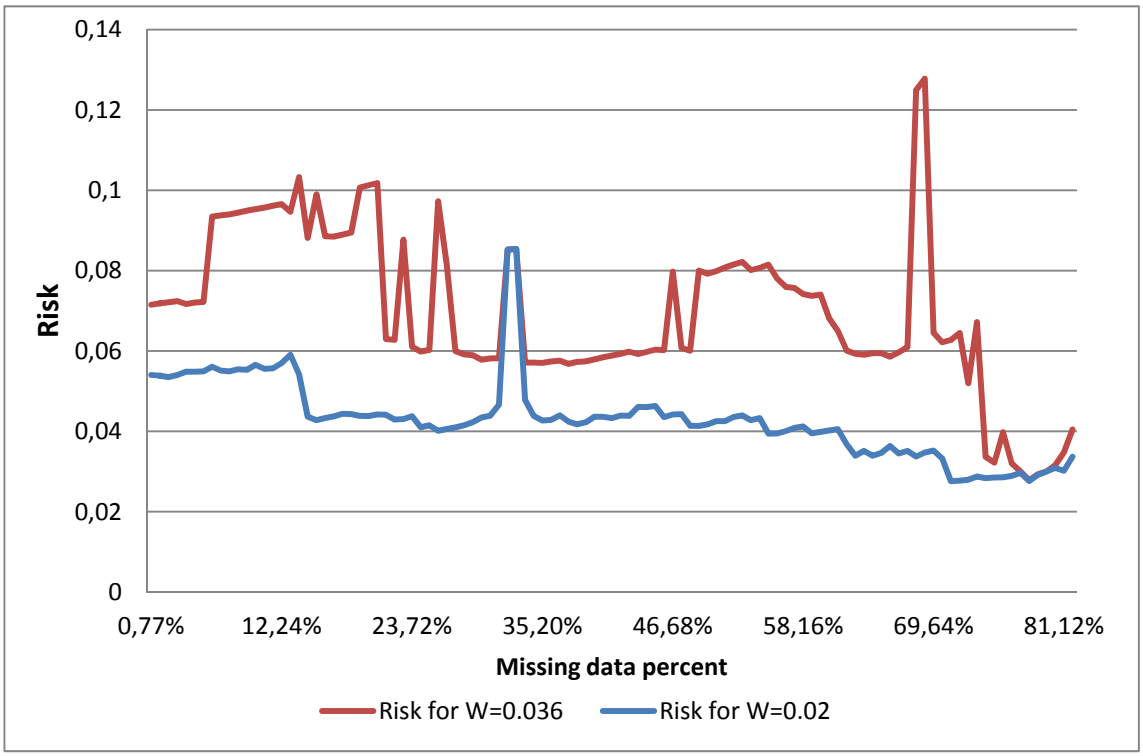

Figure 4. The graph of risk versus the missing data percent

In Figure 4 is displayed the minimum risk as a function of the percent of missing data for two values of the parameter W (W_1=0.036 and W_2=0.02). One can note that the behavior of the two graphs is random.

In Figure 5 are displayed efficient frontiers of the minimum variance models for various missing data percents. One can see that the shape of the efficient frontiers varies very much. For the models considered a missing data percent under 10% seems to be acceptable.

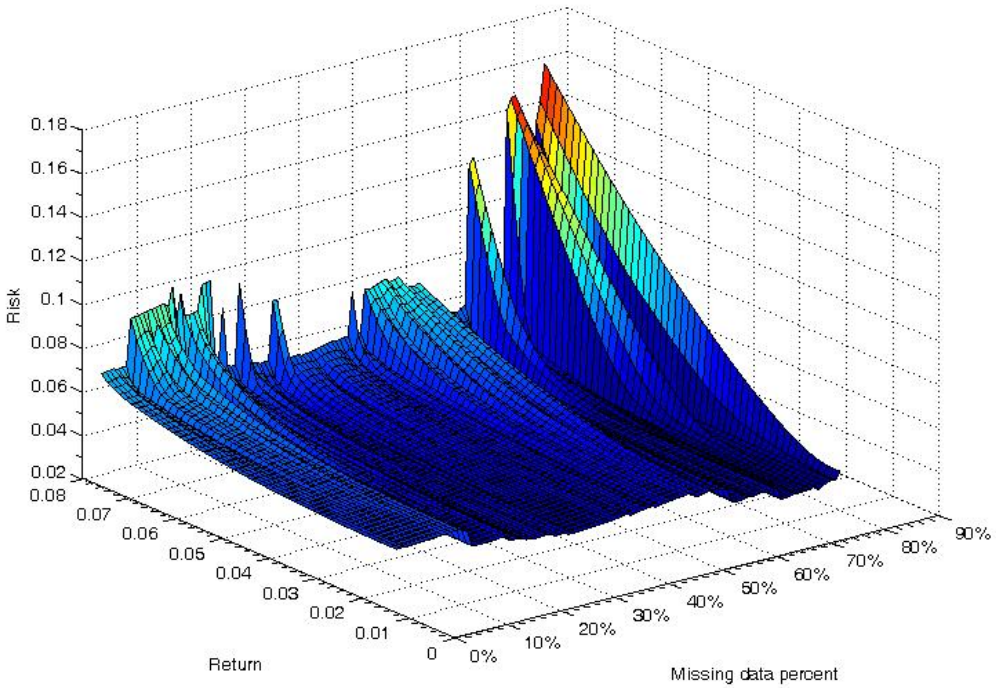

Figure 5. The efficient frontiers of the minimum variance models

5. Conclusions

The problem of treating the missing data in portfolio selection problems is very important in real applications. Our research has been focused on the study of the impact of the presence of missing data in the efficient frontiers of the minimum-variance models. Two methods for treating the missing data are used in order to build a vector and a matrix. Starting from this vector and from the matrix one considers minimum-variance models in which the vector replaces the vector of means and the matrix replaces the covariance matrix. Numerical cases are considered and efficient frontiers of the minimum variance models are displayed and the results are analyzed. The programs are written in MATLAB.

Acknowledgements

This work was supported by a grant of the Romanian National Authority for Scientific Research, CNCS – UEFISCDI, project number PN-II-ID-PCE-2011-3-0908.

REFERENCES

1. ALLISON, P. D., **Missing Data**, Thousand Oaks, CA: Sage Publications, 2001.
2. BARLADI, A. N., C. K. ENDERS, **An Introduction to Modern Missing Data Analyses**. Journal of School Psychology, vol. 48, 2010, pp. 5-37.
3. HOX, J. J., **A Review of Current Software for Handling Missing Data**, Kwantitatieve Methoden, vol. 62, 1999, pp. 123-138.
4. LEDOIT O., M. WOLF, **Improved Estimation of the Covariance Matrix of Stock Returns with an Application to Portfolio Selection**, Journal of Empirical Finance no.10, 2003, pp. 603-621.
5. LITTLE, R. J. A., D. B. RUBIN, **Statistical Analysis with Missing Data**, New York, Wiley, 1987.
6. MARKOWITZ, H. M., **Portfolio Selection**, Journal of Finance. vol. 7, 1952, pp. 77-91.
7. MARKOWITZ, H. M., **Portfolio Selection. Efficient Diversification of Investments**. John Wiley & Sons, Inc., New York, 1959.
8. MCKNIGHT, P. E., K. M. MCKNIGHT, S. SIDANI, A. J. FIGUEREDO, **Missing Data: A Gentle Introduction**, Guilford Press, New York. 2007.
9. PAGE, S., **How to Combine Long and Short Return Histories Efficiently**, Financial Analysts J., Vol. 69, no. 1, 2013, 45-52.
10. RĂDULESCU, M., C. Z. RĂDULESCU, M. TUREK RAHOVEANU, G. ZBĂGANU, **A Portfolio Theory Approach to Fishery Management**, Studies in Informatics and Control, vol. 19(3), 2010, pp. 285-294.
11. RĂDULESCU M., C. Z. RĂDULESCU, M. TUREK RAHOVEANU, **Safety-first and Chance-constrained Production Planning Models for Fish Farms**, 11th WSEAS International Conference on Systems Theory and Scientific Computation (ISTASC 2011), Florence, Italy, (2011), pp. 199-204.
12. RĂDULESCU, C. Z., M. RĂDULESCU, **A Decision Support Tool Based on a Portfolio Selection Model for Crop Planning under Ris**k, Studies in Informatics and Control, vol. 21(4), 2012, pp. 377-382.
13. RUBIN, D. B., **Multiple Imputation for Nonresponse in Surveys**, New York: John Wiley & Sons, Inc., 1987.
14. TAYLOR B., **Developing Portfolio Optimization Models**, http://www.mathworks.com/company/newsletters/articles/developing-portfolio-optimization-models.html
15. SCHAFER, J. L., **Analysis of Incomplete Multivariate Data**, New York: Chapman and Hall, 1997.