Evolutionary operators in memetic algorithm for matrix tri-factorization problem

Rok Hribar[1,2], Gašper Petelin[3], Jurij Šilc[1], Gregor Papa[1,2], and Vida Vukašinović[1]

[1] Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia
rok.hribar@ijs.si
jurij.silc@ijs.si
gregor.papa@ijs.si
vida.vukasinovic@ijs.si
[2] Jožef Stefan International Postgraduate School, Jamova 39, 1000 Ljubljana, Slovenia
[3] Faculty of Computer and Information Science, University of Ljubljana, Večna pot 113, 1000 Ljubljana, Slovenia
gapi.petelin@gmail.com

Abstract

In memetic algorithm, a population based global search technique is used to broadly locate good areas of the search space, while repeated usage of a local search heuristic is employed to locate optimum. Intuitively, evolutionary operators that generate individuals with genetic material inherited from the parents and improved performance ability should be the right option for improved performance of the algorithm in terms of time and solution quality. Evolutionary operators with such properties were devised and used in memetic algorithm for solving multi-objective matrix tri-factorization problem. It was shown, by comparing deterministic naive approach with two variants of memetic algorithm with different level of inheritance, that evolutionary operators do not improve performance in this case. Further analysis showed that even though proposed evolutionary operators inherit high fitness from its parents, local search does not perform well on such offspring which results in poor performance.

Keywords: Memetic algorithms, Non-negative matrix factorization, Multi-objective optimization, Gradient descent

1 Introduction

The level of data generated within different area of our life has drastically increased with the expand of information technology. As a consequence, an interests to extract meaningful information out of collected data significantly grew and the knowledge discovery has become widely studied research area. Within this research, we try to understand data by forming groups of instances, i.e. clusters, where instances in the same cluster are in some sense more similar to each other than the instances in other clusters. The problem studied in this paper is non-negative matrix factorization (NMF) problem which generalizes kernel k-means clustering, bipartite graph k-means clustering and spectral clustering problem [1]. Original NMF factorizes input non-negative matrix R into two non-negative matrices so that $R \approx GQ^T$, where $R \in \mathbb{R}_+^{n\times m}, G \in \mathbb{R}_+^{n\times k}$, and $Q \in \mathbb{R}_+^{m\times k}$. NMF's main objective is clustering of columns of R. While NMF can capture two types of relations, non-negative matrix tri-factorization (NMTF) $R \approx GSQ^T$ can capture more types of information [2]. Both approaches are used for revealing hidden patterns in large real-world datasets and give a good framework for simultaneously clustering the rows and columns of R. NMTF problem can be encountered in image processing, text mining, hyperspectral unmixing and bioinformatics [3, 7, 8]. Additionally, Buono and Pio proved that NMTF has several advantages compared to the original NMF approach [4].

This paper is concerned with the behavior of evolutionary operators used in solving NMTF problem with presence of additional local search strategies. In Evolutionary algorithm (EA), mechanisms inspired by biological evolution such as selection, crossover and mutation influence the evolution of population and implicitly lead the performance of evolutionary search. Mitchell and Holland analyzed promising features of genetic algorithm for its speedup and suggested that crossover in idealized genetic algorithm should create instances with higher fitness [5]. Doerr et al. provided first theoretical proof for usefulness of crossover for non-artificial problem [6]. In this work, we develop a mutation and crossover operator which, applied on matrices, are able to provide solutions with lower objective values and we compare memetic algorithms (with and without suggested crossover) and deterministic naive approach on non-negative matrix tri-factorization problem (NMTF).

2 Multi-objective non-negative matrix tri-factorization problem

The aim of NMTF is to extract insights of intra-relations of some data set. If the intra-relations are expressed by non-negative symmetric matrix R, then by NMTF of the form $R = GSG^T$, where G and S are non-negative matrices of dimensions much smaller than dimension R, some insights how data is clustered and what are relations among those clusters can be provided. Further, in a co-clustering version of NMTF problem a set of matrices R_i needs to be factored using the same G as it is shown in Eq. (1).

$$\forall i : \quad \boxed{\quad R_i \quad} = \boxed{G} \cdot \boxed{S_i} \cdot \boxed{\quad G^T \quad} \tag{1}$$

Here, the columns of G can be interpreted as clusters, while components of S_i can be interpreted as interactions among these clusters.

This problem can be stated as an optimization problem by minimizing relative square error

$$\mathrm{RSE} = \frac{\sum_i \|R_i - GS_iG^T\|_F^2}{\sum_i \|R_i\|_F^2}, \tag{2}$$

where $\|\cdot\|_F$ is the Frobenius norm. Note, that optimal solution has $RSE = 0$, while trivial solution ($G = 0$ and $S_i = 0$) has $RSE = 1$. Given that G matrix is common to all R_i tri-factorizations, all dimensions of S_i are the same. This dimensions' number can be interpreted as the number of clusters and it is not known in advance. In order to assure high ability of data relationships' interpretation, the second objective is to minimize

$$k = \dim S_i. \tag{3}$$

In this respect, RSE minimization ensures the accuracy of tri-factorization and k minimization ensures that the size of representation is as small as possible. Note that the objectives in Eq. (2) and (3) are contradictory because the capacity of GSG^T model grows with k.

3 Naive approach

If k is fixed, RSE can be minimized via gradient descent since RSE from Eq. (2) is a differentiable function of G and S_i. Libraries for automatic differentiation such as `tensorflow`, `theano` or `CNTK` can be used to calculate the gradient of RSE with respect to G and S_i and update them in direction of the gradient. It must be stressed that the second objective k is not differentiable, hence gradient descent can only be used to minimize one objective.

In this work, special version of gradient descent algorithm called Adam [9] is used. This algorithm is well suited for large problems and has two main benefits. One is is the adaptive learning rate control which changes step size during descent in response to changes in gradient magnitude. The second benefit is the use of momentum which prevents oscillations in narrow valleys of the search space and gives the descent an ability to skip shallow local minima. Both traits reduce the number of steps needed to find a local minimum.

The non-negativity constraint encountered in problem definition can be easily fulfilled if absolute value is applied to G and S_i before every RSE calculation. In this way a step of gradient descent going through the bound of the feasible region is effectively bounced back to the feasible region.

A stopping criterion for gradient descent was devised where relative differences in objective function among successive steps are used as an indicator of convergence. Median of the several past differences was found to be a reliable estimate of the pace of convergence.[4] When this pace falls far below previously encountered ones the gradient descent is stopped. Additionally, descent is also stopped if maximum number of steps is exceeded.

Even though k is not differentiable, it is possible to solve the two-objective optimization problem using only gradient descent. In naive approach (NA) Adam is performed for many different k until satisfactory approximation of Pareto front is acquired. However, there are no guidelines how large the desired k might be and the search can be concentrated to a region where k is too small. In such cases, valuable computational time is being wasted.

4 Memetic algorithm

The term memetic algorithm is used to describe a synergetic combination of an evolutionary approach and local improvement procedure. In this work memetic algorithm for above described NMTF problem is developed with the main motivation to benefit from the combination of good hereditary features of EA and efficient Adam, which is used as local improvement procedure. Stopping criteria for Adam are the same as in naive approach.

4.1 Evolutionary algorithm

Standard EA operators are adapted in such a way that offspring inherit good traits from its predecessors, i.e. no matter what is the dimension of S, a comparably low RSE value is ensured. The aim of EA is to provide good starting individuals with various dimension k for further treatment with Adam, which is usually able to further decrease RSE. In this manner evolutionary algorithm is used only to find good initial points for gradient descent which should reduce the computational load to a high degree.

The difference between suggested adapted EA and the classical one is that in the suggested algorithm the evolutionary process starts with a very small initial population which is growing linearly over time. As a selection of individuals for breeding, a tournament selection is used. We privilege individuals, where Adam was successful, hence the criteria to win the tournament is the lowest RSE. Evolutionary search is also slowly switched from exploration to exploitation; in the beginning parents are selected at random, while in the later generations individuals with lower RSE are preferably chosen to become parents. This is accomplished by setting the proportion between the tournament size and the number of individuals in the population constant over all generations.

Crossover operator used in this work combines two parents and produces one offspring. Offspring's G matrix is a concatenation of parents' G matrices along rows, while offspring's S_i matrices are a direct sum of parents' S_i matrices, see Fig. 1 for an illustration. Note, that by crossover operator individuals with enlarged dimension k are obtained but it holds that

$$\mathrm{RSE}(\text{offspring}) \leq 1/2\left(\mathrm{RSE}(\text{parent1}) + \mathrm{RSE}(\text{parent2})\right). \tag{4}$$

In order to prove statement (4) it is sufficient to show that this inequality holds for a single summand in Eq. (2). Let M_1, M_2 be GSG^T-products of the parents, while the offspring's GSG^T-product is $1/2(M_1 + M_2)$ by definition. Using this fact, it follows

$$\begin{aligned} \|R - 1/2(M_1 + M_2)\|^2 &\leq 1/4\left(\|R - M_1\| + \|R - M_2\|\right)^2 && (5)\\ &\leq 1/2\left(\|R - M_1\|^2 + \|R - M_2\|^2\right), && (6) \end{aligned}$$

where in (5) triangle inequality and in (6) the fact that $2xy \leq x^2 + y^2$ was used[5]. Clearly, an offspring inherits comparable low RSE value from its parents.

Mutation operator used in this work either deletes or adds columns to matrix G and corresponding rows and columns to matrices S_i, see Fig. 2 for an illustration. Columns and rows added by mutation are populated with small random values. In case of deletion, columns and rows that contribute the least to the $\|GS_iG^T\|_F$ are chosen. If columns of G are normalized, then by inspecting the smallest values of S_i components it is easy to determine which columns contribute

[4] Mean was found to be too susceptible to outliers which are also encountered during gradient descent.

[5] $2xy \leq x^2 + y^2$ is equivalent to $0 \leq (x - y)^2$

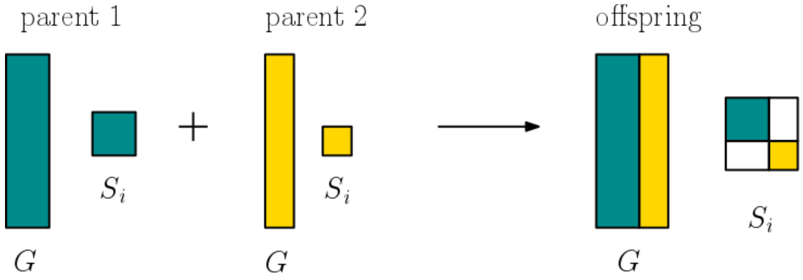

Fig. 1: Crossover of two parents.

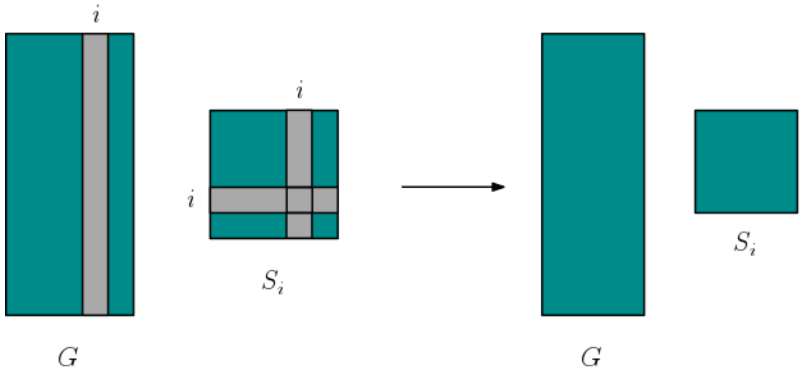

Fig. 2: Mutation where deletion of one column of G was applied.

the least. In this regard, those rows and columns of S_i and corresponding ones in G are deleted that are least significant. By applying a mutation on the selected individual, new offspring with different k and minimally altered RSE is obtained.

New individuals, constructed by crossover and mutation operators, are improved using Adam algorithm at the end of every generation.

Offspring created by mutation or crossover both inherit good RSE from their predecessors. In this regard, the main purpose of proposed evolutionary operators is their ability to construct good initial points for gradient descent from already descended individuals from the population. By using evolutionary operators, information about good clusters and their interactions is able to flow across individuals that have matrices of different dimensions. More importantly, good clusters found in lower dimensions can be passed to higher dimensional individuals for which computational load of using Adam is more pronounced. By passing this information to large individuals, the number of gradient descent steps should be reduced to a large degree.

Two versions of memetic algorithm were used in this work. The M1, performs only mutations and the second, M2, performs both crossovers and mutations.

5 Experiments

A test problem was constructed with 5 matrices R_i of dimension 800 that have a known minimum at RSE $= 0$ and $k = 50$. Approximately $1/3$ of R_i components were non-zero and their magnitude was around one. Three algorithms were used to solve this problem, i.e. M1, M2, and NA. Each algorithm was run 12 times due to limited computational resources.

Basic component of all algorithms in this work is Adam. The parameters of Adam algorithm were manually tuned beforehand, starting with values proposed in the literature [9]. The optimal parameters found were $\alpha = 0.001$, $\beta_1 = 0.9$ and $\beta_2 = 0.99$ where notation from [9] is assumed. In order to ensure reasonable execution time, maximum number of steps for Adam was chosen to be 5000. Convergence criterion was fulfilled when median of last 150 relative differences dropped below one third of the worst seen median. This convergence criterion was devised by observing how gradient descent progresses for this type of problems. Initial experiments showed that when this criterion is fulfilled, there is a small probability that continuing gradient descent will bring further improvements.

Initial matrices G and S_i for all algorithms were chosen randomly where matrix components were drawn from uniform distribution on interval $[0, 0.01]$. In this way initial matrices are close to the trivial solution ($G = 0$ and $S_i = 0$) with RSE ≈ 1. Initial experiments have shown that taking larger initial components of the matrices results to a larger number of steps before gradient descent converges.

Algorithm NA started with $k = 1$ and incremented k by one in each generation. For each k Adam is executed starting from initial random matrices. NA was stopped when it encountered an individual with RSE < 0.01.

Both M1 and M2 started with a population of 4 individuals whose k was chosen from a uniform distribution on set $\{1, 2, \ldots, 7\}$. M1 preforms only mutations, while M2 preforms both crossovers and mutations. For each crossover M2 performs two mutations. The number of columns deleted or added during mutation was drawn from geometric distribution with expected value equal to 3.0. At the end of each generation gradient descent was performed on all new individuals. Crossover of a parent with itself was prevented due to the inability of gradient descent to improve such an offspring. The stopping criterion is the same as for NA which is fulfilled when RSE < 0.01 for some individual in the population.

6 Results

A comparison was done among M1, M2 and NA algorithms with regard to the hypervolume and the number of evaluations. Fig. 3 and 4 depict the distributions of these two indicators. A depiction of Pareto front approximations obtained over all runs by each algorithm is shown if Fig. 5.

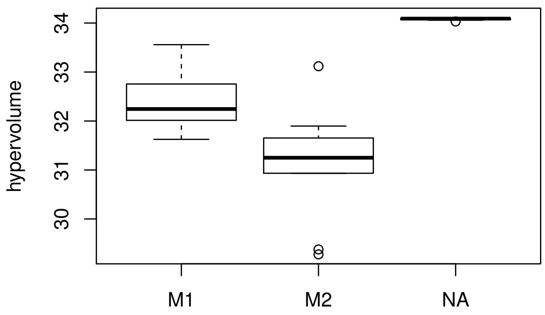

Fig. 3: Box plot of hypervolumes for 12 runs of algorithms M1, M2 and NA.

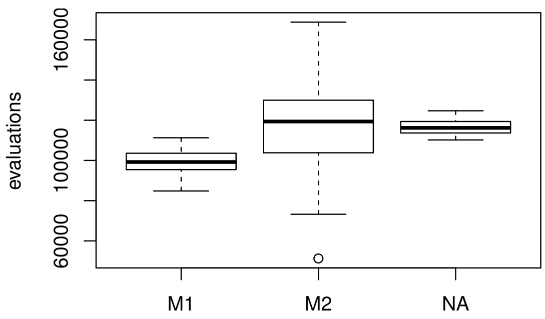

Fig. 4: Box plot of number of evaluations for 12 runs of algorithms M1, M2 and NA.

The number of runs is rather small for statistical comparison, however it can give indications for further work. The data for both comparisons was analyzed using the traditional approach [10] and the Deep Statistical Comparison (DSC) approach [11]. Using the traditional approach, the normality condition (checked with the one-dimensional Kolmogorov-Smirnov test for normality) with regard to the number of evaluations was satisfied, while with regard to the hypervolume was not satisfied. For both comparisons, the homoscedasticity of the variance was checked with Levene's test and for both comparisons the condition was not satisfied. For this reason, the Kruskal-Wallis test was selected as an appropriate omnibus statistical test.

With regard to the hypervolume, there is a statistical significance between the three algorithms, and this significance comes from the difference between the pairs M1, NA and M2, NA. For the same comparison the DSC approach showed that there is a statistical significance between the three algorithms M1, M2 and NA, and they are ranked as 2, 3, and 1, respectively, which can also be seen in Fig. 3.

Regarding the number of evaluations, the Kruskal-Wallis test showed there is a statistical significance between the three algorithms, however the post hoc test according to Dunn showed that the significance comes with regard to the pairs M1, M2 and M1, NA, while there is no difference

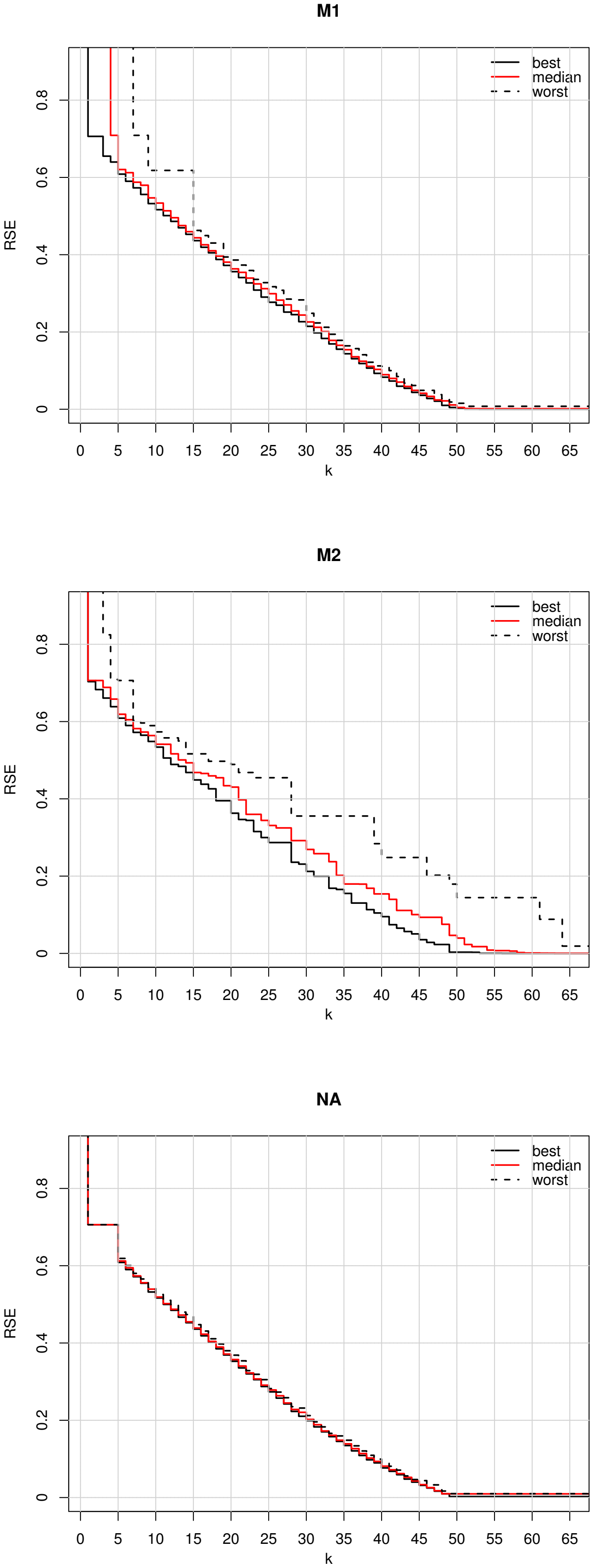

Fig. 5: Summary attainment surfaces of Pareto fronts over 12 runs for algorithms M1, M2 and NA.

between M2, NA. However the traditional approach with Kruskal-Wallis test is made with regard to the medians and not taking into account different standard deviation of the distribution. For this reason, a recently proposed DSC approach was used, where the comparison is made taking into account the whole distribution of the data. The result is that there is a statistical significant difference among M1,M2 and NA, and they are ranked as 1, 2, and 3, respectively, from which it follows that the NA needs most evaluations on average.

The fact that NA finds better quality solutions compared to M1 and M2 is very surprising. Even though evolutionary operators generate initial points with lower values of RSE, it seems that those initial points do not lead gradient descent to good regions. Evidently, starting with low RSE does not guarantee good convergence. This indicates that low RSE should not be the sole trait to be inherited in order to ensure efficient evolutionary operators. To further explore this counterintuitive behavior, the data gathered during optimizations was analyzed. All instances of individuals with $k = 10$ that was generated during M1, M2 or NA was gathered. Such individuals can be produced by crossover or mutation followed by a gradient descent or it can be produced solely by gradient descent starting from random initial point.

Fig. 6 shows the progression of gradient descent for individuals generated by crossover and for randomly generated individuals. Because crossover combines previously optimized parents, the offspring has low initial RSE compared to the random point whose RSE ≈ 1 at the start of gradient descent. When gradient descent is used on individuals created by crossover, RSE steeply falls but shortly after the convergence becomes very slow. It seems that gradient descent enters a region of slow convergence which could indicate a plateau in the optimization landscape. On the other hand, when gradient descent starts from random initial point, the convergence is quite even and no quick changes in steepness are present. It seems that crossover introduces such initial points for gradient descent that are drawn to a plateau with very small gradient.

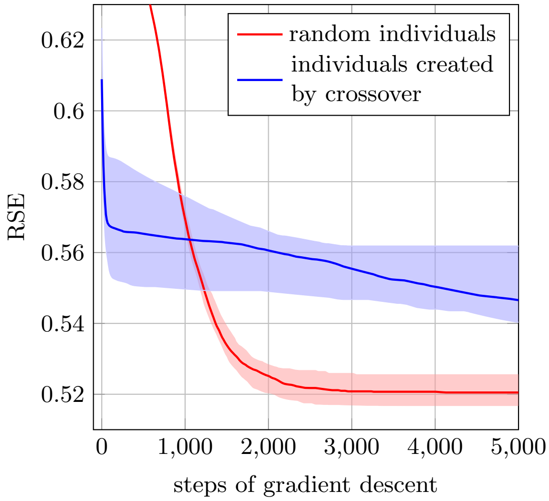

Fig. 6: Progress of gradient descent when initial points are individuals created by crossover compared to the random individuals. All solutions here have $k = 10$ and crossovers were performed using already descended individuals with $k = 3, 4, 5, 6, 7$ ($3 + 7$, $4 + 6$ and $5 + 5$). Full lines are the medians and the shaded areas represent the range of central 66% of runs.

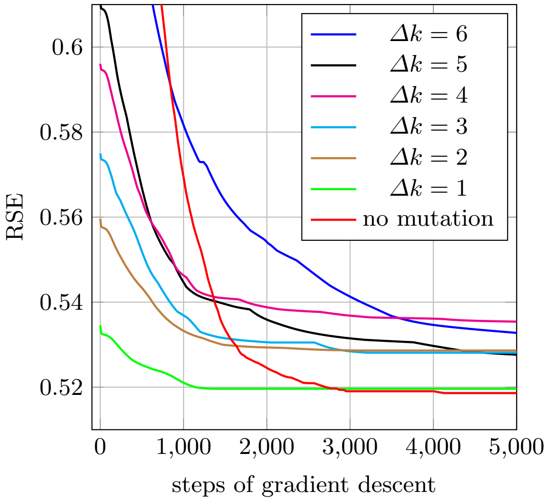

Fig. 7: Median progress of gradient descent when initial points are random individuals compared to mutated individuals. All solutions here have $k = 10$ and mutations were performed using already descended individuals with $k = 4, 5, \ldots, 9$. Quantity Δk tells how many columns were added via mutation and represents the extent to which an individual was mutated.

Fig. 7 shows the progressions of gradient descent for individuals generated by mutation. Only mutations where columns were added were considered. The extent to which an individual is mutated is measured by the number of columns that was added to an individual Δk. Gradient descent performed on mutated individuals converges to higher values of RSE compared to the one performed on random individuals. The only mutation that leads gradient descent to RSE values close to the ones in nonmutated case is the addition of one column ($\Delta k = 1$). The number of steps needed to reach convergence in this case is also approximately three times smaller compared to nonmutated case. This is one explanation why M1 requires less evaluations compared to M2 and NA. The distributions of RSE values after gradient descent for mutated individuals is shown in Fig. 8. Like with crossover, the mutation seems to lead gradient descent to unfavorable regions of the search space.

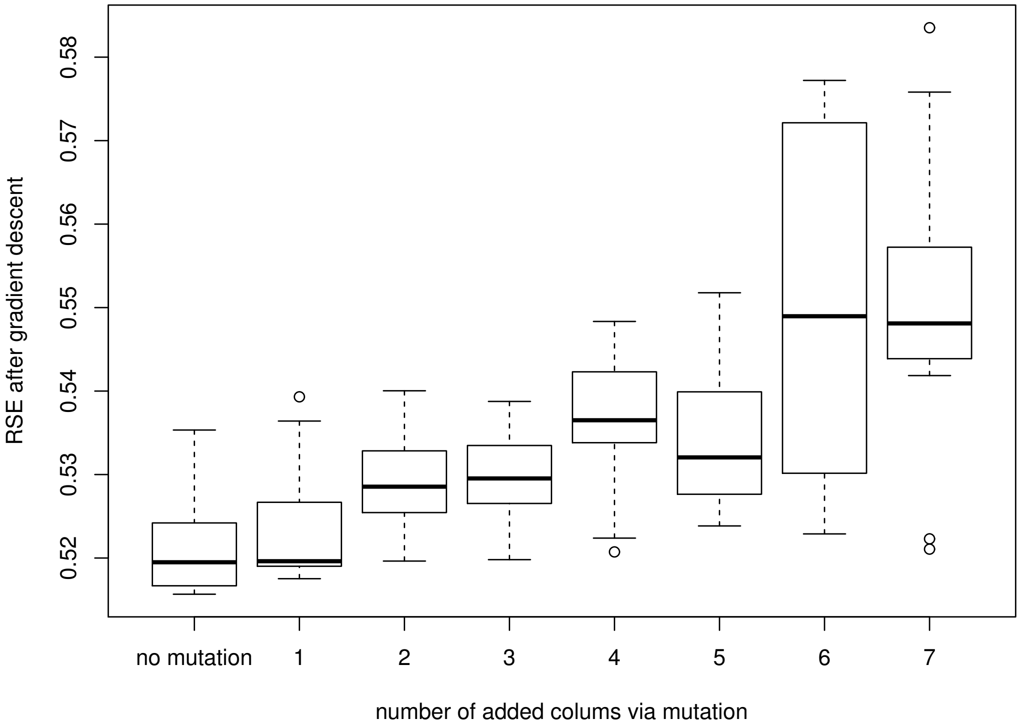

Fig. 8: Distributions of RSE after performing gradient descent on mutated individuals compared to nonmutated ones. Mutation considered here is the addition of specific number of columns (x axis) to matrices G and S_i to already descended individuals. All solutions here have $k = 10$, therefore mutations were performed on already descended individuals with $k = 3, 4, \ldots, 9$.

7 Conclusion and future work

The comparison of the three algorithms showed that the naive approach consistently surpasses memetic algorithms in the quality of solutions. Even though, memetic algorithms can generate individuals which inherit low RSE, these offspring do not lead Adam to solutions of the same quality compared to random starting points. This was further proved by analyzing gradient descent progress of individuals that were generated via crossover, mutation or generated randomly. This can indicate an interesting property of the fitness landscape that needs to be further explored. In particular, Adam in memetic algorithms starts on matrices with few dominant non-zero entries provided by evolutionary operators, while in naive approach Adam starts its search from matrices with entries which are uniformly distributed between $[0, 0.01]$. This might be the reason for low local search performance and might indicate that the choice of evolutionary operators should also be dependent of the local search used and not only on good hereditary features. Also, it seems that inheritance of traits from good individuals of lower dimension somehow guides the gradient descent to regions where gradient has very small magnitude.

Non the less, memetic algorithms proved to be significantly faster than the naive approach.

For the future work we will continue our studies of efficient evolutionary operators for the research problem. Since further testing of evolutionary operators behavior for this problem requires great computational resources, acceleration of gradient descent step will be implemented on GPGPU.

Evolutionary algorithm presented in this work will be further developed so that matrices G have orthogonal columns. This constraint restricts the problem and enforces classical interpretation of clustering. This algorithm will also be adapted so that asymmetric matrices R_i can be used which in consequence introduces several different G matrices of different dimensions, one for each data type. In this regard, the algorithm will be able to perform both classical and soft clustering for possibly heterogeneous data.

Acknowledgements

We would like to thank Tome Eftimov for guidance in statistical comparison of obtained results and Tea Tušar and Peter Korošec for useful conversations. This research was partially supported by Slovenian Research Agency program No. P2-0098, project No. J1-8155 and project No. PR-07606. It is also partially founded by the European Union's Horizon 2020 research and innovation program under Grant Agreement No. 692286.

References

1. Chris Ding, Xiaofeng He, and Horst D Simon. On the equivalence of nonnegative matrix factorization and spectral clustering. In *Proceedings of the 2005 SIAM International Conference on Data Mining*, pages 606–610. SIAM, 2005.
2. Chris Ding. Orthogonal nonnegative matrix tri-factorizations for clustering. In *In SIGKDD*, pages 126–135. Press, 2006.
3. Yulong Pei, Nilanjan Chakraborty, and Katia Sycara. Nonnegative matrix tri-factorization with graph regularization for community detection in social networks. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, pages 2083–2089. AAAI Press, 2015.
4. N. Del Buono and G. Pio. Non-negative matrix tri-factorization for co-clustering: An analysis of the block matrix. *Information Sciences*, 301:13 – 26, 2015.
5. Melanie Mitchell and John H. Holland. When will a genetic algorithm outperform hill climbing? In Stephanie Forrest, editor, *Proceedings of the 5th International Conference on Genetic Algorithms, Urbana-Champaign, IL, USA, June 1993*, page 647. Morgan Kaufmann, 1993.
6. Benjamin Doerr, Edda Happ, and Christian Klein. Crossover can provably be useful in evolutionary computation. *Theoretical Computer Science*, 425:17 – 33, 2012. Theoretical Foundations of Evolutionary Computation.
7. Nicolas Gillis. The why and how of nonnegative matrix factorization. *Regularization, Optimization, Kernels, and Support Vector Machines*, 12(257), 2014.
8. Vladimir Gligorijević, Noël Malod-Dognin, and Nataša Pržulj. Integrative methods for analyzing big data in precision medicine. *Proteomics*, 16(5):741–758, 2016.
9. Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
10. Salvador García, Daniel Molina, Manuel Lozano, and Francisco Herrera. A study on the use of non-parametric tests for analyzing the evolutionary algorithms' behaviour: a case study on the cec'2005 special session on real parameter optimization. *Journal of Heuristics*, 15(6):617, 2009.
11. Tome Eftimov, Peter Korošec, and Barbara Koroušić Seljak. A novel approach to statistical comparison of meta-heuristic stochastic optimization algorithms using deep statistics. *Information Sciences*, 417:186–215, 2017.