MF - Matrix Factorization Techniques for Data Mining

Proposal of Visualization Methods for MF Library

Marinka Zitnik
marinka@zitnik.si
http://helikoid.si/mf
http://github.com/marinkaz/mf

September 13, 2011

Abstract

In this document we propose a number of visualization methods that will be supported in the near future release of the MF - Matrix Factorization Techniques for Data Mining library. Seeing how a matrix decomposition reveals structure in a dataset is sometimes complicated. Each decomposition reveals a different kind of implicit structure and for each decomposition there are four different related ways to intrepret the results. These are: (i) a factor interpretation, (ii) a geometric interpretation, (iii) a component interpretation, (iv) a graph interpretation. Hence, for each dataset, there are many possible avenues for exploration which often provide new insights into the structures implicit in the data. In addition, visualization of some important quality and performance measures alleviates user in choosing critical parameters for the factorization or assessing the quality of the results.

1 Introduction and theory

1.1 Notation

The notation used throughout this document follows the one used in the MF - Matrix Factorization Techniques for Data Mining library. To revise, factorization can be described by an equation of the following forms. Let V be an $n \times m$ (nonnegative) matrix and r a positive integer.

1.1.1 Standard model

Matrix factorization (MF) of V is commonly defined as a pair of matrices such that

$$V \equiv W\,H,$$

where W and H are $n \times r$ and $r \times m$ matrices respectively (with nonnegative entries). Relation $\equiv$ must be understood with respect to some loss function. Common examples of used loss functions are based on Kullback-Leibler divergence or Frobenius norm. Integer r is called factorization rank. As MF approach has been successfully applied to several fields, the column components of W and H can have different names. Some of the used names are:

- W **columns:** metagenes, source, image basis, factors, basis vectors,
- H **columns:** metagene expression profiles, weights, mixture coefficients.

Here, the matrix factor W is noted as basis matrix, factor H as mixture matrix and their columns as basis vectors and mixture coefficients, respectively. The rows or the target matrix V are called features and the columns of the target matrix are named samples.

1.1.2 Nonsmooth model

The nonsmooth model differs from the standard model by defining factorization as

$$V \equiv W\ S(\theta)\ H,$$

where $S(\theta)$ is a $r \times r$ square matrix whose entries depend on an extra parameter, θ. Matrix S is called smoothing matrix, other terminology is applied from the standard model.

1.1.3 Multiple model

The multiple model modifies the standard model by capturing multiple mixture matrices and multiple target matrices. It can be defined as

$$V_1 \equiv W\ H_1, \quad V_2 \equiv W\ H_2,$$

where H_1 is the first mixture matrix and H_2 is the second mixture matrix and the rest of the terminology remains the same as in standard model. The shape of the target matrices V_1 is $n \times p$ and V_2 is of $n \times m$ size, therefore the shape of the first mixture matrix is $r \times p$ and of the second mixture matrix is $r \times m$. Modification of the standard divergence or Euclidean based methods by introducing new matrices is required to support multiple MF algorithms.

1.2 Interpretation

Presented interpretations are mathematically equivalent. However, they provide different views of the hidden structure revealed by the factorization of the target matrix. For each application domain and specific dataset one or more of these are especially representative and natural, but it is usually instructive to consider all interpretations.

1.2.1 Factor interpretation in NMF – hidden sources

The underlying assumption is that the rows of the mixture matrix represent r hidden factors with inherent significance. Therefore features in the target matrix V are the result of mixing these underlying factors in different proportions given by the corresponding entries in basis matrix. Entries in the rows of the basis matrix should be thought of as quantities of underlying factors that have been mixed together to produce the observed sample values of each feature.

1.2.2 Geometric interpreation in NMF – hidden clusters

The rows of target matrix can be interpreted as coordinates in m-dimensional space. After factorization is transformed objects are described by a set of new coordinates, the entries of the corresponding rows of basis matrix with respect to a set of axes given by the r rows of the mixture matrix.

1.2.3 Component interpretation in NMF – hidden processes

Assumption is that each dataset entry is a blend of values from different processes that have contributed to the dataset. The component interpretation separates such contributions. Consider the product of the ith column of basis matrix, possibly the ith entry of the diagonal of smoothing matrix and the ith row of mixture matrix. The target matrix can then be approximated by the pointwise sum of these outer product matrices with i ranging from 1 to r. Each entry in the target matrix is the sum of the corresponding entries in each of the outer product matrices. Outer product matrices can be viewed as layers or components whose pointwise sum is the original target matrix. Each layer corresponding to an outer-product matrix can be examined to see if it can be identified with a known process, noise with some structural component or contains fundamental structure to the modelling task.

1.2.4 Graph interpretation in NMF – hidden connections

Vertices are associated with each of the samples and features of the target matrix, with edges joining vertices of different types, weighted by the matrix entries. We can think of target matrix V as a bipartite graph and the edge weights as measures of the strength of the association between particular sample and feature or the pull between the vertices at each end. Another useful way to think of edge weights is that weights can be regarded as permeability of the edges, meaning that similar objects have many "easy" paths between them. It is possible to consider how close (that is, how similar) two vertices are in terms of commute time – the average "time" it takes to get from one to the other and back over all possible paths between them. There is connection between commute distance metrics and higher-order structure of the fitted factorization model.

A matrix factorization can be thought of as replacing this bipartite graph by a tripartite graph. The first type of vertices corresponds to the features, the second type of vertices corresponds to the samples and the third type of vertices corresponds to the "middle" dimension of the factorization model. The number of vertices in middle dimension is r. We can think of these middle vertices as waystations on paths between the points corresponding to samples and features. For any given feature and sample there are r different paths. Edge weights can be assigned to each of the edges in tripartite graph in the following way:

- w_{ij} for the edges from the feature vertices to the middle vertices,
- h_{jk} for the edges from the middle vertices to the sample vertices.
- In nonsmooth model the effect of the smoothing matrix can be included by multiplying the weight on each edge by the square root of the corresponding element of the diagonal.

Therefore, using tripartite interpretation, the edge v_{ik} from the bipartite version has been smeared across all of the r paths that join the ith feature to the kth sample via any of the middle layer vertices. This relationship is:

$$v_{ik} = \sum_j w_{ij} h_{jk},$$

where this is not an arbitrary decomposition of the v_{ik}.

2 Factor interpretation visualization

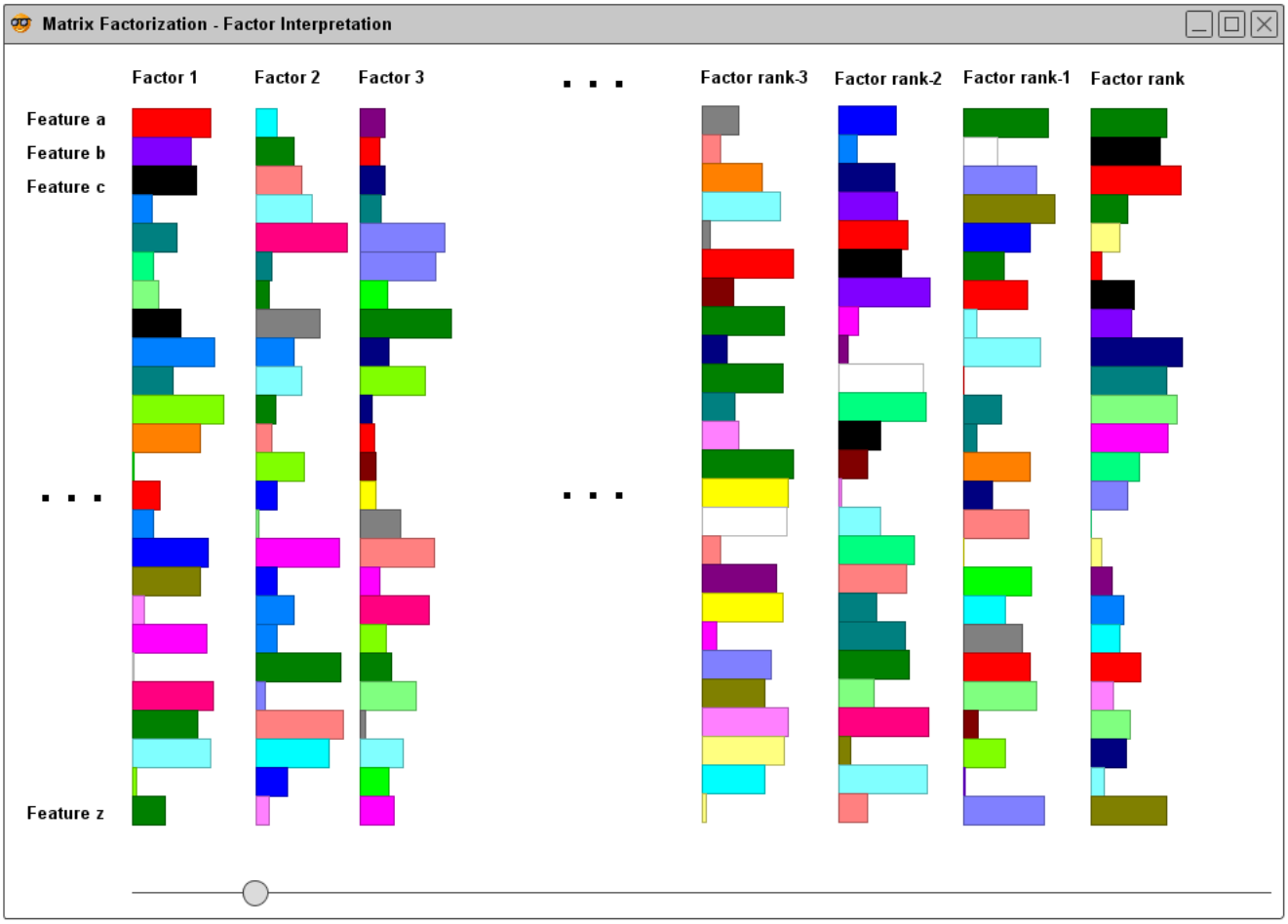

Figure 1: Factor interpretation is the natural way of NMF interpretation. NMF defines a set of factors and a mixing of those to produce the observed data. Factors can be understood as parts and the mixing as addition of nonnegative components. This interpretation is successful when the underlying data are images or signals. The difficulty of the interpretation of produced factors depends on the application domain.

3 Geometric intrepretation visualization

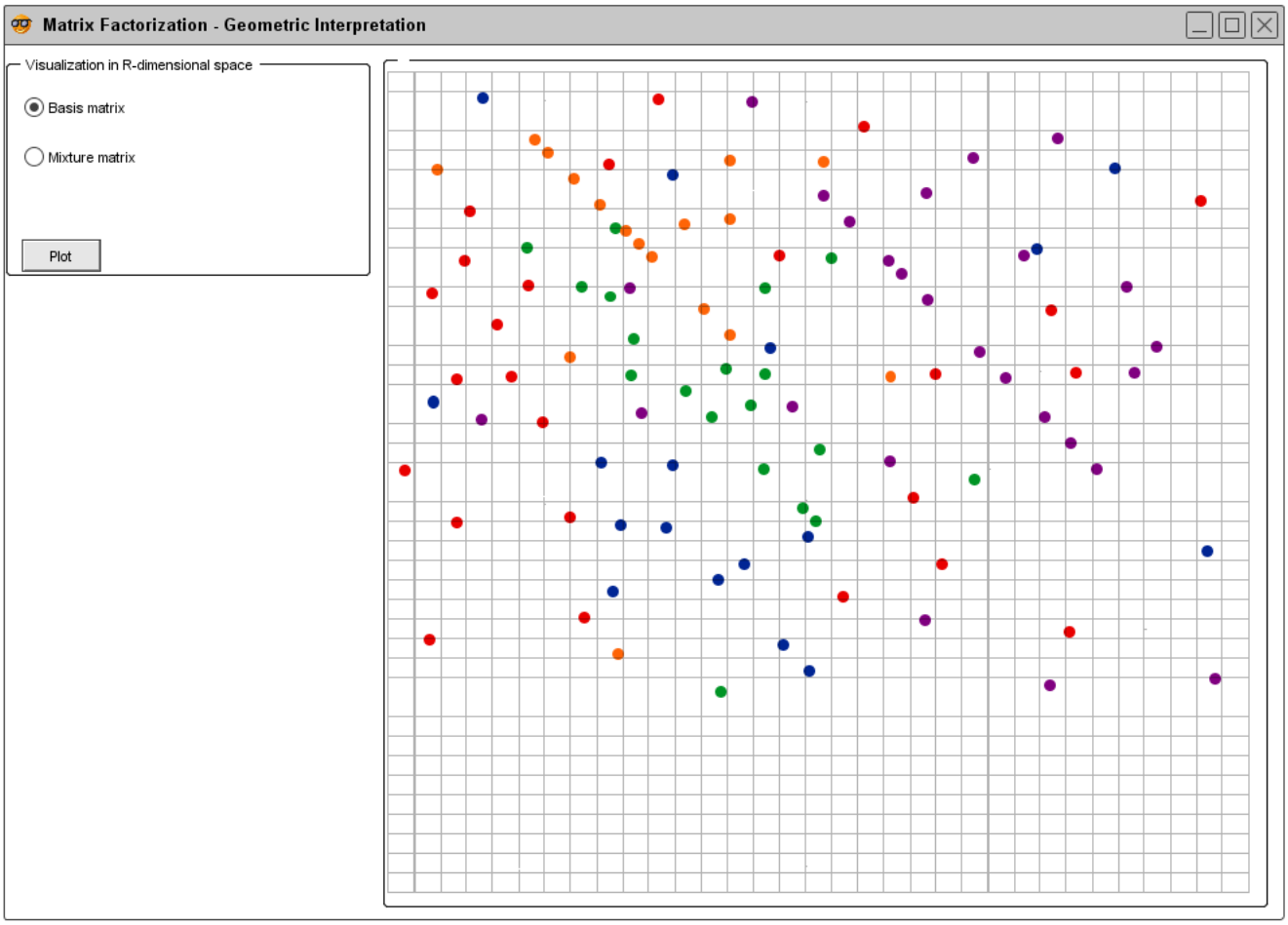

Figure 2: Geometric interpretation. Since neither rows of basis matrix or columns of mixture matrix have no natural interpretation as axes, there is sometimes no natural geometric visualization. However, plotting the matrix entries as coordinates in r-dimensional space can be useful. Two points that are located far apart are dissimilar; two points of close proximity are not necessarily similar. Distance from the origin can be exploited because of the additive nature of NMF. Point is far from the origin if it uses parts with large magnitude entries or mixing coefficients. Point is close to the origin if both parts and mixing coefficients are small.

4 Component intrepretation visualization

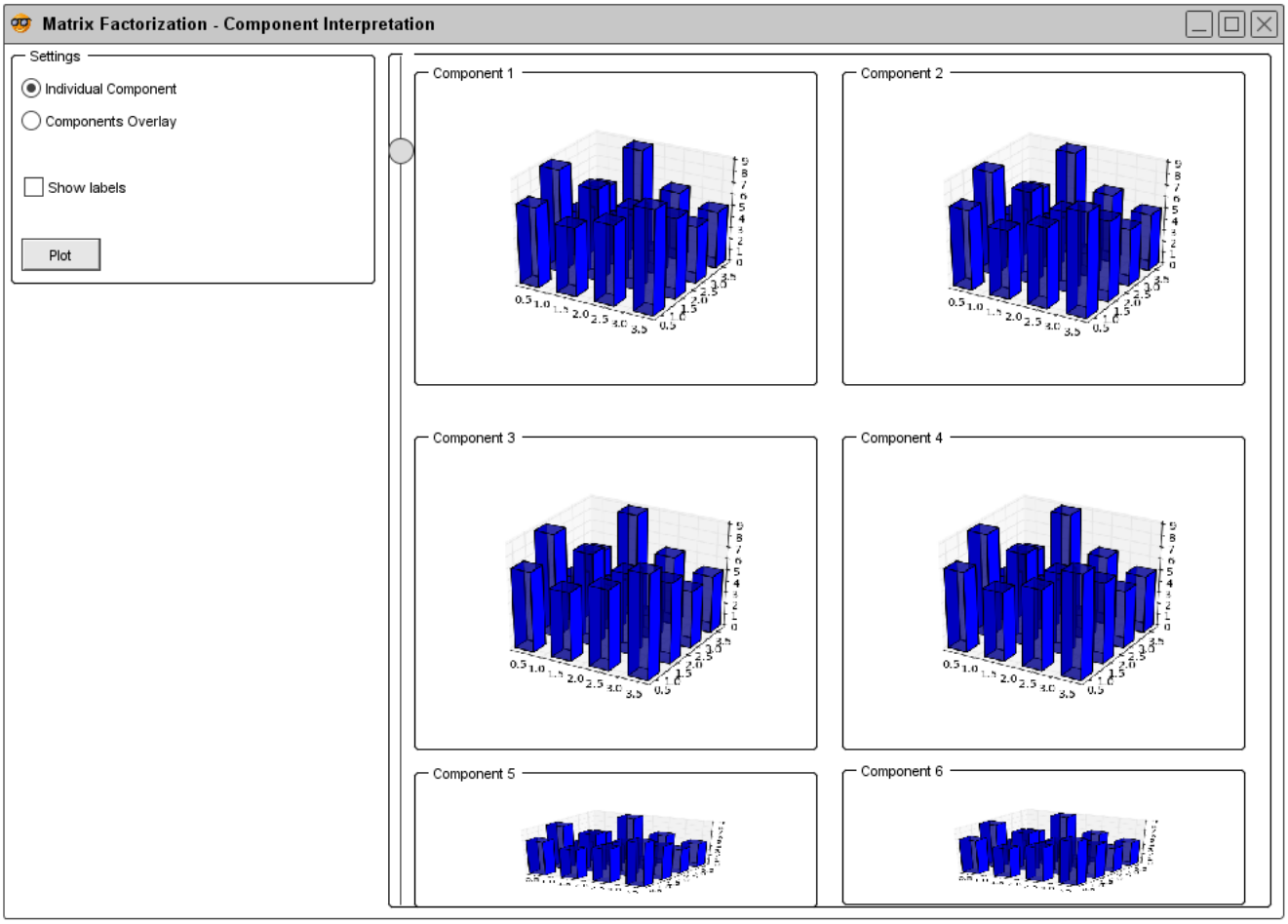

Figure 3: Component interpretation. Individual component visualization or overlay is possible. If component has an interesting structure, it can be seen as a biclustering. These biclusters may be easy to relate to application domain because of nonnegativity. This interpretation main usage is text analysis, microarray analysis to find biclusters of genes and conditions, facial recognition etc.

5 Graph intrepretation visualization

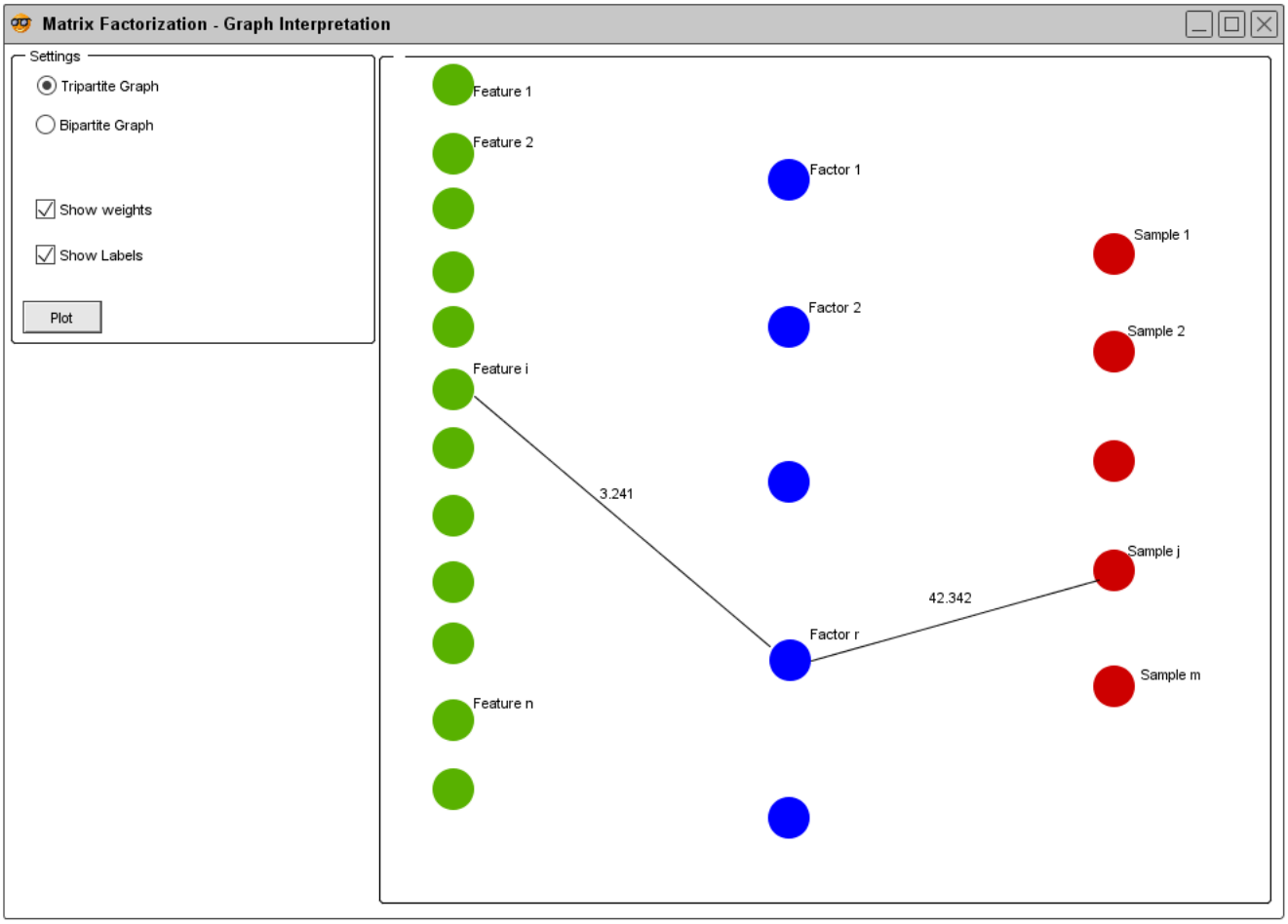

Figure 4: Graph interpretation as a tripartite (or bipartite) graph. One set of nodes corresponds to features, a second set corresponds to samples and a third to r intermediate components. Constraints on the weights of the edges are based on sums of nonnegative quantities. In case of sparse factorization, the graph is sparse as well. Understanding is alleviated because of positive weights.

6 Factorization rank estimation

A list of features supported in scheme 5.

- Factorization rank estimation.
- Overfitting resolution.

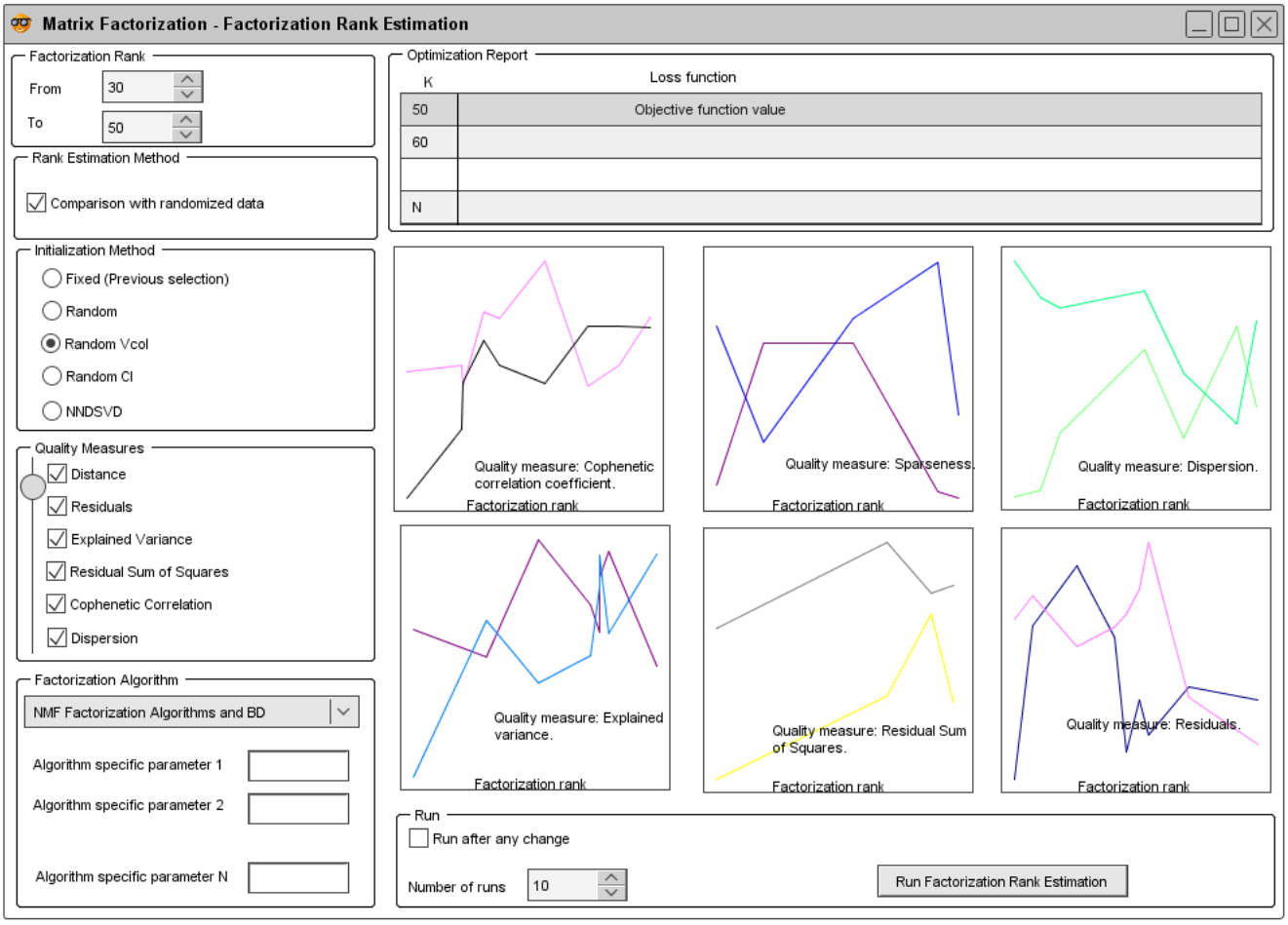

Figure 5: Factorization rank estimation and overfitting resolution. Factorization rank r is a vital parameter to choose in any factorization. The MF library provides several measures proposed to choose the optimal value of r. Although rank estimation can be lengthy process, 30–50 runs are usually sufficient for robust results. Plotting quality measures (e. g. RSS, explained variance, cophenetic correlation coefficient, sparseness, dispersion, residuals) helps in assessing the optimal rank value as suggested by approaches in [Brunet2004], [Hutchins2008] and [Frigyesi2008]. Approach from [Frigyesi2008] may be used to detect overfitting by considering the results for unstructured, randomized original data. Increasing the factorization rank usually leads to decrease in the residuals, as more variables are available to approximate the target data matrix. This gives raise to the overfitting problem. [Frigyesi2008] approach requires computing quality measures for the randomized original data and applying rank estimation procedure for the actual and randomized data.

7 Quality and performance measures tracking across multiple runs visualization

A list of features supported in scheme 6.

- Tracking error (residuals) across single run.
- Tracking quality and performance measures across multiple runs.

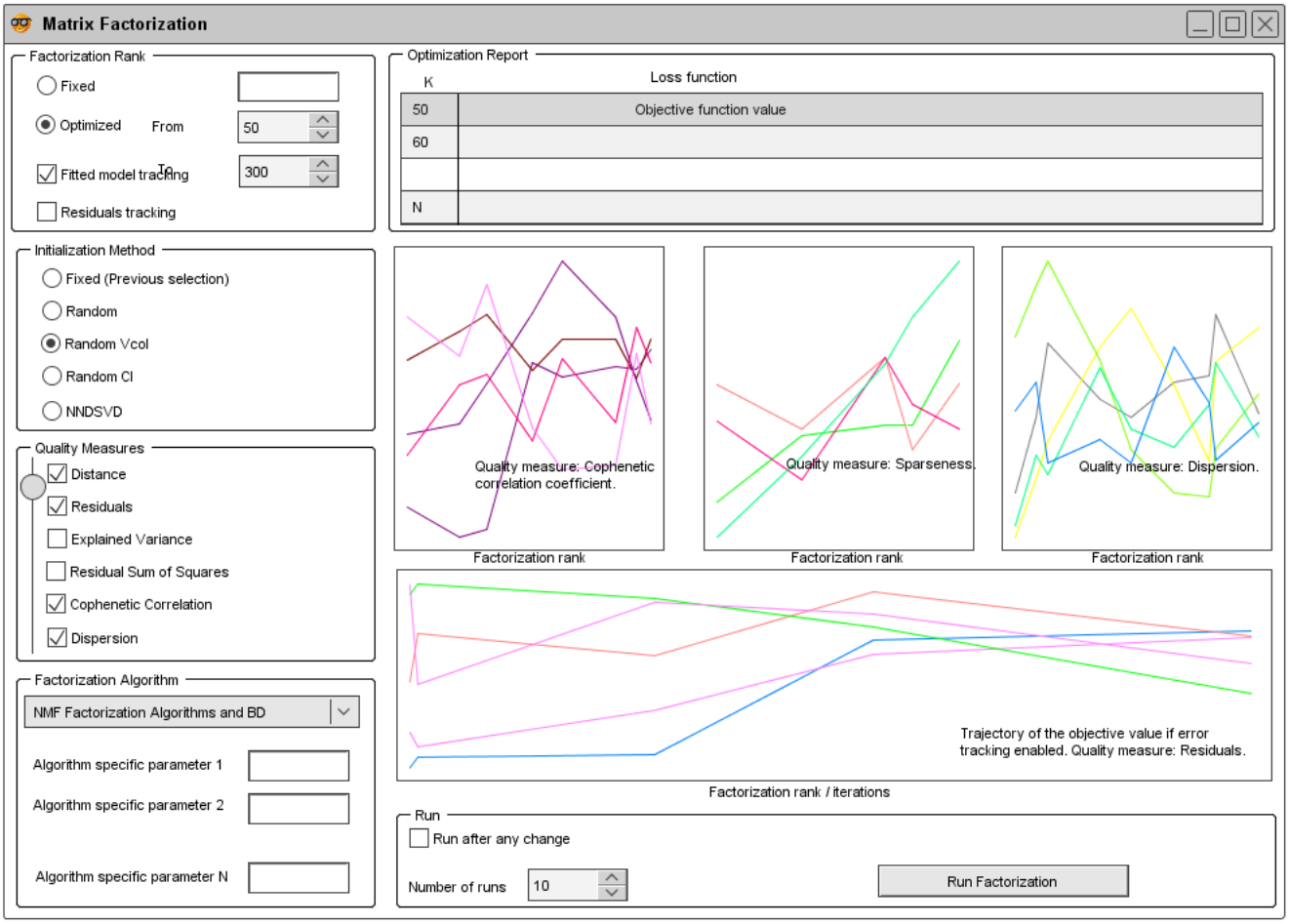

Figure 6: General settings. Quality and performance measures tracking across multiple (single) runs.