

Országos Tudományos Diákköri Konferencia 2023



Non-Negative Matrix Factorization for Many-objective Decision Support

Készítette: Ipkovich Ádám

Témavezető: Dr. Abonyi János

Pannon Egyetem, Mérnöki Kar, Folyamatmérnöki Tanszék

A pályamunka lezárásának dátuma: 2021.04.19

Contents

1	Introduction	2
2	Methods	7
2.1	Curse of dimensionality in Multi-objective decision making	7
2.2	Dimension Reduction	9
2.2.1	Principal component analysis	10
2.2.2	Non-negative matrix factorization	11
2.3	Performance Metrics	13
2.4	Comparing methods – SRD	14
3	Results	15
3.1	Leiden CWTS 2020 Database	15
3.2	Ranking	15
3.2.1	Distinguishing the differences of ranking	17
3.2.2	The correlation structure of objectives and samples	18
3.2.3	NNMF based Leiden ranking	21
4	Discussion and Conclusion	25
4.1	Discussion	25
4.2	Conclusion	29
A	Appendix	36
A.1	Multicriteria analysis - TOPSIS	36
A.2	The interpretation of the principal components	39
A.3	Nomenclature	41
A.3.1	Mathematical Notation	41
A.3.2	Leiden database variables - Scientific Impact Indicators	42
A.3.3	Leiden database variables - Collaboration Indicators	42
A.3.4	Leiden database variables - Gender Indicators	43
A.3.5	Leiden database variables - Open Access Indicators	43

Abstract

This study proposes a method to solve many-objective ranking problems by non-negative matrix factorization-based structural grouping of the objectives. Engineering work contains an abundance of complex multicriteria decision problems where the determination of optimal solutions lies in evaluating many-aspect decisions with various viewpoints. When the number of objectives is too high, many alternative solutions are optimal at least from one objective, so the Pareto front of non-dominated alternatives becomes over-saturated. In multi-objective optimization, as long as there are only three or four conflicting viewpoints, an optimal solution can be determined by finding the proper Pareto front, the theoretical border of non-dominated solutions that outperform the superior members of the solution space. Suppose the number of viewpoints increases and the multi-objective problem evolves into a many-objective optimization task. In that case, the Pareto front becomes over-saturated. This work aims to propose a method that reduces the dimensionality of the problem by providing interpretable groupings of the objectives. As Pareto-ranking is performed in this lower-dimension space, the groups of aggregated objectives ranks similarly. The key idea of the work is that non-negative matrix factorization (NNMF) generates interpretable factors of the objectives, and the Sum of Ranking Differences (SRD) method provides additional information on their similarity. The applicability of the method is illustrated on the ranking of 1176 universities based on 46 variables of the Leiden CWTS 2020 database, as this problem is complex yet easily understandable. The method is compared to principal component analysis (PCA) and sparse non-negative matrix factorization-based solutions. The results illustrate that NNMF and SRD forms a proper method to filter and combine objectives that form interpretable many-objective rankings even in problems with more than 30 criteria.

Keywords: Multicriteria decision analysis, Post-Pareto optimization, non-negative matrix factorization, sum of ranking differences, targeted ranking of universities, selection of best descriptors for a good publication policy

Chapter 1

Introduction

Motivated by Einstein's quote, "Everybody is a genius. But if you judge a fish by its ability to climb a tree, it will live its whole life believing that it is stupid.", *i.e.* everything should be measured against appropriate standards. This quote inherently incorporates the core problem of the present research: the higher the number of objectives is, the more and more samples can become optimal from a particular perspective. The selection of the optimal solutions based on three or more criteria is called multi-objective optimization (MOO).

Many engineering disciplines consist of complex tasks with multiple conflicting parameters to provide an optimal solution to the problem. MOO aims to solve these problems by developing comprehensive decision support systems; although the discipline lacks a general method, several approaches exist. MOO is a fine addition to the skillset of an engineer; an example is material selection based on material properties for additive manufacturing processes [1], resulting in a cleaner and more sustainable production. Additionally, subtractive manufacturing is affected. Xiao *et al.* [2] conducted a research on optimizing the parameter processing in CNC machines, where the cutting parameters are optimized based on the cost and energy need. Furthermore, some of the optimization problems focus on global warming and the climate change aspect of the industry as well. The method proposed by Khan *et al.* [3] reduced specific energy costs by 18.10 %, carbon emission by 22.17 % and product cost by 16.25 %.

A classical approach to multi-objective optimization problems is to find the Pareto front, *i.e.* to determine non-dominated solutions, which are optimal according to a particular objective set. The set of non-dominated solutions is called the Pareto front of the optimization problem, which can be nicely drawn on a scatter plot with two or three objectives (dimensions). One example of these Pareto optimality based methods are the evolutionary algorithms, *e.g.* non-dominated sorting genetic algorithm III (NSGA-III) [4]. However, the performance of most multi-objective evolutionary algorithms significantly

decreases after four conflicting objectives [5], as the number of non-dominated solutions rises exponentially with each objective added. This leads to a state where a considerable portion of the solutions is optimal from only a single objective. The previously described problem is the evolved version of the MOOs, called the *many*-objective optimization problem (MaOO problem).

A possible solution to MaOO is multicriteria decision analysis (MCDA), a set of methods or frameworks for analyzing complex problems. Most methods focus on evaluating the objectives through a set of feasible samples. Many employ aggregation or feature selection, the former groups, and evaluates objectives with an aggregation function such as mean or sum. The latter selects the dominant objectives and ranks accordingly. An example method commonly employed in MCDA is the Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) [6], which utilizes aggregation based ranking by calculating the distance of samples to positive and negative ideal references.

In the world of MCDA, a consensus between the methods is rare, as the same results are obtained only in few (somewhat random) cases. Wątróbski *et al.* [7] analyzed 56 available MCDA methods and arranged the characteristics in a hierarchy, selected rules to build a framework supporting the selection of an MCDA method for a specific decision-making situation. Leoneti [8] did thorough research on the ranking inconsistency phenomena, MCDA methods ranked the data differently, an utmost of 20% of the initial ranging order is predicted correctly. An alternative approach to MCDA is to extend existing methods *e.g.* Shih *et al.* [9] created a robust model for TOPSIS to be used in group decision making. Although fast and efficient, the comprehensive visualization of data and the relationship of objectives remains unsolved.

Several methods have been applied to determine the fair ranking of the institutions, *e.g.* Sziklai [10] performed university ranking by a novel selection method, the Weighted Top Candidate (WTC) method; another algorithm by Kong *et al.* [11] proposed the University Profiling Framework with the usage of deep learning and clustering to compare universities and Vavříková *et al.* [12] employed MCDA on public facilities by assessing research and education.

To address the difficulties of MaOO problems, I suggest the application of non-negative matrix factorization (NNMF) [13]. NNMF reduces the problem to lower dimensions. In the special case of two or three dimensions, the Pareto front can be accessed, providing proper visualization of the data, and grouping the objectives pointing into the same direction together. The method has excellent clustering properties, *e.g.* Li *et al.* [14] solved consensus clustering with NNMF based K-means clustering.

The main profile of NNMF is the field of *recommender systems* [15]. An example is social media applications recommending people based on relationships and preferences [16].

Recommender systems also employ a technique called collaborative filtering for the efficient extraction of information from the Internet. It consists of explicit and implicit rating; the former depends on what other users prefer; the latter focuses on the preferences of the user. Collaborative filtering is a popular research topic, as it provides efficient *recommender systems*, and a possible implementation can be handled by matrix factorization.

An example is the robust collaborative filtering based on non-negative matrix factorization by Zhang *et al.* [17]. Recently, recommender systems based on the blend of matrix factorization and other techniques [18] have already appeared and impacted on the overall performance and efficiency. Other applications of NNMF include face recognition by feature extraction [19] and social network analysis by Snášel *et al.* [20]. NNMF based optimization has also been stimulated with Pareto fronts [21]. Although NNMF is a common method to implement decision-making algorithms, according to our knowledge, it has not been utilized in MaOO type *many*-objective university ranking as thoroughly as this before.

During the ranking of institutions, it should be noted that ranking the universities ideally is impossible. According to the different rankings based on different sets of objectives, the order of universities tend to match at the end and the beginning, but differences in-between depend on the applied method. Many ranking is dependent on an immense amount of objectives, which might be contradicting as well. Some of them may rank in a reversed order compared to the expected order. To understand the underlying structure of the objectives that determine the outcome of the ranking, there is a need for distinguishing conflicting or reversely ranking objectives. The sum of ranking differences [22–24] (SRD) performs exceptionally well in this topic. The SRD is able to determine the optimality of an objective compared to an ideal ranking by calculating the city block (Manhattan) distances between the ranked values and the reference. The optimal objectives are closer to the ideal ranking and obtain an SRD value closer to zero, while the less optimal objectives for ranking are far from the ideal with high SRD values. This method is a jack of all trades as validation, filtering, ranking, and multicriteria analysis can be done through this one method. Lourenço *et al.* [25] proved that, with SRD, a consensus of eight MCDA methods can be met, and in multi-objective optimization, a Pareto front can be established with its help. In the present work, I employ SRD for filtering.

Fig. 1.1 demonstrates the clustered version of recent publications. The figure consists of the keywords (marked by the nodes) of 432 publications queried from Scopus with the search expressions matrix factorization and ranking, multi-objective optimization, decision support, multi objective-, or multicriteria decision making. Two nodes are connected if the keywords cooccur in publications more than or equal to four times. The figure can

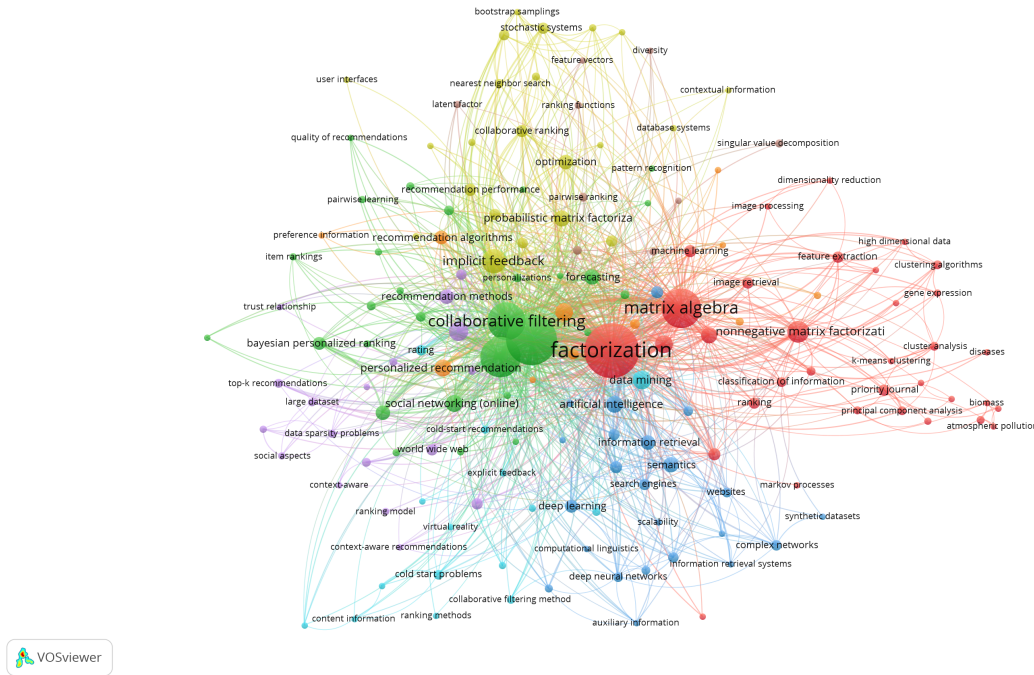


Figure 1.1: **Literature of Matrix factorization and MOO.** The figure shows the cooccurrences of article keywords from 432 publications queried from Scopus. The search expressions/keywords were matrix factorization and ranking, multi-objective optimization, decision support, multi-objective decision-making, or multicriteria. Unique keywords, marked by the nodes, are connected if they cooccur at least four times. Only publications before the middle of December 2020 is included.

be split into groups of clusters indicating more related publication fields. The different colours of the nodes mark them. To the right side, the red group consists of clusters of methods to solve MOOs. The clusters of matrix algebra and factorization stand out the most. The bottom (blue) clusters are related to artificial intelligence with keywords such as deep learning and neural networks. The middle left (green) clusters represent recommendation system related concepts *e.g.* collaborative filters and item rankings.

The goal of applying non-negative matrix factorization is to explore the structure and relation behind the factors and create aggregated variables from which the Pareto front is formed. According to the front, the optimal solutions are ranked to determine university ranking based on the purely bibliographic CWTS Leiden database 2020 [26], which holds 1146 universities and can be ranked according to 46 objectives. In addition, several different techniques (principal component analysis (PCA) [27], NMF, and sparse NMF [28, 29]) are compared. SRD is applied to validate the results.

Many-objective optimizations (MaOOs) are becoming influential to the world of optimization; thus, by using non-negative matrix factorization, I propose a fast and efficient solution to cope with the problems of MaOOs.

As for the contribution of this research, the novelty of the application of NNMF is the following:

- I present the efficiency of the NNMF for the dimension reduction of ranking problems, as the method enforces a non-negativity constraint on the data; thus, conflicting or inverse objectives cannot be grouped (unlikely to other PCA-like traditional dimension reduction techniques). Due to the solely positively aligned objectives, the low-dimensional projection of the samples can be covered by a Pareto front, and the analysis of the group of objectives can be introduced.
- The application of the NNMF groups, not just the objectives, but the method clusters the samples, in this case, the universities, as well. I present how the method can detect the cultural differences of different nationalities by analyzing university ranks. As a result, the reduced dimension ranking problem incorporates all aspects of the original objective space and highlights these differences in the rankings.

The sparse NNMF is provided by the Non-Negative Matrix Factorization Toolbox [30]. All results stated below were implemented in MATLAB.

In the following sections, first, the methodology is discussed, in which I introduce the problem of the classical approach. Then, the dimension reduction techniques, performance metrics and SRD are described. The section afterwards introduces the CWTS Leiden 2020 database and the results of the methods. Lastly, the outcome is discussed and concluding remarks are provided. Supporting information on TOPSIS and the composition of the principal components can be found in the appendix. A nomenclature section is also provided to support the reader in a better understanding of the paper.

Chapter 2

Methods

2.1 Curse of dimensionality in Multi-objective decision making

Let us suppose that an $\mathbf{X}_{N \times n}$ matrix is derived from an arbitrary database, where N is the number of samples, and n is the number of objectives.

$$\mathbf{X}_{N \times n} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \dots & x_{Nn} \end{bmatrix}$$

An objective is defined as $\mathbf{o}_k = [x_{k1}, x_{k2}, \dots, x_{kN}]^T, k = 1, 2, 3, \dots, n$ column vectors, while a data sample is represented by the $\mathbf{x}_i = [x_{i1}, x_{i2}, \dots, x_{in}]^T, i = 1, 2, 3, \dots, N, \mathbf{x}_i \in \mathbf{X}$ row vectors. The objective function $f_k(\mathbf{o}_j), j = 1, 2, 3, \dots, n$ ranks $\mathbf{X}_{N \times n}$ amongst the k -th objective, the result of which is matrix $\mathbf{R}_{N \times n}$

$$\mathbf{R}_{N \times n} = \begin{bmatrix} f_1(\mathbf{o}_1) & f_2(\mathbf{o}_2) & \dots & f_n(\mathbf{o}_n) \end{bmatrix}$$

Multi-objective optimization problems can be described as to find the optimal solution to all objectives:

$$\min \{f_1(\mathbf{o}_k), f_2(\mathbf{o}_k), \dots, f_n(\mathbf{o}_k)\} \quad (2.1)$$

where the number of objective functions is equal to the number of objectives, $k = 1, 2, 3, \dots, n$.

Though there are many methods to minimize objective functions, when several objectives negatively correlate with each other, there is often no proper way to minimize them

at once.

The classical approach to solve Eq. 2.1 is to find the set of Pareto optimal solutions to the problem. If a solution can be found on the theoretical border built by non-dominated solutions called the Pareto front, it is optimal. These solutions cannot be improved beyond the border; any more steps to optimize the objectives shift the position of the solution on the front. However, to find Pareto fronts, the Pareto dominance must be established first by selecting the non-dominated solutions. One solution Pareto dominates [31] ($\mathbf{x}_i \prec_P \mathbf{x}_j$) another, if it is no worse than the other in any objectives, and is strictly better in at least one objective.

$$\forall_k f_k(\mathbf{x}_i) \leq f_k(\mathbf{x}_j) \wedge \exists k : f_k(\mathbf{x}_i) < f_k(\mathbf{x}_j) \quad (2.2)$$

where $k = 1, 2, 3 \dots n$ is the index of the objective functions, $i = 1, 2, 3 \dots N; j = 1, 2, 3 \dots N; i \neq j$

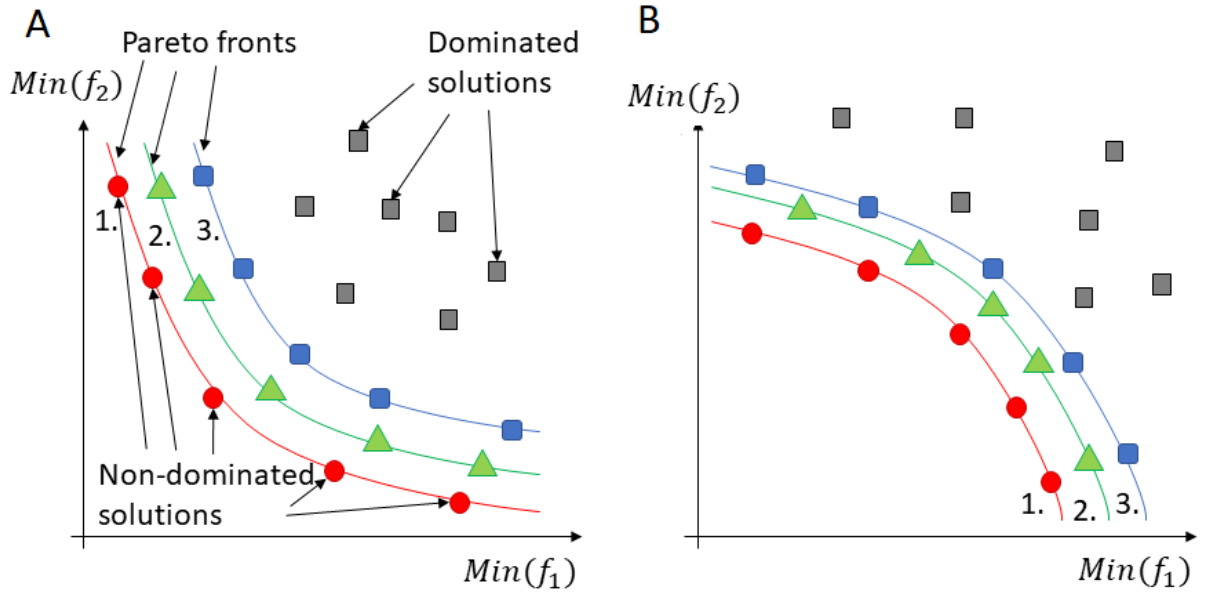


Figure 2.1: **Pareto front representation.** Pareto fronts can be convex (A), concave (B), or mixed. Additionally, the set of solutions can have multiple Pareto fronts, which can be found by ignoring previously determined non-dominated solutions and repeating the search for Pareto front. Think of an onion; peel off the first layer to see the second, which will then be the outermost layer. Clustering the Pareto front is also possible [32,33].

Pareto fronts can take up convex and concave forms or their mixtures as well. The border proposes different trade-offs. Concave has strong, and convex has relaxed trade-offs [34]. There can be more than one Pareto front; in case of discarding the first set of non-dominated solutions, the second set emerges, rinse and repeat until there are not enough solutions to form another Pareto front. Multi-layered Pareto optimality [35] is analogous to an onion. When you remove a layer, you get another one, the next set

of non-dominated solutions; this repeats until you get to the core. The two extrema of Pareto fronts, the concave and convex cases, are illustrated in Figure 2.1. The non-dominated solutions are marked by red circles, while the members of the second and third non-dominated fronts are marked by green triangles and blue squares, respectively. The members of such Pareto fronts can be further clustered for post-Pareto optimality analysis [32, 33].

Finding the Pareto front can be the key to solve the multi-objective optimization problems since it represents the optimal solutions and reduces the number of samples that have to be processed. When the number of objectives increase drastically, the difficulties of MaOO problems [36] are enhanced because of:

- The deterioration of Pareto-dominance based search,
- The escalation in the number of solutions required to estimate the entire Pareto front,
- Lack of visualization

The collective name for the difficulties as mentioned above is the *curse of dimensionality*. When the problem is to deal with several hundreds of objectives, inflation of ranks happens, as more and more solutions obtain the same rank, and hundreds will be firsts, seconds, *etc.* Because of this, the number of non-dominated solutions expand exponentially, which causes the first two difficulties. The reason for the lack of visualization is the high dimensionality.

2.2 Dimension Reduction

Dimension reduction is a way to eliminate the effects of the *curse of dimensionality*. The main focus of dimension reduction techniques is to reduce the number of objectives but retain as much meaningful information as possible.

$$\mathbf{y}_i = g(\mathbf{x}_i) \quad (2.3)$$

where $\mathbf{x}_i = [x_{i1}, x_{i2}, \dots, x_{in}]$ $\mathbf{x}_i \in \mathbf{X}_{N \times n}$, $\mathbf{y}_i = [y_{i1}, y_{i2} \dots y_{ip}]$, $\mathbf{y}_i \in \mathbf{Y}_{N \times p}$, $p < n$. To calculate errors, the original matrix is to be reconstructed to see the error difference between the two. The main focus of dimension reduction is to minimize the difference between the original and reconstructed matrix, in the form the square of the L_2 norm:

$$\min ||\mathbf{X} - \hat{\mathbf{X}}||_2^2 \quad (2.4)$$

where $\hat{\mathbf{X}}$ is the reconstructed matrix from $\mathbf{y}_i$ by an appropriate function.

In the present study, three dimension reduction technique is applied: PCA, NMF, sparse NMF. The essential difference between them is depicted in Fig. 2.2. As the figure represents, PCA (A) generates a hyperplane that covers all solutions; NMF (B) takes two vectors and forces them in between. Notice that PCA is capable of processing negative input as well and that the NMF and sparse NMF only utilizes the first (positive) quarter of the Cartesian coordinate system. Sparse NMF (C) enforces sparsity and forces the solutions on the determined main vectors (or in their close vicinity). Changing the representations from A to C gradually fewer (and) fewer solutions are generated, ensuring the sparseness form among the virtually endless number of solutions.

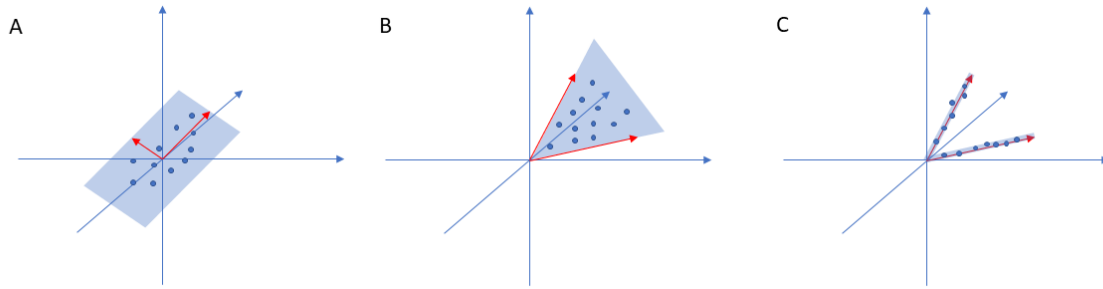


Figure 2.2: **Graphical illustration of PCA(A), NMF(B), and sparse NMF(C).** The latter is two reduced from three to two dimensions. PCA covers the solutions hyperplane, NMF constraints them between two vectors, and sparse NMF forces the solutions to be close to the vectors.

2.2.1 Principal component analysis

Principal component analysis (PCA) [27] is a classic dimension reduction technique. By calculating the eigenvalues of the normalized matrix, the principal components (PC) are determined, and a hyperplane that spans the solutions is created. PCA focuses on maximizing the amount of variance captured by the components sequentially. Therefore, the first explain most of the variance, but this amount is reduced with each new PC. With the help of singular value decomposition [37], the PCs can be determined.

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{M}^T \quad (2.5)$$

where $\mathbf{U}$ is the matrix of eigenvectors, $\mathbf{\Sigma}$ is the matrix of square root of the eigenvalues. The $\mathbf{U}\mathbf{\Sigma}$ product, called the score matrix, contain the principal components. $\mathbf{M}$ is the matrix of loading vectors. The variance of the individual PCs can be determined as:

$$\eta_i = \frac{\Sigma_{ii}^2}{N} \quad (2.6)$$

where η_i is the variance, $i = 1, 2, 3 \dots N$ is the amount of samples.

To reduce the dimensions of the PCs, I discard those that have a negligible percentage; the reduced dimension may result in severe information loss.

2.2.2 Non-negative matrix factorization

Non-negative matrix factorization (NNMF) [13] establishes a low-rank approximation to the input matrix to a low dimensional space. To do so, an input matrix $\mathbf{V}_{N \times n}$ is decomposed as $\mathbf{V} = \mathbf{WH}$, where $\mathbf{W}$ is an $N \times p$ matrix, and $\mathbf{H}$ is a $p \times n$ matrix, and p is significantly less than n ($p \ll n$). $\mathbf{W}$ comprises the sample vectors (the rows of input matrix $\mathbf{V}$), and $\mathbf{H}$ is the factor matrix comprising the objective vectors (the columns of input matrix $\mathbf{V}$). For the elements of $\mathbf{W}$ and $\mathbf{H}$ a non-negativity constraint is introduced, which only allows positively correlating objectives and creates non-orthogonal basis vectors. Additionally, this constraint can be perceived as vectors in between the solutions are placed. The choice of input matrix $\mathbf{V}$ can be arbitrary, depending on the aim of application: the normalized/rescaled form of the sample matrix $\mathbf{X}$ (normalization is required for ensuring non-negative input) provides information on the relationship of the samples, while its ranked form $\mathbf{R}$ describes the relationship of the different ranking objectives. The optimization problem is mathematically formulated as the following:

$$\min_{\mathbf{W}, \mathbf{H}} \|\mathbf{V} - \mathbf{WH}\|_2^2, \text{ subject to } \mathbf{W} \geq 0, \mathbf{H} \geq 0 \quad (2.7)$$

where $\mathbf{V}$ is an $N \times n$ matrix, $\mathbf{W}$ is an $N \times p$ matrix, $\mathbf{H}$ is a $p \times n$ matrix and $p \ll n$.

NNMF is also capable of clustering; coefficients of the objectives in $\mathbf{H}$ each give the similarity of an objective to the basis vectors. $\mathbf{W}$ returns the weights of the data; the clustering happens naturally as one weight element may be more significant than the other concerning their respective row vector of $\mathbf{H}$. As such, the Pareto front placed on the reduced data points is also clustered. To construct $\hat{\mathbf{V}}$, linear combination of the weights and the coefficients is taken.

By reducing the n of objectives of matrix $\mathbf{V}$ to p , the *curse of dimensionality* disappears, yet all objectives are represented. The reduction to such dimension is only available if the residual of the reconstruction is not significant. If $p = 2$ or 3 , then the Pareto fronts are no longer crowded by the non-dominated solutions. The efficiency of the Pareto dominance based search returns and the problem of visualization is solved.

Moving further with the restriction of the matrix factorization, not just non-negativity but constraints of sparsity can be defined. Sparsity measures how many zero values there are in a matrix and constraints the solutions to be placed on, or be in the vicinity of the main vectors defined by non-negativity, similarly depicted in part C of Figure 2.3. In a

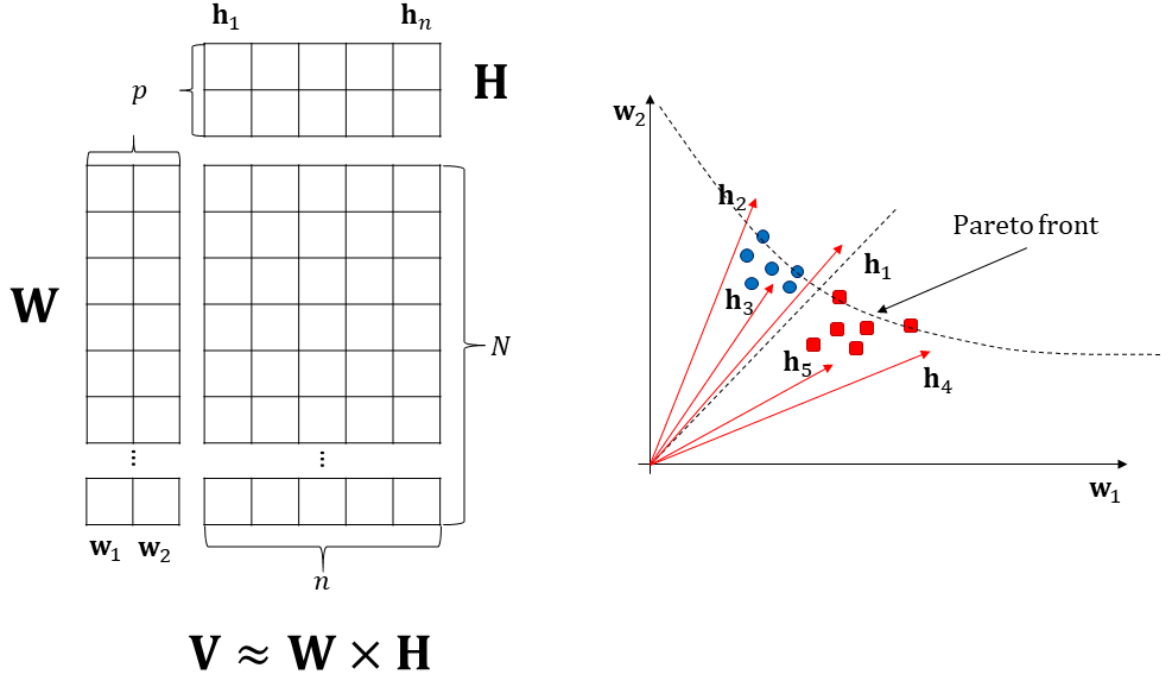


Figure 2.3: **Illustration of non-negative matrix factorization in two dimension.** When the linear combination of $\mathbf{W}_{N \times p}$ and $\mathbf{H}_{p \times n}$ is taken, the approximation ($\hat{\mathbf{V}}$) of $\mathbf{V}_{N \times p}$ is returned. Moreover, the technique has inherent clustering properties, with which the objectives are automatically clustered in $\mathbf{W}$ as are the samples and the Pareto Front. $\mathbf{h}_k$ are the coefficients in relation to the basis vectors, which provides the relation of the objectives.

sparse matrix, the non-zero values can effectively represent the samples, a consequence of which is the emphasis of non-dominated objectives.

To acquire a sparsity level, the usage of sparse non-negative matrix factorization (sparse NNMF) [28, 29] is preferable. The method may enforce sparsity to both $\mathbf{W}$ and $\mathbf{H}$.

$$\min_{\mathbf{W}, \mathbf{H}} \|\mathbf{V} - \mathbf{WH}\|_2^2 + \alpha \|\mathbf{W}\|_2^2 + \beta \sum_{k=1}^n \|\mathbf{h}_k\|_2^2, \quad (2.8)$$

subject to $\mathbf{W} \geq 0, \mathbf{H} \geq 0$

where α and β are parameters to shrink, and regularize the trade-off of sparsity in both $\mathbf{W}$ and $\mathbf{H}$.

Several iterations are to be performed to calibrate sparse NNMF, whose update function based on the alternating least squares is similar to a variant of the Tikhonov regularization, (L_2) two-norm calibration or maintenance without reference samples [38].

2.3 Performance Metrics

To comprehend the over-saturation of the ranking or Pareto fronts, it is recommended to measure how many of the ranks match each other. Relative entropy/Kullback–Leibler divergence (RE) [39] is a descriptive but straightforward measure:

$$RE(D) = \frac{\sum_r \frac{D(r)}{N} \log \frac{D(r)}{N}}{\log \frac{1}{N}} \quad (2.9)$$

where $D(r)$ is number of samples with rank r ,
 $(r = 1, 2, 3 \dots N)$.

A possible misunderstanding maybe that entropy is considered a measure of disorder; it is the reverse from the viewpoint MOO. The less the entropy is, the more similar the values are, reducing the overall efficiency of the ranking and hindering optimal solutions. Conversely, if the relative entropy is high, the ranking is more elaborate and efficient. Naturally, I strive for a higher relative entropy value, though methods in a similar range are not be adequately differentiated from the other even with a higher entropy value.

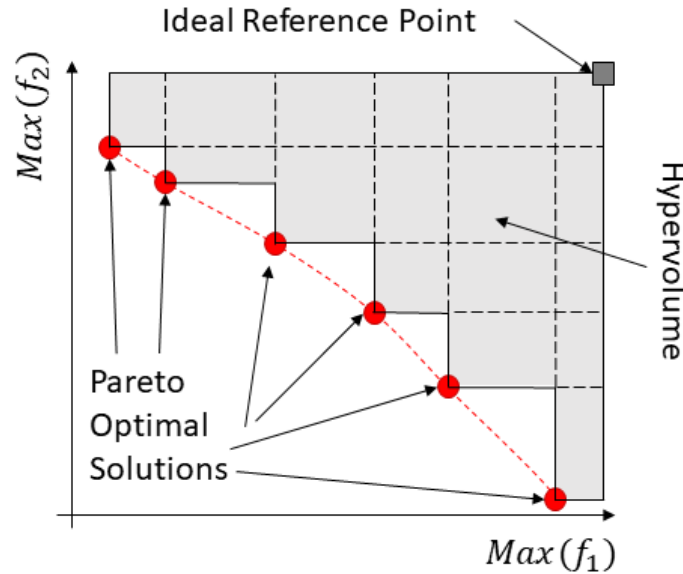


Figure 2.4: **Hypervolume.** The hypervolume indicator measures the area of the Pareto front from a reference. If the reference is the maximum, and the Pareto front consists of maximum values, then a preferable value for hypervolume is close to zero.

Accompanying relative entropy, the hypervolume [40] is to solve the latter problem. The method measures the area between the front and an ideal reference to measure the convergence of the Pareto front. In this publication, the maximum of the objectives is used as reference points. The area, or hypervolume value, between the ideal point and the non-dominated set is to be minimized if the Pareto front is based on the maximum values, and *vice versa*. The algorithm provided by Yi Cao [41] estimates the hypervolume

by Monte Carlo [42] simulation by placing several evenly distributed points to the feature space to approximate the ratio of dominated and the total points, returning the percentage as the measure of performance.

The reason behind employing these measures is for the description of the Pareto front. Higher hypervolume carries the meaning of an inaccurate Pareto front as it can be considered a far cry from the ideal front. The more the number of objectives is, the less value the RE yields, a possible interpretation of which is that the Pareto front becomes over-saturated and demonstrate the *curse of dimensionality*.

2.4 Comparing methods – SRD

The sum of ranking differences [22–24] is a novel statistical method that compares the values with a common reference point; these can be either ideal or a chosen reference if the former is unknown. Though it is not necessary, the input of SRD can be preprocessed. After which, the objectives are ranked, and the (City Block) Manhattan distances are calculated from the reference points. This method is an easy-to-use measure; the smaller the sum is, the closer the sample is to the ideal ranking. The SRD is not a simple distance metric as Spearman footrule; it contains partial ranks (ties) and two validation steps: i) a randomization test and ii) a leave-many-out cross-validation.

If there is no proper ideal ranking, substituting is also possible. Several methods can be applied; these include but not limited to row average, median, maximum, minimum, *etc.* The steps of the SRD can be observed in Figure A.1 along with TOPSIS. Samples and objectives are interchangeable from the viewpoint of the SRD. The proximity of SRD values suggests the relationship of the variables; the closer they are to one another, the more similar they are. Having distance between SRD values provides information on the difference of the variables. Additionally, SRD provides a consensus of eight MOO algorithm by Lourenço and Lebensztein [25].

Chapter 3

Results

3.1 Leiden CWTS 2020 Database

I used the CWTS Leiden Ranking 2020 database [26], which is based exclusively on bibliographic data from the Web of Science database evaluating 1176 universities worldwide based on 46 objectives. The universities are selected from the Web of Science database based on the number of their indexed publications between 2015 and 2018. It is known that Web of Science, *ab ovo*, is biased toward the English speaking scientific community, the start of some rankings based on their data mainly includes English speaking countries.

I applied two different methods for the preprocessing of the dataset: the min-max scaled (0,1), called the scaled input data and the one called the ranked input form, ranked with tied rank, if the ranks are tied, their mean is given to each sample (university).

The names of the objectives are shortened to the combination of a letter and number, where the letter indicates the type of the indicator, *e.g.* scientific impact indicator, while the number describes the variable. For example, S2 is the abbreviation of the second of the scientific impact variables, which is 'the total number of citations of the publications of a university'. The list of abbreviations is provided in the appendix section.

3.2 Ranking

Coming to the topic of many-objective ranking, several straightforward questions can arise in those interested: How to rank fairly? By being naïve and in the name of equity considering all institutions firsts? By taking into consideration all the positive and negative aspects of a university? From the viewpoint of MaOO, if all solutions are optimal solutions, nothing has been done in terms of optimization. However, if all objectives are incorporated to provide an agglomerated ranking, all solutions might be good at something. In that case, everyone is the first, and yet again, I have nothing to show for ranking.

To present the problem, I have created Fig. 3.1 by selecting the combination of two objectives with the highest relative entropy (calculated according to Equation 2.9), providing the most different ranks and hence, the most easily interpretable solution. Therefore, the ranking is only based on the total number of open access publications (O2) and the number of green open access publications (O6), which are available to read without the journal. Without considering the other 44 objectives, the two yields a narrow Pareto front and a decent relative entropy value ($RE(D) = 0.7925$).

These two objectives are positively correlated with each other. When the number of publications increases, more and more scientists decide to self-archive their publications. These publications are accessible to everyone; such a publication may target more researchers than those with fees. If more researchers are familiar with the work of the particular author, the higher the chance of collaboration or citing/referencing is. Even though these objectives have the most distinctive ranking, a contradiction is proposed as one of the most correlated objective pair is chosen. The Pareto fronts are visually separated. The non-dominated solutions in decreasing order of optimality: Harvard University, University College of London, University of Oxford and John Hopkins University

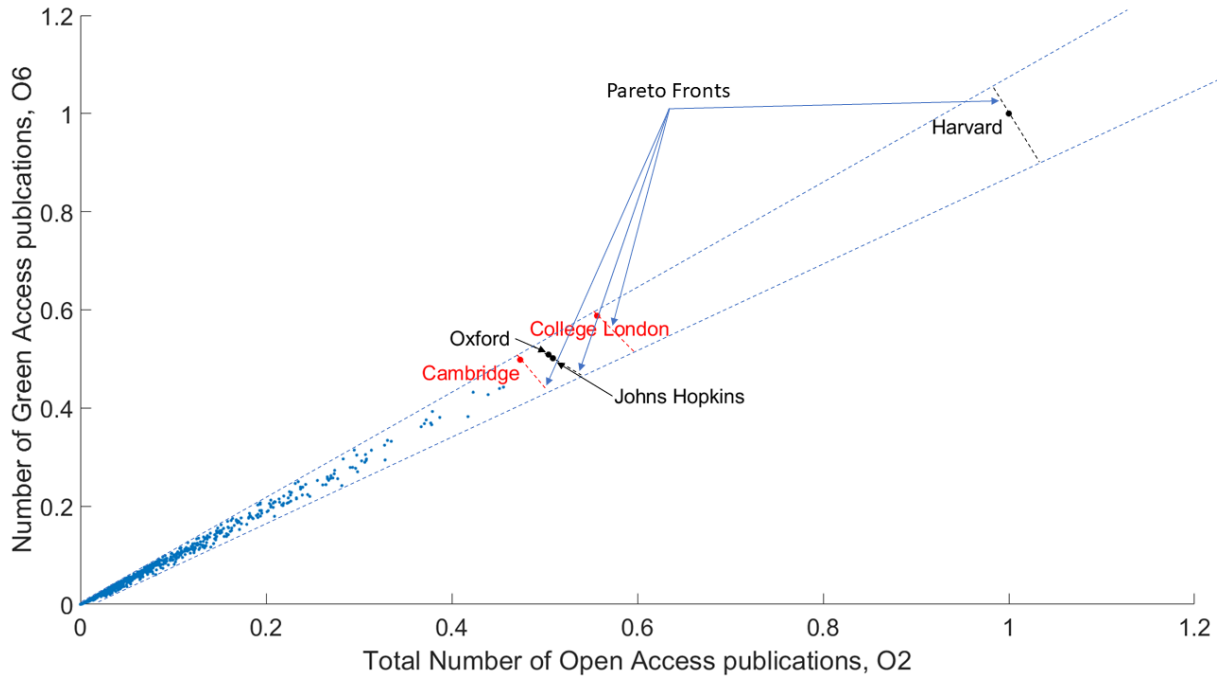


Figure 3.1: **Pareto Front of the two objectives with the highest entropy.** The Pareto fronts only take account of the non-dominated solutions that are optimal from the viewpoint of the 'total number of open access publications' (O2) and the 'number of green access publications' (O6). Coincidentally, the chosen objectives are one of the most correlating pairs. The other 44 objectives have no input to the ranking. In the case of taking another two dominant objectives, I would obtain different non-dominated solutions.

Table 3.1: **SRD without preprocessing.** The worst objectives are proportional indicators; short distance collaboration (C10), male (G7) and female (G8) authorships in proportion to the total number of authorships and the female authorships (G9) in proportion to known gender.

Rank	Worst Obj.	SRD value	Best Obj.	SRD value
1	O13	78.57	G1	8.17
2	G7	78.0	S7	11.28
3	G8	70.3	S2	11.36
4	G9	68.37	S1	11.81
5	C10	67.82	C1	11.84

(the latter two is on the same front) and the University of Cambridge.

On the other hand, when all objectives are used as inputs, the ranking inflates, and the *curse of dimensionality* appears, resulting in all solutions being placed on the first Pareto front. In this case, the relative entropy is zero. The hardship of optimizing *many*-objective problems is that the less the entropy is, the more similar the values are, and differentiating between them becomes impossible. The question is, how it is possible to determine the best when all objectives are ranked first.

3.2.1 Distinguishing the differences of ranking

In the present work, the SRD is the perfect tool to discover the relationship of the different objectives and hence, underpin, validate and interpret the results of the NNMF-based solution. The application of the SRD provides a list of objectives that rank close to the ideal rank, objectives that rank randomly and objectives ranking reversely to the ideal rank.

Presumably a university is better, if publicizes more, receives more citations, *etc.* The algorithm is extremely straight forward and interpretable, *i.e.* the reversely ranked objectives turned out by the SRD analysis.

Figure 3.2 illustrates the results of the SRD analysis. The information is provided by the position/grouping of lines rather than their length. The solid black line indicates a cumulative Gauss approximation (Normal Distribution) of the ranking, its value is indicated on the second vertical axis, the cumulative random frequencies of SRD vectors (CrFr% of SRD(RndV)). The objectives, whose positions are close to zero rank are similar to the golden/ideal rank. The objectives, whose lines are intersected by the probability curve (between XX1 and XX19) cannot be distinguished from random ranking. The objectives to the right side of the line XX19 tend to rank reversely to the ideal rank.

The 'wrong' objectives that are located on the right side of the XX1 line (randomly or reversely ranking indices) and, in a similar manner, the 'good' objectives are on the left

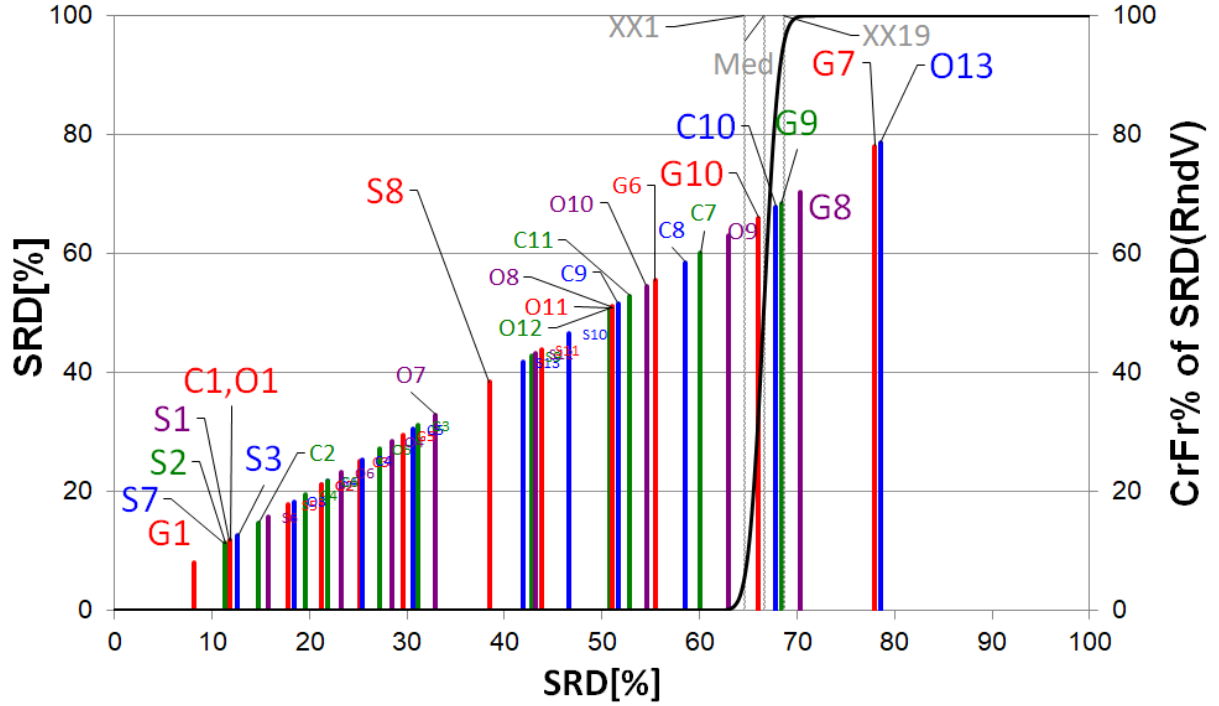


Figure 3.2: **Representation of the objectives by SRD.** The SRD range between 0 and 8 is empty, i.e. the total number of authorships (G1) is close to the ideal/gold reference. The range between the lines XX1 (5 %) and XX19 (95 %) shows the random ranking. Objectives to the right side of line of XX19 rank reversely (G7-G9, O13). The solid line on the second vertical axis indicates the cumulative random frequencies of SRD vectors (CrFr % of SRD (RndV)).

hand side of XX1 line. The first five of each is provided in Table 3.1.

3.2.2 The correlation structure of objectives and samples

The proposed dimension reduction-based technique is capable of solving the problem of the inflated ranks of many objective ranking. The main idea is to reduce the number of dimensions, simultaneously keeping as much information of the original ranks as possible. This way, not just an aggregated information of the numerous objectives serves as the basis of the ranking, but the subtle differences are reflected in the low-dimensional values.

As many dimension reduction methods diminish the *curse of dimensionality*, the choice of the type of information preserved and its optimization is not trivial. In the present work, the performance of NMF and sparse NMF are presented, but in order to provide a benchmark basis for the performance of dimension reduction, as well as present the correlation structure of the objectives and samples, the traditional PCA is applied, as well.

During the PCA analysis, every principal component, but the first two, was discarded to reduce dimensions, explaining 65.87% of the total data variance. I determined the

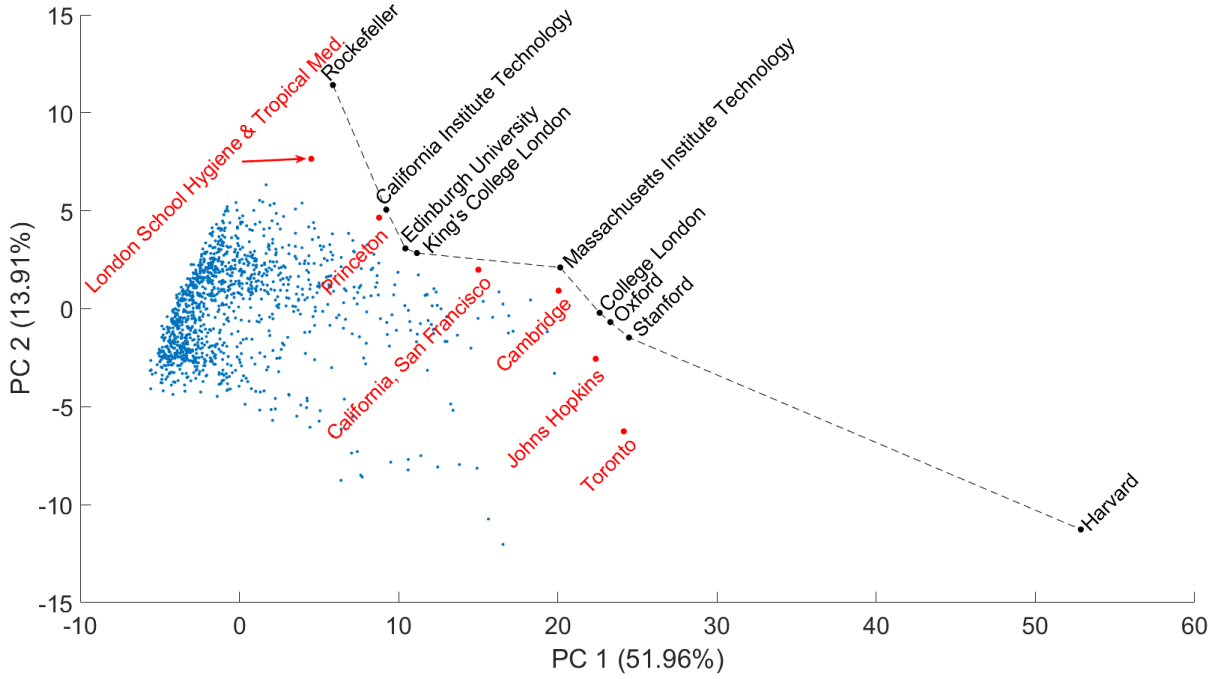


Figure 3.3: **PCA based Pareto front.** In the first PC, objectives that focus on quantity are dominant, while the second PC promotes the high proportion of publications that perform well in scientific indicators, open access and collaborations. Rockefeller is a small private university but has outstanding proportion objective values compared to other more robust Universities, while the other end is Harvard, where resources are available to create an immense amount of publications. Note that only universities from the Anglosphere are on the first two Pareto fronts.

Pareto fronts of the objectives in this reduced objective space, as presented in Figure 3.3. The relative entropy of the PCA based ranking is $RE(D) = 0.6072$. The less the entropy is, the more similar the ranks are, reducing the efficiency of the ranking. Now the question is what drives the better evaluation in this reduced objective space. Figure 3.4 depicts the correlation structure of the objectives on a biplot in the two-dimensional space (samples and objectives are displayed as points and vectors, respectively). The institutions that perform better in the first PC (51.96% explained variance) have a high number of notable publications, reflecting the quantity aspect, while the second PC (13.91% explained variance) is having a high proportion of publications with a considerable scientific impact, open access and collaboration, which collects the quality aspects of the ranking.

The coefficients of the objectives (represented by loading vectors) form six major clusters in Figure 3.4. The first two PCs deal with the quantity (PC1) and proportional (PC2) aspects of the ranking. The massive cluster on and in the vicinity of the horizontal axis contains the quantity variables of different indicators from which PC 1 is mainly composed of. PC2 summarizes the proportional components. It is positively influenced by G7 and G8, the number of male and female authorships as a proportion of a university's

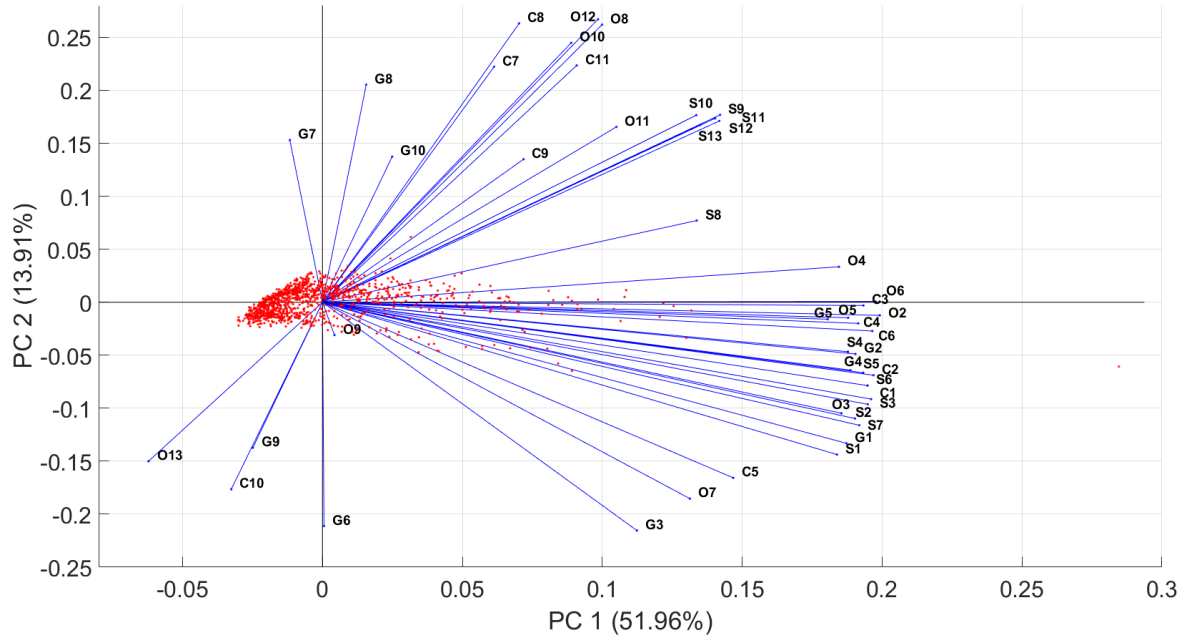


Figure 3.4: **Relation of objectives by PCA.** The biplot depicts both scores and loading, though the scores are scaled differently to Figure 3.3. The axes split the objectives; the vertical axis distinguishes the effects of the objectives either pull back or benefit the samples, in this case, universities. The horizontal axis splits into the objectives as to whether the proportion of the quantity is examined. Six clusters can be observed, the two drawback clusters, the proportion of collaboration indicators, the proportion of scientific indicators and the quantity publications of all indicators. There is an outlier objective (S8), which is more or less in the middle of SRD values without preprocessing. The drawbacks, when combined with positive factors, still pull samples away from the Pareto front.

total number of authorships, respectively, and G10, the number of female authorships as a proportion of a university's number of male and female authorships. The drawbacks point downwards; these are the objectives that disadvantageously affect the rank of an institution. These form two major clusters, the first one incorporates the unknown gender (G6), the male to female ratio (G9), the proportion of unknown open access (O13) and the proportion of short distance collaboration (C10) variables, while the members of the second cluster are the quantity equivalent of the latter three objectives from the first cluster: number of short geographical distance cooperation (C5), the number of publications of which the open access status is unknown and the number of authorships of a university for which the gender is unknown (G3). Short distance collaboration decreases the rank (and, in this sense, the judgement) of an institution; a possible explanation is provided by an idea as old as the New Testament: "Truly, I say to you, no prophet is acceptable in his hometown." Luke 4.24. Unknown open access publications cannot be appropriately cited; these have no DOI number and are generally low-level contributions to obscure

'predatory' journals. If it is unknown to the other scientists, it is as if the publication never existed. From the distribution of the objectives between the two PCs, it can be deduced that the high quantity is not sufficient to provide a better rank. The quality of performance is equally important as well.

After discussing the main driving clusters of the first two PCs, the variables in the positive quarter of the Cartesian coordinate system are left unexplained. These variables form two major clusters, and there is an outlier as well, S8, the average number of citations of the publications of a university. The first cluster is the proportion of collaborations (C7-C11) and green open access publications (O12) except for short collaboration, which almost inversely proportional to this cluster. By publishing free information, the university reaches a vivid variation of audience. In other words, more ideas may be generated, and if a collaboration happens, the rank of the institution might benefit significantly from the accredited new publications. The second cluster consists of the proportion of the most cited publications of universities (S9-S13). The following statements might be trivial, but the cluster gathers descriptors, which proposes that the less impact the publications have, the less the performance of an institution is. The proportion of scientific indicator variables (S8-S13) show that innovation is one of the central roles of academies; the ranking is heavily influenced by publications with higher impact and the average number of publications.

3.2.3 NNMF based Leiden ranking

With non-negative matrix factorization, both the original and the ranked data is examined. As can be seen in Figure 3.5, when the scaled input matrix is processed, the method clusters the universities into institutions under Asian (mainly Chinese) and Western cultural influence. The cluster of red stars and brown diamonds on the left side contains Chinese and Taiwanese institutions. The South Korean universities to the right appear first as a subcluster of the second bigger one. Under that are Iranian institutions, which are close to the lower-ranked Pareto fronts, based on their place in the figure, these universities can be conceived as a mixture of influence from both Western and Eastern style institutions. The western-influenced institutions cover the remaining space, and some nations create subclusters, *i.e.* Japan. Germany and the United States cannot be distinguished as well as the previous nations, as they scatter all over the cluster. Some academies, such as the Royal Holloway, University of London take up a much closer position to the left cluster; this can be appropriated to the openness to take in lots of students with cultural differences. Brandeis University focuses on liberal arts and has a wide variety of cultural studies. The Pareto fronts are clustered to provide a vivid multi-cultural ranking, where nationality takes precedence. In Figure 3.5 the upper section of the first

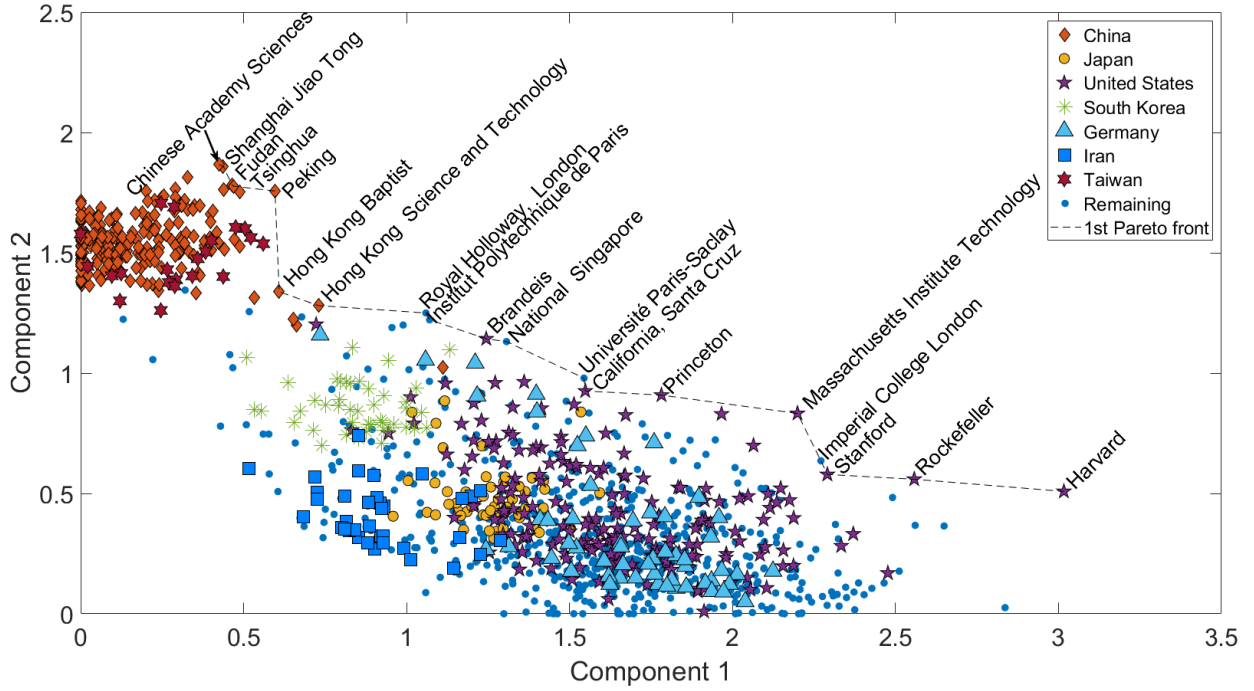


Figure 3.5: **NNMF based Pareto front with scaled data as input.** To rank the institutions, the scaled data is provided to NNMF; The method clusters the institutions to Eastern and Western cultures in education. The Chinese and Taiwanese Universities are put in a cluster on the left; the European, American and the other academies can be found on the right-side cluster. South Korean and Iranian universities connect the two dominant scientific circles.

Pareto front consists of Chinese institutions; the part of the front that is in the second cluster, not only British and American universities appear, but French and Singaporean as well. The relative entropy of the result is $RE(D) = 0.5131$.

As the ranked matrix is chosen as the input of NNMF, it does not provide information on the institutions but the ranking and the relation of ranks and objectives of the Universities. The biplot in Figure 3.6 describes these relationships. In a clockwise manner, the clusters are the following: near the second component, objectives deemed the worst by SRD can be found. In the next cluster, it can be deduced that, according to NNMF, the proportion of international collaboration is connected to open access publications; they go hand in hand. Various underlying reasons can be derived: open access publications may promote more international collaborations, or some international funding opportunities promote collaboration and open access as performance indicators. The proportional collaboration variables, short distance, long-distance, and industry, have a much closer relationship than in the PCA. S10 is an outlier in this representation but separates the proportional and quantitative variables similar to the PCA-based projection. The cluster under S10, consisting of G9, S9 and S11-13 suggests that the male to female ratio

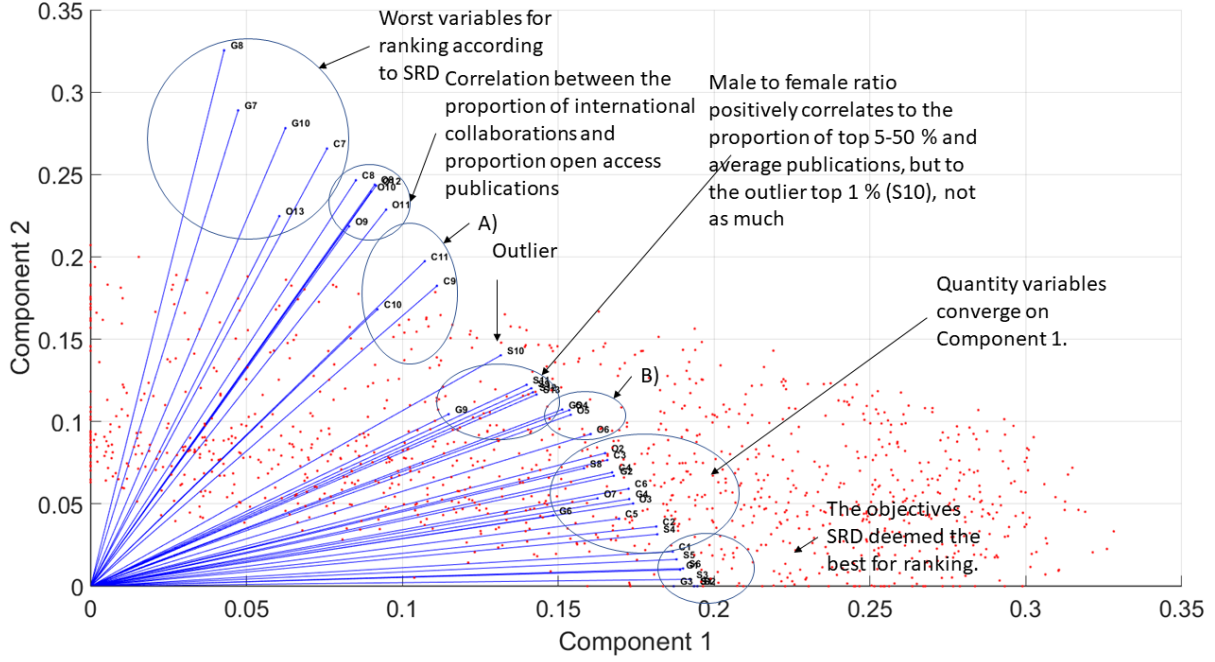


Figure 3.6: **Coefficients of the ranking by NNMF with ranked data as input.** The non-negativity constraint forces the objectives into the positive quarter of the Cartesian coordinate system. By providing ranked input to NNMF, I gained insight into the relationship of the objectives. The ones close to the axes are almost identical to the SRD result of best/worst objectives, but in-between them are the randomly given ones. Clusters appear as well. Cluster A) describes the alignment of proportional collaboration variables (C9-10), cluster B) suggests correlation between the number of female authors (G5) and the number of open access publications (O4-6)

somewhat positively correlates to the proportions of the 5-50% scientific impact and the average number of the publication.

On the other hand, the next cluster grouped (G5, O3-5) the number of female authorships and the quantity of hybrid, bronze and green open access publications. The remaining quantitative objectives are somewhat close to each other as in PCA. These objectives are the essential parts of the cumulative distribution by connecting the outlier with the best/worst objectives. The objectives of the upper clusters are in a similar relationship. NNMF selected the best objectives of ranking, the same that SRD chose, and put them in the vicinity of the first component. Because of the non-negativity constraint, all objectives correlate positively.

I enforced sparsity on the scaled data to see what non-dominated factors represent the dataset. On the first try, reducing the data resulted in all objectives converging on one axis. The objectives shifted towards both axes by removing the reverse ranking objectives, one with the proportional values and one with the quantity indicators. Figure 3.7 demonstrates the values of the objectives in relation to the two basis vectors. The proportional scientific indicators are considered transitional between the basis vectors. These

values either force the samples to be placed on higher-ranked Pareto front or with the improvement of the objectives located on one of the axes might result in the improvement in the scientific indicators.

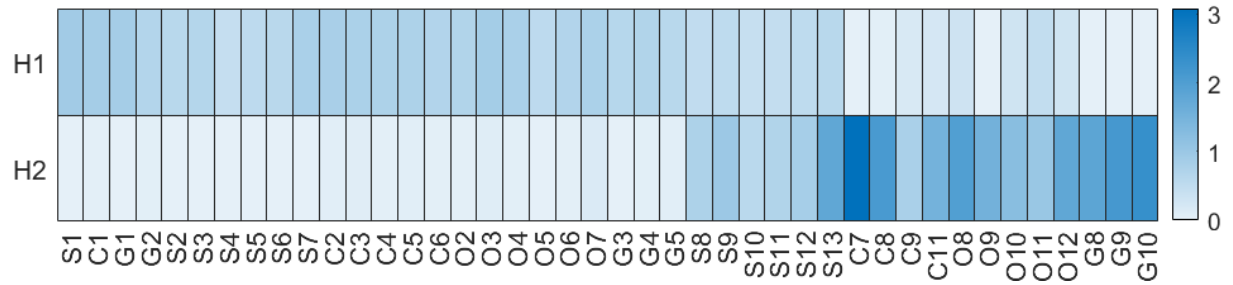


Figure 3.7: **Heatmap of the coefficients in sparse NNMF.** Sparse NNMF forces the coefficients of the objectives to take on the form of basis vectors, and sorts them into quantity (H1) and proportional (H2) objectives.

Chapter 4

Discussion and Conclusion

4.1 Discussion

It is important to admit that the reliability of the results is inherently distorted by the size and echo effects. In other words, larger universities have an advantage in almost every objective indicator, and a university that once enjoyed high prestige will look good for many years to come, even if it can no longer maintain its outstanding performance. The rankings based on different objectives agree in (and hence, are generally reliable for) the leaders and the drivers (the end of the ranking), while in the centre field of the different rankings, the difference in relative ranking is often only visible by a magnifying glass [43, 44]. These findings are well-underpinned by the solution of the SRD analysis, in which the row maximum is used as the gold (ideal) standard without preprocessing of the data.

The indicator of "unknown gender" is outstanding in the case of new universities of possibly smaller ethnicities with higher uncertainty of the gender of their researchers, which is most likely determined by automated engines. There is a probability that their universities are not ranked amongst the top universities due to their young history and the lack of publication records, and hence, the indicator of "unknown gender" ranks in the revised order.

Vavříková [12] ranked universities by taking their means, and used linear normalization, and performed a multi-objective optimization. Hypothetically speaking, in many cases of MaOO, aggregation retains some of the *curse of dimensionality*, as they do not restrict the deterioration of an objective. If the minimum of the rankings is taken, with a high amount of objectives, many solutions retain a common minimum ranked value [45]. On the other hand, NNMF excels at MaOO, as the objectives are reduced to the desired number, and the data behaves as if it is still MOO.

The extended TOPSIS method by Shih [9] *et al.* is much more fit than the latter

method to determine a ranking with a high amount of objectives but does not provide comprehensive visualization and information on the relationship of the objectives. I have shown in Figure 3.5 and Figure 3.6 that NNMF can perform the latter two; furthermore, clustering of the institutions are added to the toolbox of ranking.

To address Weighted Top Candidate (WTC), a novel selection method proposed by Sziklai [10], which recommends experts by previously determined experts, the only downside of the method is that the data has to be structured accordingly to provide preferable results and that the performance of the method is influenced heavily by the institutions as the recommendation system requires experts, to propose new experts. Compared to WTC, NNMF can be fed with differently preprocessed input, which may result in a somewhat changed structure, *e.g.* if ranking transformation is applied, then the results return the relationship of the ranks, rather than the data. Besides, there is no 'hierarchy' or 'recommendation' in NNMF; the clustered weights provide a Pareto front to be evaluated.

Compared to the University profiling Network (UPN) proposed by Kong *et al.* [11], NNMF is a somewhat less complex method, as it does not require the optimization of hidden layers. Autoencoders are prone to overfit the model, contrary to dimension reduction methods. The update function of NNMF can stop by setting the maximum number of iterations or the maximum of the error; if Equation 2.7 reaches the preferable value, the method stops. Both methods feature clustering. However, the main difference of the methods is that determining the relationships of objectives is much closer to the nature of NNMF as it provides coefficients, while autoencoders are complex methods, which makes it difficult to pry into and visualize as easily as NNMF can be.

Jiani *et al.* [46] proposed trade-off ranking a method that utilizes the distance from ideal points as a means to determine the Pareto set. It is a sensitive ranking: the choice of metrics, the existence of a well-distributed Pareto set are both a major factor for performance. On the difference between trade-off ranking and NNMF, the former is not limited to non-negativity due to being a distance-based method and focuses on the samples that outperform others in each component. Simultaneously, the latter reduces the number of objectives to the desired amount, clusters the Pareto fronts and provides coefficients to explore the underlying structure of data.

As such, I determined the Relative Entropy values of the set, the highest entropy objective pairs and each method in Figure 4.1. These represent the saturation of the Pareto fronts, and if the RE is too small, it indicates that the Pareto front is over-saturated, similarly to the upper left histogram which provides the saturation of the dataset. In this case, the Pareto optimality cannot distinguish the samples and puts each one of them on the Pareto front. The bottom-right histogram is the histogram of the highest objective pair depicted in Figure 3.1. The figure represents well separated Pareto

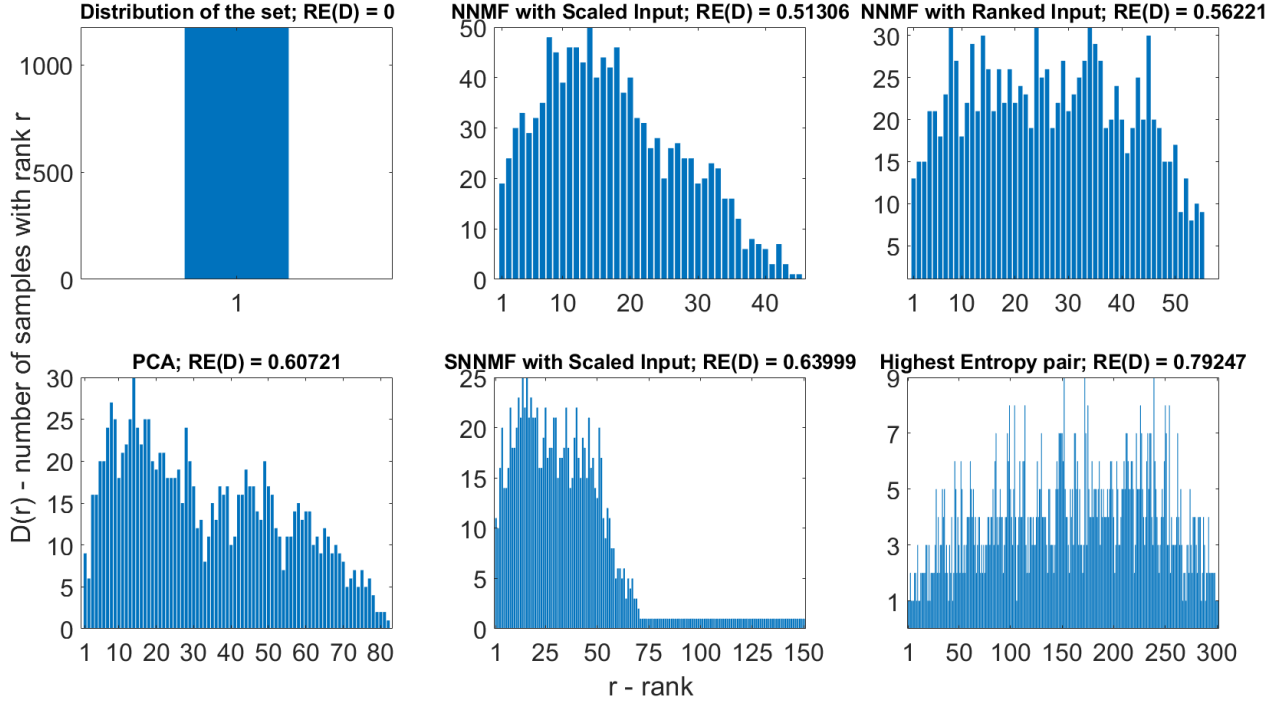


Figure 4.1: **Relative Entropies of the methods.** The smaller the range of the y axis (samples with rank), the bigger the range of the x axis (ranks) is and *vice versa*. Data with high RE have more ranks and fewer samples with the same rank. The histograms represent the saturation of the Pareto fronts. From the viewpoint of Entropy, PCA and sparse NNMf preforms more sufficiently than the core NNMf. As PCA has more ranks, the sparse NNMf clarifies a more specific set of optimal solutions which SRD validates. Furthermore, Bimodal distributions can be observed (or suspected), except for the top-left figure.

fronts with few members on them, 1 sample for the first front, 1 for the second, 2 for the third, 1 for the fourth. On the histogram, the same number of samples with ranks can be seen. The other histograms show the saturation of the dimension reduction methods. RE is only used as a measure to prove that the Pareto fronts are not oversaturated.

According to the hypervolume and entropy, I evaluated the performance of the methods in figure 4.2. The ideal solution has high relative entropy, but minimal hypervolume, as high relative entropy provides the measure of the efficiency of the ranking, and the hypervolume is the measure of the closeness to the ideal rank. The Pareto front on the set includes all points; thus, both entropy and hypervolume are close to zero. The ideal solution is selecting the highest entropy pair objective, as it has only one solution on the Pareto front. These are almost two directly proportional objectives, which is a coincidence, but in practice, it does not include any other objectives and conversely cannot solve a many-objective optimization problem. In between these two are the dimension reduction methods. NNMf with scaled input comes as an intermediate in hypervolume concerning sparse NNMf and PCA. The latter is not capable of clustering, though it performs better

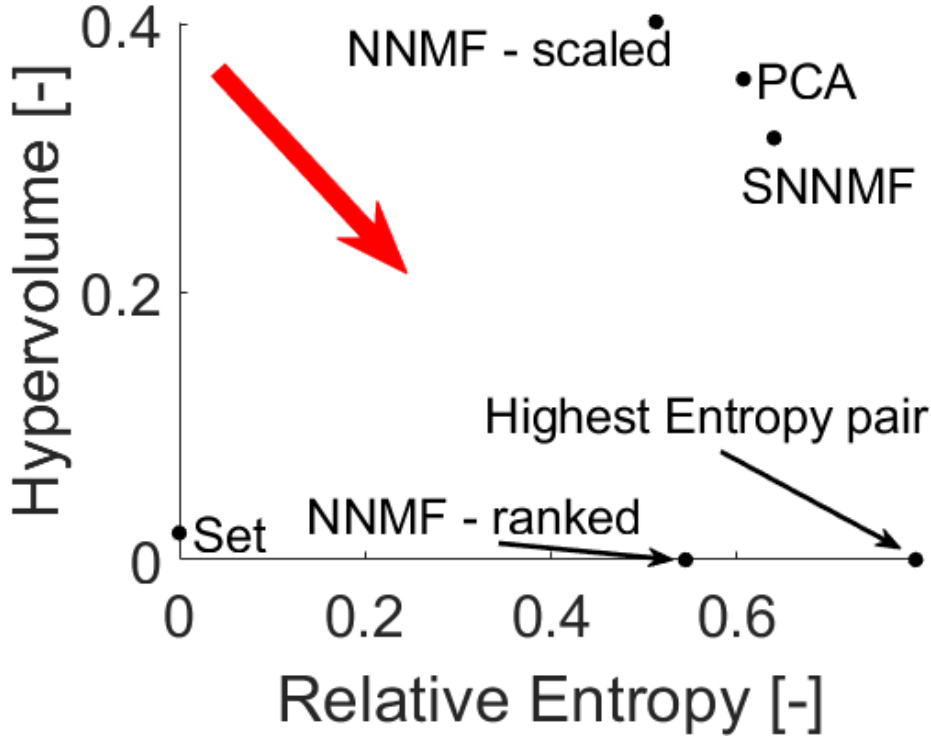


Figure 4.2: **Evaluation of Performance.** The more the entropy is, the more efficient the ranking is. The reference point for the hypervolume was the maximum of the dataset; thus, lesser values are preferable. The optimal method would be selecting the Highest Entropy pair of objectives, but the other objectives are not included in the ranking as only two are selected.

in determining the ideal Pareto fronts. Although it has a more accurate approximation of the ideal Pareto front, PCA has reversely correlating objectives, as well as conflicting ones. To visualize the data, most principal components is to be omitted. The NNMF with scaled input and the sparse NNMF do cluster; if the worst objectives are taken out from the data, sparse NNMF has higher relative entropy and less hypervolume. Sparsity can also be calibrated to control the number of solutions on the Pareto front, which comes in handy when the number of first ones is less than a specified number. The perfect method would compress the objectives but retain high entropy, and reduce the hypervolume, and stay close to the ideal Pareto front. The determined Pareto front has hardly any false non-dominated solutions.

The entropy value of NNMF with ranked input is not extraordinary, yet the hypervolume is zero. In the case of ranked NNMF, it might not provide a valid value, for the number of non-dominated solutions is not one, meaning that the reference point is constructed from the maximum of each objective and is through the border since a non-dominated solution must be better in at least one objective than its neighbours. If there is more than one sample on the Pareto front, then the area between the reference point

and the Pareto front cannot be zero as the Pareto front does not fit the reference point. They can only be zero when there is no distance between the reference and the Pareto front. An example is selecting the highest entropy pair, as there is only one solution on the Pareto front, which is the ideal reference point, the maximum of the database.

Multi-objective optimization is a discipline that helps deliver a more precise and better decision. By applying methods capable of *many*-objective optimization, an optimal decision can be made in all aspects of life, including engineering, where decision support systems can maximize performance and quality whilst decreasing costs, emission, and electricity consumption. In other words, the optimization of complex systems with 30 or more conflicting objectives is becoming available.

4.2 Conclusion

Present work does not simply confirm the trivial view that quantity of publications and visibility criteria (number of citations) is essential to receive a better ranking, but established some reversely ranking objectives, which should be avoided or diminished: publications without DOI, low male-female ratio, short distance cooperation.

A recommended publication policy would include complete utilization of open access possibilities and long-distance cooperations. Naturally diminishing the female contributions is *not recommended, neither is increasing the number of co-authors*. This empirical investigation makes the conservatism of the academic world visible.

Similarly, the empirical clustering of Asian, Western and "in-between" style universities does not involve any value judgment but unambiguously shows the different outcome and clustering of the universities. It should also be emphasized that the intermediate position of the top 1 % and 5 % citations suggest that the role of these indices are overweighted, and they are determined by random factors largely.

It is reassuring that the analyses return the well-known tendencies and commonplaces. On the other hand, the latter strengthens our surprising or non-common conclusions, make them probable.

This paper provides a complex analysis of ranking with high dimensionality and *many*-objective optimization. Methods such as PCA, NNMF, and sparse NNMF can determine a proper ranking even if the objectives are controversial and rank in reverse. SRD is a handy tool in presampling, as it detects unfavourable objectives that can be removed later.

In university ranking, the grouping of universities by NNMF is heavily influenced by national publication policies, clusters depending on where a university belongs in these (inter)national communities. The thorough analysis of the objectives returned trivial

statements that the amount of publications and citation is a driving force for a higher rank. However, some not-so-trivial results may help developing an advanced plan for an improved publication policy. I propose the endorsement of long-distance and international collaborations, as to promote the latter, the increase of open access publication may spark up many collaborations publications.

By employing NNMF for *many*-objective optimization, the constraints of the previously mentioned techniques can be eased, as only a non-negativity constraint is to be met. NNMF provides positively correlating objectives that can be further analyzed for the determination of the underlying relationships. This method clusters the data and the Pareto front. To answer the question of 'how to rank fairly?', I propose using SRD for filtering and sparse NNMF to select universities from throughout the world and simultaneously take as many viewpoints as possible into consideration.

Acknowledgments

Funding provided by the National Laboratory for Climate Change (NKFIH-872 project)

I would like to give my heartiest appreciation to my supervisors, János Abonyi, Gyula Dörgő and Károly Héberger. Gyula, I am extremely grateful for the help and care you have shown me the past months. Also, I had the pleasure of working with Károly, whose on-the-spot comments, unwavering precision and encouraging presence is a huge factor in my improvement. Last, but no least, János, your finesse in science amazed me, the opportunities you provided me ushered me to the next level.

Bibliography

- [1] Rohit Agrawal. Sustainable material selection for additive manufacturing technologies: A critical analysis of rank reversal approach. *Journal of Cleaner Production*, 296:126500, 2021.
- [2] Yongmao Xiao, Zhigang Jiang, Quan Gu, Wei Yan, and Ruping Wang. A novel approach to cnc machining center processing parameters optimization considering energy-saving and low-cost. *Journal of Manufacturing Systems*, 59:535–548, 2021.
- [3] A.M. Khan, L. Liang, M. Mia, M.K. Gupta, Z. Wei, M. Jamil, and H. Ning. Development of process performance simulator (pps) and parametric optimization for sustainable machining considering carbon emission, cost and energy aspects. *Renewable and Sustainable Energy Reviews*, 139:110738, 2021.
- [4] K. Deb and H. Jain. An evolutionary many-objective optimization algorithm using reference-point-based nondominated sorting approach, part i: Solving problems with box constraints. *IEEE Transactions on Evolutionary Computation*, 18(4):577–601, 2014.
- [5] L.S. Batista, F. Campelo, F.G. Guimaraes, and J.A. Ramirez. A comparison of dominance criteria in many-objective optimization problems. pages 2359–2366, 2011.
- [6] Ching-Lai Hwang and Kwangsun Yoon. *Methods for Multiple Attribute Decision Making*, pages 58–191. Springer Berlin Heidelberg, Berlin, Heidelberg, 1981.
- [7] Jarosław Wątróbski, Jarosław Jankowski, Paweł Ziemba, Artur Karczmarczyk, and Magdalena Ziolo. Generalised framework for multi-criteria method selection. *Omega (United Kingdom)*, 86:107–124, 2019.
- [8] Alexandre Bevilacqua Leoneti. Considerations regarding the choice of ranking multiple criteria decision making methods. *Pesquisa Operacional*, 36(2), 2016.
- [9] Hsu-Shih Shih, Huan-Jyh Shyur, and E. Stanley Lee. An extension of topsis for group decision making. *Mathematical and Computer Modelling*, 45(7):801–813, 2007.

- [10] Balázs R. Sziklai. Ranking institutions within a discipline: The steep mountain of academic excellence. *Journal of Informetrics*, 15(2):101133, 2021.
- [11] X. Kong, J. Li, L. Wang, G. Shen, Y. Sun, and I. Lee. Recurrent-dc: A deep representation clustering model for university profiling based on academic graph. *Future Generation Computer Systems*, 116:156–167, 2021.
- [12] L. Vavříková. Multicriteria decision making and rankings based on aggregation operators (application on assessment of public universities and their faculties). *Acta Polytechnica Hungarica*, 8(3):79–90, 2011.
- [13] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.
- [14] T. Li, C. Ding, and M. I. Jordan. Solving consensus and semi-supervised clustering problems using nonnegative matrix factorization. In *Proceedings - IEEE International Conference on Data Mining, ICDM*, pages 577–582, 2007.
- [15] A. Jain and C. Gupta. *Fuzzy logic in recommender systems*, volume 749 of *Studies in Computational Intelligence*. 2018.
- [16] B. Thirunavukarasu, N. Richi, and C. Yuen. People to people recommendation using coupled nonnegative boolean matrix factorization. In *ICSNS 2018 - Proceedings of IEEE International Conference on Soft-Computing and Network Security*, 2018.
- [17] F. Zhang, Y. Lu, J. Chen, S. Liu, and Z. Ling. Robust collaborative filtering based on non-negative matrix factorization and ℓ_1 -norm. *Knowledge-Based Systems*, 118:177–190, 2017.
- [18] N. Sivaramakrishnan, V. Subramaniaswamy, A. Vilorio, V. Vijayakumar, and N. Senthilselvan. A deep learning-based hybrid model for recommendation generation and ranking. *Neural Computing and Applications*, 2020.
- [19] H. H. Abbas, B. Z. Ahmed, and A. K. Abbas. 3d face factorisation for face recognition using pattern recognition algorithms. *Cybernetics and Information Technologies*, 19(2):28–37, 2019.
- [20] V. Snášel, Z. Horák, J. Kočíbová, and A. Abraham. Reducing social network dimensions using matrix factorization methods. In *Proceedings of the 2009 International Conference on Advances in Social Network Analysis and Mining, ASONAM 2009*, pages 348–351, 2009.

- [21] Kevin Foley, Derek Greene, and Padraig Cunningham. Optimizing conflicting objectives in nmf using pareto simulated annealing. pages 112–121, 01 2010.
- [22] K. Héberger. Sum of ranking differences compares methods or models fairly. *TrAC - Trends in Analytical Chemistry*, 29(1):101–109, 2010.
- [23] Károly Héberger. Method and model comparison by sum of ranking differences in cases of repeated observations (ties). *Chemometrics and Intelligent Laboratory Systems*, 127:139–146, 08 2013.
- [24] Károly Héberger and Klára Kollár-Hunek. Sum of ranking differences for method discrimination and its validation: Comparison of ranks with random numbers. *Journal of Chemometrics*, 25:151 – 158, 04 2011.
- [25] J.M. Lourenço and L. Lebensztajn. Post-pareto optimality analysis with sum of ranking differences. *IEEE Transactions on Magnetics*, 54(8), 2018.
- [26] Nees Jan Van Eck. Cwts leiden ranking 2020. August 2020.
- [27] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1):37–52, 1987.
- [28] P. O. Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, 5:1457–1469, 2004.
- [29] H. Kim and H. Park. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics*, 23(12):1495–1502, 2007.
- [30] Y. Li and A. Ngom. The non-negative matrix factorization toolbox for biological data mining. *Source Code for Biology and Medicine*, 8, 2013.
- [31] P. Ngatchou, A. Zarei, and A. El-Sharkawi. Pareto multi objective optimization. In *Proceedings of the 13th International Conference on, Intelligent Systems Application to Power Systems*, pages 84–91, 2005.
- [32] H. A. Taboada, F. Baheranwala, D. W. Coit, and N. Wattanapongsakorn. Practical solutions for multi-objective optimization: An application to system reliability design problems. *Reliability Engineering and System Safety*, 92(3):314–322, 2007. Cited By :103.
- [33] Oswaldo Aguirre; Heidi Taboada. A clustering method based on dynamic self organizing trees for post-pareto optimality analysis. *Procedia Computer Science*, 6, 2011.

- [34] Iraes Rabbers, Johan Heerden, Niclas Nordholt, Herwig Bachmann, Bas Teusink, and Frank Bruggeman. Metabolism at evolutionary optimal states. *Metabolites*, 5:311–43, 06 2015.
- [35] M. Davoodi Monfared, A. Mohades, and J. Rezaei. Convex hull ranking algorithm for multi-objective evolutionary algorithms. *Scientia Iranica*, 18(6):1435 – 1442, 2011.
- [36] H. Ishibuchi, N. Tsukamoto, and Y. Nojima. Evolutionary many-objective optimization: A short review. In *2008 IEEE Congress on Evolutionary Computation, CEC 2008*, pages 2419–2426, 2008.
- [37] G. H. Golub; C. Reinsch. Singular value decomposition and least squares solutions. *Numerische Mathematik*, 14, 04 1970.
- [38] John H. Kalivas. Overview of two-norm (l_2) and one-norm (l_1) tikhonov regularization variants for full wavelength or sparse spectral multivariate calibration models or maintenance. *Journal of Chemometrics*, 26, 2012.
- [39] S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Statist.*, 22(1):79–86, 03 1951.
- [40] C. M. Fonseca, L. Paquete, and M. Lopez-Ibanez. An improved dimension-sweep algorithm for the hypervolume indicator. In *2006 IEEE International Conference on Evolutionary Computation*, pages 1157–1163, 2006.
- [41] Yi Cao. Hypervolume indicator. *MATLAB Central File Exchange*, 2021.
- [42] Johannes Bader, Kalyanmoy Deb, and Eckart Zitzler. *Faster Hypervolume-Based Search Using Monte Carlo Sampling*, volume 634, pages 313–326. 01 2010.
- [43] Zs. T. Kosztyán, É. Orbán-Mihálykó, Cs. Mihálykó, V. V. Csányi, and A. Telcs. Analyzing and clustering students’ application preferences in higher education. *Journal of Applied Statistics*, 47(16):2961–2983, 2020.
- [44] Péter Mihályi. The current state of knowledge about university rankings. *Magyar Tudomány*, 181(10):1298–1308, 2020.
- [45] S. Kukkonen and J. Lampinen. Ranking-dominance and many-objective optimization. In *2007 IEEE Congress on Evolutionary Computation, CEC 2007*, pages 3983–3990, 2007.
- [46] N. Jaini and S. Utyuzhnikov. Trade-off ranking method for multi-criteria decision analysis. *Journal of Multi-Criteria Decision Analysis*, 24(3-4):121–132, 2017.

- [47] E. C. Fieller, H. O. Hartley, and E. S. Pearson. Tests for rank correlation coefficients.
i. *Biometrika*, 44, 12 1957.
- [48] Mike Thelwall, Carol Bailey, Catherine Tobin, and Noel-Ann Bradshaw. Gender
differences in research areas, methods and topics: Can people and thing orientations
explain the results? *Journal of Informetrics*, 13(1):149–169, 2019.

Appendix A

Appendix

A.1 Multicriteria analysis - TOPSIS

TOPSIS [6] is an aggregation-based method that transforms the samples with many objectives into a single value, whose distance from two ideal points is measured; this determines the sample's rank. The positive ideal solution can be chosen as $\max(f)$, and the negative ideal solution as $\min(f)$, TOPSIS can rank according to the best and worst of the data matrix. TOPSIS is a multicriteria analysis method, as it aggregates the multiple conflicting objectives to find optimality.

First, the input matrix $\mathbf{X}$ is to be normalized.

$$\mathbf{o}_k = \frac{x_{ik}}{\sqrt{\sum_{j=1}^N x_{jk}^2}}, i = 1, 2, 3 \dots N, k = 1, 2, 3 \dots n$$

Then the ideal points are determined, and the Euclidean distance to them is calculated for each sample.

$$\begin{aligned}\rho_- &= \min(\mathbf{o}_1, \mathbf{o}_2, \mathbf{o}_3 \dots \mathbf{o}_n) \\ \rho_+ &= \max(\mathbf{o}_1, \mathbf{o}_2, \mathbf{o}_3 \dots \mathbf{o}_n) \\ \mathbf{d}_{i-} &= \sqrt{\sum_{k=1}^n (x'_{ik} - p_k^-)^2}, i = 1, 2, 3 \dots N, k = 1, 2, 3 \dots n \\ \mathbf{d}_{i+} &= \sqrt{\sum_{k=1}^n (x'_{ik} - p_k^+)^2}\end{aligned}$$

Then the distances are scaled between $[0; 1]$

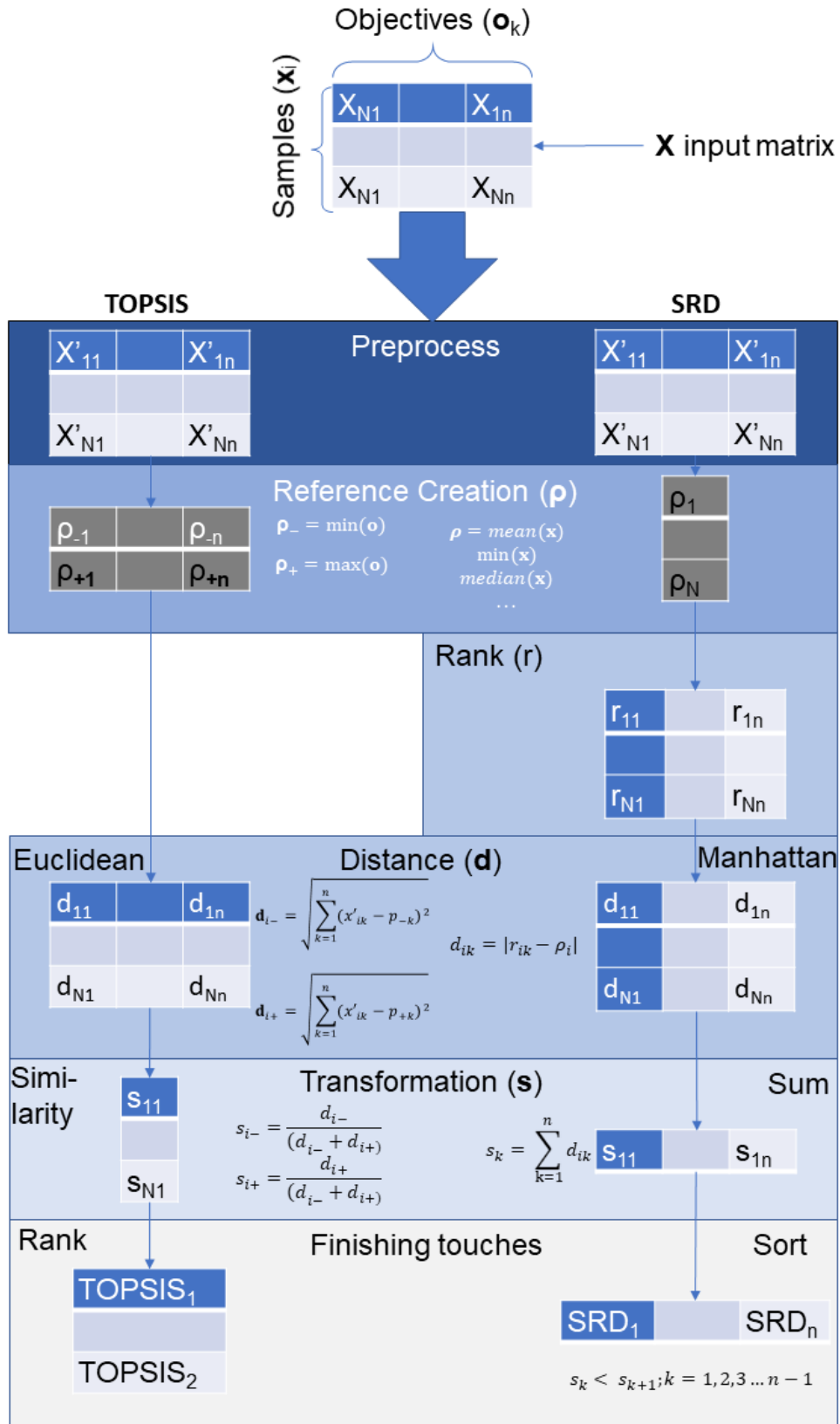


Figure A.1: **TOPSIS and SRD Flowchart.** The methods have similar steps, though they yield fundamentally different results.

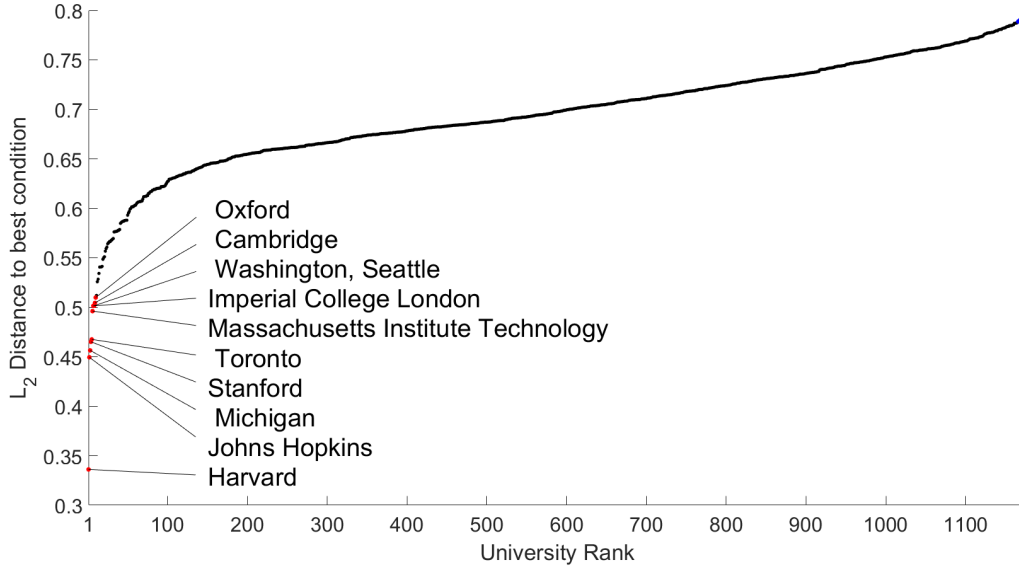


Figure A.2: **TOPSIS based ranking.** The universities are ranked by their distance to the positive ideal solution, *i.e.* best solution. The red samples are the first ten in the ranking, and the blue ones are the last ten. The samples draw an almost perfect shape suggest an almost perfect Spearman's rank correlation [47] value and indicate an unknown relationship. Notice that the best universities, according to TOPSIS, are in the Anglosphere.

$$s_{i-} = \frac{d_{i-}}{(d_{i-} + d_{i+})}$$

$$s_{i+} = \frac{d_{i+}}{(d_{i-} + d_{i+})}$$

where $s_{i+} = 0; s_{i-} = 1$ is the best sample, $s_{i+} = 1; s_{i-} = 0$ is the worst.

In order to have a ranking based on TOPSIS, we rank s_{i+} or s_{i-} . All objectives contribute towards the final score of the samples, but the correlation between them cannot be interpreted appropriately. The structure of the data lies undiscovered. The steps are described in Fig. A.1 along with SRD.

We applied TOPSIS to the Leiden CWTS 2020 database, and the result is in Figure A.2. TOPSIS ranks by two ideal points. To examine the relation of the samples, we calculated the Spearman's rank correlation [47], a result of which is 1. The first ten Universities are from the Anglosphere, as the figure represents (red solutions). On the other hand, the closest solutions (blue solutions) to the worst ideal solution are:

1. University of Shahrood
2. North University of China
3. Annamalai University.

4. Malek Ashtar University of Technology
5. Hebei University
6. Xinjiang University
7. Indian Institute of Technology (Indian School of Mines)
8. Indian Institute of Technology Kanpur
9. Yanshan University
10. Urmia University

A.2 The interpretation of the principal components

To provide better insight to the PCA analysis, we extended section 3.2.2 with the description of the five other PCs and a heatmap of the first seven PC. Although not a direct measure of performance, except for the total number of authorships (G1), the gender indicators provided help establish links between genders and other indicators and the relative role of genders in current academics. We created Figure A.3 to provide information on the principal components. According to PC3 (8.49%), the proportion of female authors (G8, G10) goes hand in hand with the higher number of open access publications resulting in more frequent collaboration and accessible information. This trend shows the modern wave of openness in the traditionally conservative world of the academy. Thelwall *et al.* [48] that there are significant differences in the fields that genders are a majority in. Female authors might be more oriented towards people, although each profession in a field may yield different proportions in gender *e.g.* nurses and neurosurgeons. Male authors tend to focus on abstractions, objects, systems and STEM rather than social callings, education and healthcare. PC4 (5.36%) promotes that the proportion of male authors (G7-G9) correlate with the proportion of industrial collaboration (C9), yet none of the genders has a significant relation to scientific indicators. On the other hand, on PC5 (4.45%), the proportion of authors with unknown gender correlates with all of the above, and again, there is no significant connection to scientific indicators. The following PC (2.58%) provides information on the correlation of the proportion of long-distance (C11) and the proportion of international collaborations (C8). PC7 (2.23%) indicates that with the proportion of short distance collaboration, there is a weak connection to the scientific correlation and industry collaboration, as companies often do joint projects/researches with universities that promote collaboration and scientific research. The empirical data provided by PC3-5 implies that there is no direct positive connection between scientific

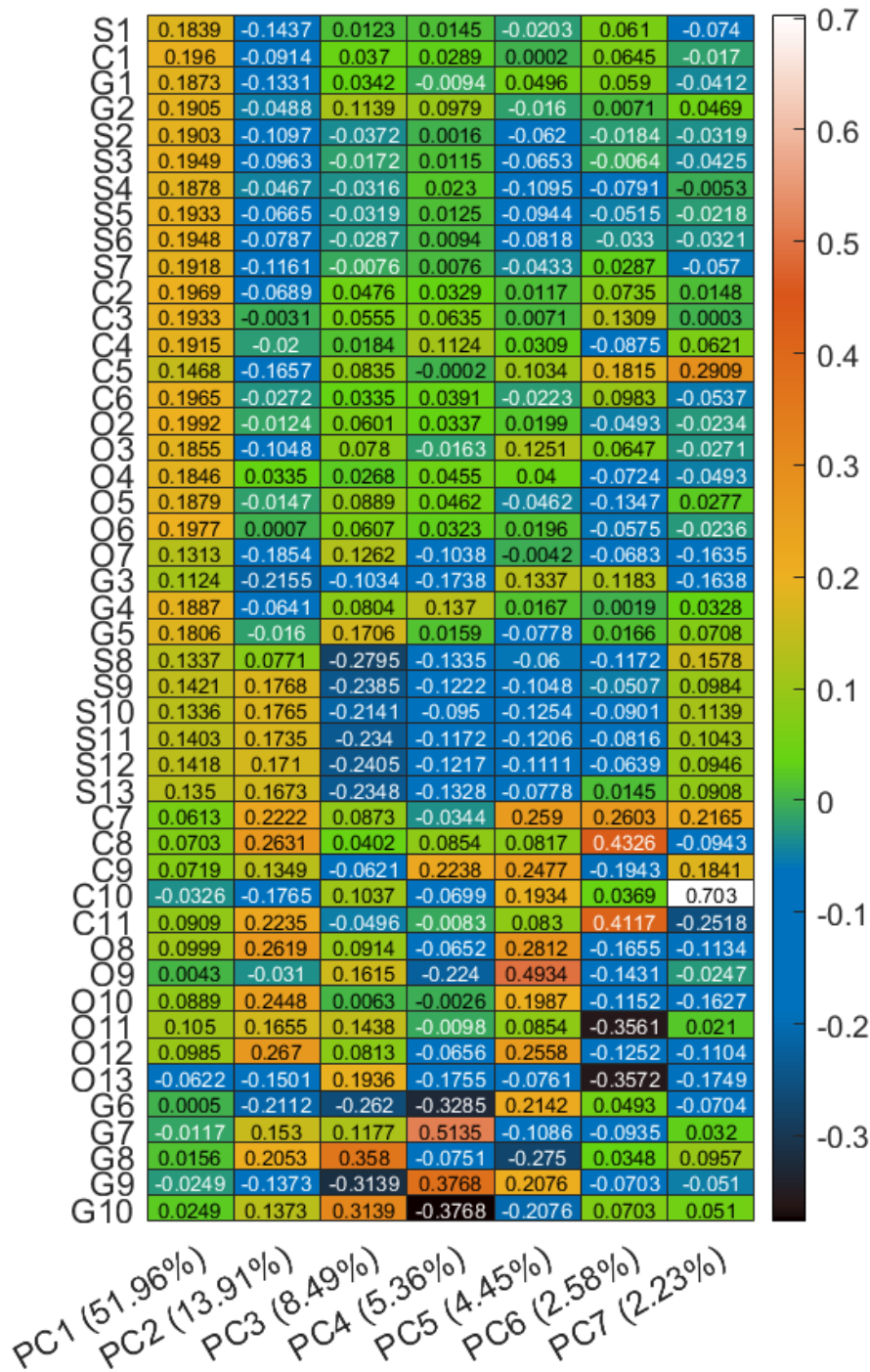


Figure A.3: **Heatmap of the first seven Principal Component** The first PC incorporates most of the quantity vindicators, the second does the same with proportional indicators, except for the proportion of unknown gender (G6) and male to known gender ratio (G9). The third takes on female gender indicators as the main objectives. The fourth does the same with male indicators. The fifth PC indicates the relationship between open access and collaboration indicators. The sixth and seventh PC mainly features one or two indicators.

impact and gender. The correct indirect correlations are an open question, as a hierarchy of relationships underlies all indicators.

A.3 Nomenclature

A.3.1 Mathematical Notation

$\mathbf{X}$ - Numerical 2D (row and column) data extracted from the database.

N - Number of rows

n - Number of columns

$\mathbf{x}$ - A sample, row vector of $\mathbf{X}$, a university

i - Index of a sample, $i = 1, 2, 3 \dots N$

$\mathbf{o}$ - An objective, column vector of $\mathbf{X}$, ranking variable, performance descriptor, indicator
(*e.g.* S1)

k - Index of an objective, $k = 1, 2, 3 \dots n$

r - Ranks, $1, 2, 3 \dots N$

$f_k(\mathbf{o}_i)$ - Objective function that ranks $\mathbf{o}_k$

$\mathbf{R}$ - Rank(ed) matrix, created by applying $f_k(\mathbf{o}_k)$ to $\mathbf{X}$

$D(r)$ - Number of objects with rank r

$RE(D)$ - Relative entropy of an input matrix

p - Number of objectives after dimension reduction, less than n .

$\mathbf{Y}$ - The matrix with reduced dimensions ($N \times p$)

$g(\mathbf{x}_i)$ - Dimension reduction function that reduces $\mathbf{x}_i$ sample with n amount of objectives
to $\mathbf{y}_i$ with p amount of objectives.

$\hat{\mathbf{X}}$ - Reconstructed matrix from $\mathbf{y}_i$ -s

$\mathbf{U}$ - Matrix of eigenvectors

Σ - Matrix of the singular values

$\mathbf{M}$ - Matrix of the loading vectors

η_i - Variance of a principal component

$\mathbf{V}$ - Input matrix of non-negative matrix factorization

$\mathbf{W}$ - $N \times p$ representation of the reduced samples by NNMF

$\mathbf{H}$ - $p \times n$ coefficient matrix by NNMF

α - Regularization parameter to suppress $\mathbf{W}$ and $\mathbf{H}$

β - Regularization parameter to suppress $\mathbf{W}$ and $\mathbf{H}$

ρ - Reference vector for either SRD or TOPSIS. Can be either column or row vector.

$\mathbf{d}$ - Distance from the respective reference value (ρ). Can be either column or row vector.

$\mathbf{s}$ - Similarity (in case of TOPSIS), or Sum (in case of SRD). Can be either column or row vector.

A.3.2 Leiden database variables - Scientific Impact Indicators

S1 - The total number of publications of a university with scientific impact.

S2, S8 - The total and the average number of citations of the publications of a university.

S3, S9 - The total and the average number of citations of the publications of a university, normalized for field and publication year.

S4, S10 - The number and the proportion of a university's publications that, compared with other publications in the same field and in the same year, belong to the top 1% most frequently cited.

S5, S11 - As above, belong to the top 5% most frequently cited.

S6, S12 - As above, belong to the top 10% most frequently cited.

S7, S13 - As above belong to the top 50% most frequently cited.

A.3.3 Leiden database variables - Collaboration Indicators

C1 - The total number of publications of a university in collaboration with other universities.

C2, C7 - The number and the proportion of a university's publications that have been co-authored with one or more other organizations.

C3, C8 - The number and the proportion of a university's publications that two or more countries have co-authored

C4, C9 - The number and the proportion of a university's publications that have been co-authored with one or more industrial organizations. All private sector for-profit business enterprises, covering all manufacturing and services sectors, are regarded as industrial organizations.

C5, C10 - The number and the proportion of a university's publications with a geographical collaboration distance of less than 100 km.

C6, C11 - The number and the proportion of a university's publications with a geographical collaboration distance of more than 5000 km.

A.3.4 Leiden database variables - Gender Indicators

G1 - The total number of authorships of a university. Consider, for instance, a publication with five authors, of which three report university X as their affiliation and two report university Y as their affiliation.

G2 - The number of male and female authorships of a university, that is, a university's number of authorships for which the gender is known.

G3, G6 - The number and proportion (to G1) of authorships of a university for which the gender is unknown

G4, G7 - The number and proportion (to G1) of male authorships of a university.

G5, G8 - The number and proportion (to G1) of female authorships of a university.

G6 - The number of authorships for which the gender is unknown as a proportion of a university's total number of authorships.

G9 - The number of male authorships as a proportion (to G2) of a university's number of known male and female authorship.

G10 - The number of female authorships as a proportion (to G2) of a university's number of known male and female authorship.

A.3.5 Leiden database variables - Open Access Indicators

O1 - Removed, as it consists of the same values as C1.

O2, O8 - The number and proportion of open access publications of a university.

- O3, O9** - The number and proportion of gold open access publications of a university.
Gold open access publications are publications in an open-access journal.
- O4, O10** - The number and proportion of hybrid open access publications of a university.
Hybrid open access publications are publications in a subscription journal that are open access.
- O5, O11** - The proportion of bronze open access publications of a university. Bronze open access publications are publications in a journal that is open access without a license.
- O6, O12** - The number and proportion of green open access publications of a university.
Green open access publications are publications in a journal that are also available in an open access repository.
- O7, O13** - The proportion of a university's publications for which the open access status is unknown. These publications typically do not have a DOI in the Web of Science database.