Knowledge-Based Systems 251 (2022) 109180

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Network-based dimensionality reduction of high-dimensional, low-sample-size datasets

Zsolt T. Kosztyán [a,b,c,*], Marcell T. Kurbucz [a,d,e], Attila I. Katona [a]

[a] Department of Quantitative Methods, University of Pannonia, Egyetem Street 10, Veszprém, H-8200, Hungary
[b] Institute of Advanced Studies (iASK), Chernel Street 14, Kőszeg, H-9730, Hungary
[c] MTA-PE Budapest Ranking Research Group, Eötvös Loránd Research Network (ELKH), Piarista Street 4, Budapest, H-1052, Hungary
[d] Wigner Research Centre for Physics, Department of Computational Sciences, Konkoly-Thege Miklós Street 29-33, Budapest, H-1121, Hungary
[e] Corvinus University of Budapest, Department of Statistics, Fővám Square 8, Budapest, H-1093, Hungary

ARTICLE INFO

Article history:
Received 9 June 2021
Received in revised form 3 March 2022
Accepted 30 May 2022
Available online 4 June 2022

Keywords:
Nonparametric methods
Dimensionality reduction
Community detection
Communality analysis

ABSTRACT

In the field of data science, there are a variety of datasets that suffer from the high-dimensional, low-sample-size (HDLSS) problem; however, only a few dimensionality reduction methods exist that are applicable to address this type of problem, and there is no nonparametric solution to date. The purpose of this work is to develop a novel network-based (nonparametric) dimensionality reduction analysis (NDA) method, that can be effectively applied to HDLSS data. First, with the NDA method, the correlation graph of variables is specified. With a modularity-based community detection method, the set of modules is specified. Then, the linear combination of variables weighted by their eigenvector centralities (EVCs), defined as LVs, is determined. In the optional phase of variable selection, variables with low EVCs and low communality are ignored. Then, the set of LVs and the set of indicators belonging to the LVs are specified using the NDA method. NDA is applied to publicly available databases and compared with principal factoring with community analysis (PFA) methods. The results show that NDA can be effectively applied to HDLSS datasets as it outperforms the existing methods in terms of interpretability. In addition, the application of NDA is easier, since there is no need to specify the number of latent variables due to its nonparametric nature.

1. Introduction

In high-dimensional data, the number of variables is usually much larger than the sample size. Since in such a dataset,[1] many dimensions are irrelevant or redundant, dimensionality reduction has become an essential element of the engineering approach to mathematical modeling [1,2]. Many model reduction methods have been developed and successfully implemented in applications, such as approximation by manifolds, graphs, and complexes [3,4]; low-rank tensor network decompositions [5,6]; deep generative models, such as the variational autoencoder (VAE) [7–9]; ensemble learning methods, such as boost [10], random forest [11], forest of extreme learning machines with rule-base transferring (FELM-RT) [12], ensemble of radial basis function neural networks in decision tree structure with knowledge transferring (ERDK) [13], and forest of decision trees with radial basis function networks and knowledge transferring (FRDK) [14], as well as various versions of factor analysis (FA) [15,16] and principal component analysis (PCA) [17–20].[2]

Note that while both PCA and PFA are widely used methods of dimensionality reduction, they have a major disadvantage. In these methods, it is assumed that data are linearly separable. However, the linear model is not always reliable in capturing nonlinear relationships in real-world problems, especially with limited samples [7,22]. To solve this problem, Schölkopf et al. [23,24] proposed kernel PCA (KPCA), which is a non-linear extension of PCA that uses kernel methods. The application of KPCA in the field of HDLSS data analysis is an active research area. For instance, Liu et al. [25] and Reverter et al. [26] applied this method for the analysis of HDLSS gene expression data, while in their recent paper, Nakayama et al. [20] tested clustering performance on HDLSS microarray datasets.

An additional difficulty in applying ordinary PCA is that principal components are usually linear combinations of all input

* Corresponding author at: Department of Quantitative Methods, University of Pannonia, Egyetem Street 10, Veszprém, H-8200, Hungary.
E-mail address: kosztyan.zsolt@gtk.uni-pannon.hu (Z.T. Kosztyán).

[1] These are often termed high-dimensional, low-sample-size (HDLSS) datasets.

[2] Furthermore, the combination of multiple feature selection methods is a widely used approach. The performance of various parallel and serial combination techniques using HDLSS data was examined and compared by Tsai and Sung [21].

https://doi.org/10.1016/j.knosys.2022.109180

variables, which makes it difficult to interpret the results, especially in the case of HDLSS datasets (see, e.g., [27]). To address this problem, Zou et al. [28] proposed sparse PCA (SPCA), which aims at producing easily interpreted models through sparse loadings; i.e., the principal components are linear combinations of a subset of the original variables [29]. Moreover, Jiang et al. [30] proposed the approximated gradient flow (AgFlow) method to lower the computation complexity of the aforementioned problem under HDLSS settings. In contrast to PCA, there is an iterative feature (i.e., variable) selection step in PFA called communality analysis, in which irrelevant indicators (for which the square correlation[3] between an LV and the indicator is below a certain threshold), and after factor rotation, common indicators (where the communality values for different LVs are similar) can be ignored. The interpretation of a factor depends on which group of variables is correlated with it. In other words, the interpretation of a factor reveals which indicators belong to the factor. Therefore, it must be clearly determined to which factor a variable belongs. Otherwise, the factors will be difficult to interpret [31]. For this reason, common variables should also be ignored. In this study, we show that this variable selection approach can be adapted to the proposed network-based (nonparametric) dimensionality reduction analysis (NDA).

Explanatory FA (EFA) [32] and PCA are statistical methods that are widely applied to simplify complex sets of data and to describe the covariance relationships among variables. While both the EFA and PCA methods seek to approximate the covariance matrix, the EFA model is more complex, as its major question is whether the data are consistent with a prescribed structure. The two methods are used to reduce a large number of variables to a smaller number of factors called latent variables (LVs). Common FA (CFA) [33], also called principal FA (PFA) [34] or principal factoring, can be considered a combination of FA and PCA. PFA also seeks the fewest factors that can account for the common variance (correlation) of a set of variables. PCA can be defined as follows: $\mathbf{Z} = \mathbf{FL}$, where $\mathbf{Z}$ is the standardized original data matrix (denoted as $\mathbf{D}$), $\mathbf{F}$ is the standardized factor score matrix, and $\mathbf{L}$ is a factor multiplied by the variable weight matrix (factor $\times$ variable weight matrix) [35]. The columns of $\mathbf{L}$ are multiplied by the square root of the corresponding eigenvalues, that is, the eigenvectors are scaled up by the variances. It is assumed that $\mathbf{Z}$ represents the variance of a variable, $\mathbf{F}$ represents a variance of common factors (or a common variance among $\mathbf{Z}$ and other variables in the analysis), and $\mathbf{L}$ represents coefficients showing how $\mathbf{F}$ and $\mathbf{Z}$ are related [34]. The PFA equation is $\mathbf{Z} = \mathbf{FL} + \mathbf{U}$. The difference between the two equations is the last component (i.e., $\mathbf{U}$). $u_j \in \mathbf{U}$ is the unique variance of a variable j. $\mathbf{U}$ contains both the unreliable variance of the measurement error and the reliable variance, which does not overlap with the common variance [36]. PCA assumes that the communality ($h_i^2 = \sum_{j=1}^{k} p_{ij}^2$), i.e., the common variance, becomes maximized and that there is no unique variance in each variable. In contrast, PFA assumes that there is a substantial amount of unique variance ($u_i = 1 - h_i^2$) and reliable common variance [34]. PCA condenses the original variables into a smaller number of components, thus performing a model or dimensionality reduction. PFA finds a factor model (factor structure) that best reproduces the observed correlation and thus is aimed at explaining the correlation between variables [34]. Since PFA specifies variables, this method usually involves variable selection, where multicollinear and irrelevant variables must be ignored.

Despite the widespread use of PCA- and FA-related methods, they have several shortcomings, which have to be improved:

1. Both EFA and PFA assume that the number of observations should be at least 10 times the number of variables [37]. It has also been reported that the use of PCA in the analysis of the HDLSS dataset did not yield satisfactory results [38,39]. Unlike SPCA, these methods cannot be used if the number of observations is equal to or lower than the number of variables.
2. PFA handles only scale variables. Despite the fact that several methods, such as confirmatory FA (CoFA) [40], are already able to handle nonscale variables such as ordinal variables [41], arbitrary (e.g., nonlinear, ordinal-scale) interdependencies between indicators cannot be addressed.
3. Although an explanation of the total variance of LVs and eigenvalues can help us to determine the maximal number of LVs, there is no exact method to specify the number of LVs, which is problematic because the number of LVs should be specified before performing FA. PCA and its extensions are also typically parametric methods.

The proposed method fills these gaps.

1. First, the correlation graph between indicators is specified. A modularity-based community detection method, such as the Louvain [42] or Leiden [43] method, specifies modules in the correlation graph. These modules specify the set of indicators, where the interdependencies within a module are higher than those between modules. This modularity-based community detection method can be applied to large-scale correlation graphs, where the nodes are variables and the arcs between nodes are the correlations between variables.
2. By specifying the correlation graph through communality analysis to specify the factor loadings and factor scores, instead of Pearson's correlation, the Kendall and Spearman correlations can be used if ordinal variables need to be addressed. However, in the case of arbitrary detection, such as the detection of nonlinear interdependencies, the distance correlation (DC) [44] can also be used in the proposed method. The advantage of applying the DC is that its value is zero if and only if there is no dependency between variables.
3. After carrying out modularity-based community detection, the number of modules specifies the number of LVs. The linear combination of the eigenvector centrality (EVC) and indicators specifies the LV in a module. The optional iterative feature selection phase refines the LVs while ignoring indicators with low EVC and/or low communality.

In other words, the contribution of this paper is the development of an NDA method that can be applied to HDLSS data and has the following properties:

1. Non-parametric;
2. Robust (scaleable);
3. Handles non-linear measures;
4. Supports geometric representations;
5. Includes embedded feature selection.

This paper is organized as follows. Section 2 introduces the data utilized in this study and the methodology. Section 3 details the calculation process. Section 4 presents and discusses the results of the NDA method, and also compares the results of the NDA method with the results of the PFA, PCA, and SPCA methods. Finally, Section 5 provides the summary and conclusions of this work and, proposes future research directions.

[3] This is referred to as the communality value.

2. Materials and methods

2.1. Data employed

Two publicly available data sources are employed. The first is the CWTS Leiden Ranking 2020 database, which includes 42 indicators and 1176 universities from all around the world.[4] The Leiden Ranking offers a multidimensional perspective on university performance. The Leiden Ranking provides indicators of scientific impact, collaboration, open-access publishing, and gender diversity. Size matters when comparing universities: performance can be viewed from an absolute or a relative perspective (e.g., the number versus the percentage of highly cited publications). Thus, size-dependent and size-independent indicators are consistently presented together in the Leiden Ranking. Hence, both types of indicators need to be taken into account. The CWTS Leiden Ranking 2020 provides statistics not only at the level of science as a whole but also at the level of the following five main fields of science: (1) Biomedical and health sciences, (2) Life and earth sciences, (3) Mathematics and computer science, (4) Physical sciences and engineering, (5) Social sciences and humanities. The Leiden Ranking is based on publications in the Web of Science database produced by Clarivate Analytics. The most up-to-date statistics made available in the Leiden Ranking are based on publications in the period 2015–2018, but statistics are also provided for earlier periods. Since, the number of observations (i.e., universities) is more than ten times greater than the number of variables, PFA and the proposed NDA can be compared. The CWTS Leiden 2020 university ranking database (herein, the CWTS'2020 database) contains several time periods and scientific fields. We consider only the latest (2015–2018) time period and include all scientific fields. To calculate the impact indicators, we perform fractional counting.[5] The list of applied indicators are in Appendix A.

The second database is a joint dataset of official COVID-19 reports and the governance, trade and competitiveness indicators of World Bank Group platforms [45]. This cross-section includes the governmental, trade, and competitiveness relationships among official COVID-19 reports. It contains 18 COVID-19 variables generated based on the official reports of 138 countries, as well as an additional 2160 governance, trade, and competitiveness indicators from the World Bank Group's GovData360 and TCdata360 platforms.[6] With the exception of the 18 COVID-19 indicators, all indicators are used from this database. The list of indicators can be found in Appendix B. In both cases, the original values of the given indicators of these two datasets are used without any data cleaning or data manipulation. This database includes 2160 governance, trade, and competitiveness indicators (herein, GovDB'20) are used for the 138 countries, the number of indicators is ten times greater than the number of observations. Therefore, PCA and PFA cannot be used.

2.2. Methods employed

2.2.1. Applied correlation coefficients

The first step of the proposed NDA method is to specify the bivariate correlation graph between indicators. Several bivariate correlation coefficients exists. One of the most widely used is Pearson's correlation coefficient (see Eq. (1)), which measures the linear correlation between two variables (v_i, v_j).

$$\rho_{v_i,v_j} = \frac{E[(v_i - \mu_{v_i})(v_j - \mu_{v_j})]}{\sigma_{v_i}\sigma_{v_j}}, \tag{1}$$

where $E(v_i) = \mu_{v_i}$ is the mean value and σ_{v_i} is the standard deviation of variable v_i.

Pearson's correlation coefficient measures only the linear correlation between metric (i.e., continuous) variables. The Spearman correlation coefficient is defined as the Pearson correlation coefficient between ordinal variables. The Kendall correlation is defined as:

$$\tau_{v_i,v_j} = \frac{2}{n(n-1)}[n_c(v_i, v_j), n_d(v_i, v_j)], \tag{2}$$

where $n_c(v_i, v_j)$ is the number of concordant pairs and $n_d(v_i, v_j)$ is the number of discordant pairs.[7]

The Spearman and Kendall correlation coefficients are suitable for ordinal values but can also be used on metric values; in this way, not only linear but also monotonic relationships between variables can be measured.

The distance correlation [44,46] is a measure of the association strength between nonlinear random variables. It surpasses Pearson's correlation because it can be used to discern other associations in addition to linear ones and can work multidimensionally. The distance correlation ranges from 0 to 1, where 0 implies independence between variables v_i and v_j and 1 implies that the linear subspaces of v_i and v_j are equal. Define $X = v_i$ and $Y = v_j$. The distance correlation is specified as:

$$dCor(v_i, v_j) = dCor(X, Y) = \frac{dCov(X, Y)}{\sqrt{dVar(X) \cdot dVar(Y)}}, \tag{3}$$

where $dVar(X) = dCov(X, X)$, and

$$dCov^2(X, Y) = \frac{1}{N^2}\sum_{J=1}^{N}\sum_{K=1}^{N} A_{J,K}B_{J,K}, \tag{4}$$

where N is the number of observations. Here,

$$A_{J,K} = a_{J,K} - \bar{a}_{J\cdot} - \bar{a}_{\cdot K} - \bar{\bar{a}}_{\cdot\cdot} \tag{5}$$

$$B_{J,K} = b_{J,K} - \bar{b}_{J\cdot} - \bar{b}_{\cdot K} - \bar{\bar{b}}_{\cdot\cdot}, \tag{6}$$

where $a_{J,K} = \|X_J - X_K\|$, $b_{J,K} = \|Y_J - Y_K\|$, $J, K = 1, 2, \ldots, N$; $\|.\|$ is the distance measure; and $\bar{a}_{J\cdot} = \frac{1}{n}\sum_{K=1}^{N} a_{J,K}$, $\bar{a}_{\cdot K} = \frac{1}{n}\sum_{J=1}^{N} a_{K,J}$, and $\bar{\bar{a}}_{\cdot\cdot} = \frac{1}{n^2}\sum_{J,K=1}^{N} a_{J,K}$.

The distance correlation is also an extension of Pearson's correlation and the Spearman correlation; however, Székely and Rizzo [46] stated that the distance correlation is 0 if and only if $X = v_i$ and $Y = v_j$ are independent.

NDA can apply all the above correlation coefficients; however, when analyzing the nonlinear relationship between variables, the distance correlation is recommended. Furthermore, the square correlation between variable v_i and variable v_j is denoted as $r_{i,j}$.

2.2.2. Modularity-based community detection

After the correlation graph is determined from the variables, a modularity-based community detection method can be used to specify the group of highly intercorrelated variables.

In this study, we use the modularity measure introduced by Newman [47], represented in Eq. (7), to uncover modules and

[4] The database can be downloaded for free from https://www.leidenranking.com/ranking/2020/list (retrieved: May 14, 2021).

[5] The fractional counting method gives less weight to collaborative publications than to noncollaborative ones.

[6] More information about the GovData360 and TCdata360 platforms can be found at https://govdata360.worldbank.org/ (retrieved: May 14, 2021) and https://tcdata360.worldbank.org/ (retrieved: May 14, 2021).

[7] Any pair of observations (x_i, y_i) and (x_j, y_j), where $i < j$, are said to be *concordant* if the sort order of (x_i, x_j) and (y_i, y_j) agrees, that is, if either both $x_i > x_j$ and $y_i > y_j$ hold or both $x_i < x_j$ and $y_i < y_j$ hold; otherwise, they are said to be *discordant*.

to set up the modular structure of the network. A network module is a subgraph whose vertices are more likely to be connected (i.e., variables are more likely to be correlated) than those outside the subgraph.

$$M = \frac{1}{2L}\sum_{ij}\left(r_{i,j} - \gamma\hat{r}_{i,j}\right)\delta(C_i, C_j), \tag{7}$$

where M is the modularity value; $r_{i,j}$ is the edge weight between node i and node j; $\hat{r}_{i,j}$ is the expected weight between node i and node j (the so-called null model); L is the total number of weights in the network; γ is a constant (default value is 1); and the Kronecker δ is 1 if the community of variable i and the community of variable j are the same and 0 otherwise. In the correlation graph, the nodes represent variables and the edges represent the bivariate square correlations between variables. As a null model, we use Newman's [47] configuration model, which here is $\hat{r}_{i,j} = \frac{r_i r_j}{2R}$, where r_i and r_j are the outdegree and indegree of nodes i and j, respectively. In addition, $\hat{r}_{i,j} = \frac{2}{n(n-1)}\sum_{i:=1}^{n-1}\sum_{j:=i+1}^{n} r_{i,j}$ and $\hat{r}_{i,j} = r_{\min}$ are allowed.

Modularity reflects the extent of expected edge weights relative to the random configuration network (where the edge weight is $\hat{r}_{i,j}$). The modularity maximization algorithm organizes the vertices into groups such that the edge weights within the modules are as large as possible compared to the configuration model. We used both the Louvain [42] and Leiden [43] algorithms. The Leiden algorithm takes more time but outperforms the popular Louvain algorithm [43].

2.2.3. Calculating the latent variables with eigenvector centralities

In network science, EVC is a measure of the influence of a node in a network [48]. Relative scores are assigned to all nodes in the network based on the concept that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes. A high eigenvector score means that a node is connected to many nodes who themselves have high scores. The EVC for variable v_i (i.e., for node i) can be calculated as follows:

$$c_i = \frac{1}{R}\sum_{j} r_{i,j}c_j, \tag{8}$$

where R is a constant and $r_{i,j}$ is the edge weight (i.e., square correlation) between nodes (i.e., variables). The centrality measure has several important properties: the score value is independent from the node properties, and it depends only on the network structure. Nodes with low centrality values are called *peripheral* nodes, while nodes with high centrality values are called *core* nodes.

The EVC has additional important properties, e.g., anonymity,[8] symmetry,[9] positive homogeneity,[10] and robustness.[11]

The LV for the module C_I is calculated as follows:

$$LV_I = \frac{\sum_{i\in C_I} c_i z_i}{\sum_{i\in C_I} c_i}, \tag{9}$$

where LV_I is the LV for module C_I; c_i is the EVC of variable v_i and $z_i = (v_i - \mu_{v_i})/\sigma_{v_i}$ is the standardized variable of variable v_i.

[8] The scores of nodes are unaffected by the way in which they are labeled.
[9] Symmetric nodes receive the same score.
[10] This means homogeneity for all positive values.
[11] This indicates invariance after adding an average node.

3. Calculation

The calculation is performed as follows. Steps 1–3 of the calculation specify the LVs. These steps are mandatory. In contrast, steps 4–5 are supplementary steps. These steps are good for the variable selection process and the refinement of the LV. While steps 4–5 are optional, they help us to satisfy the following assumptions:

Assumption 1. The minimal centrality value or minimal communality value of variables must be greater than a certain threshold.

This assumption ensures that the LV describes the variables in a module and that indicators in a module belong to its LV.

Assumption 2. There are no common variables.

This assumption ensures that there is no indicator that belongs to more than one LV.

Denote the set of variables as $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\} = \mathbf{V}$. $\mathbf{v}_1, \ldots, \mathbf{v}_n \in \mathbb{R}^m$, where m is the number of observations and n is the number of variables.

Step 1: Specification of the correlation graph

Denote $G(\mathcal{N}, \mathcal{A}, \mathcal{W})$ as the undirected weighted graph of the correlation graph, where $\mathcal{N}$ is the set of nodes. The node i represents the variable i, $i = 1, 2, \ldots, n$. $\mathcal{A}$ is the set of arcs, and $\mathcal{W}$ is the set of arc weights. The weight of an arc is a square correlation (denoted as $r_{i,j} = w_{i,j} \in \mathcal{W}$) between two variables. The minimal square correlation $r_{\min}$ can be specified to make the graph sparse and to ignore low correlations. Formally, $a_{i,j} \in \mathcal{A} \iff i, j \in \mathcal{N}$ and $w_{i,j} = r_{i,j} \geq r_{\min}, i \neq j$. In this graph, we neglect loops ($a_{i,i} \notin \mathcal{A}$).

The layout applied for visualization of the correlation graph was the Force Atlas II (FA2) algorithm [40]. This algorithm has a force-directed layout: it simulates a physical system to spatialize a network. Noack [49] showed that force-directed layouts optimize the modularity measure. Communities appear as groups of nodes. FA2 implementation of the adaptive local and global speeds gives good performances for networks of fewer than 100,000 nodes [49].

Step 2: Community detection

Modularity-based community detection algorithms minimize Eq. (7). The result of community detection is a partition of the graph. Isolated nodes and variables with low correlation are classified in a small community; therefore, the minimal number of a variables within a community ($n^c_{\min}$) is specified. The result of the modularity-based community detection is N modules. Modules are subgraphs of the correlation graph. Formally, $C_1, C_2, \ldots, C_N \in G$, $C_I \cap C_J = \emptyset, I, J \in \{1, \ldots, N\}$, $\cup_{I:=1}^{N} C_I \subseteq G$, which is true for all $C_I(\mathcal{N}_I, \mathcal{A}_I, \mathcal{W}_I) \in G \Rightarrow |\mathcal{N}_I| \geq n^c_{\min}$.

Step 3: Specification of the latent variables

LV_I is specified within module I as a linear combination of the centrality variable and the standardized variable v_i (see Eq. (9)). The authors suggested the EVC because of its beneficial property (see Section 2.2.3), but any other centrality measure can be applied.

Step 4 (optional and iterative): Ignore variables with low centrality values

This iterative step involves multiple feature selections. Variables with low centrality values have low weights for specifying the LV; therefore, they can be ignored if their centrality value is lower than a specified $c_{\min}$ value. Ignoring low-centrality-value variables is an iterative process because, in every step, the LV will be modified. Therefore, in one iteration, only the variable with the lowest centrality is ignored. This iterative process ends if the

minimal centrality value for variables within a module is greater than $c_{\min}$.

Step 5 (optional and iterative): Communality analysis

The alternative step of variable selection is to ignore variables based on communality analysis. This communality analysis follows that of PFA [31]. Similar to PCA, EFA and PFA, the factor loading $L_{i,l}$ is the correlation between the LV LV_l and the variable v_i ($L_{i,l} = cor(LV_l, v_i)$). The communality value of variable v_i is $h_i := \max_l L_{i,l}^2$. Similar to step 4, a variable can be ignored if its communality value is lower than a threshold ($h_{\min}$). However, the elimination process is an iterative process, which means that only the variable with the minimal communality value is ignored because, in every cycle, the LV should be recalculated for the remaining variables. This iteration ends if $h_i \geq h_{\min}$ is satisfied for all remaining variables.

Modularity-based community detection methods define distinct communities, where the correlation between communities is much more lower than the correlation within a community; as a result, the correlation between LVs has to be low, and common indicators can exist. To ensure completeness, for the variable (v_i), suppose that $|L_{i,1}| \geq |L_{i,2}| \geq \cdots \geq |L_{i,N}|$ is satisfied. v_i is not a common indicator if either $|L_{i,1}| > |L_{i,2}| + C_{\min}$ or $|L_{i,1}| > 2|L_{i,2}|$[12]; otherwise, it is a common indicator. In an iteration, that common indicator, which has the lowest communality value, is ignored. This iteration ends if there are no more common indicators.

While the assumptions are not satisfied, the iteration returns to step 3 to recalculate the LVs.

At the end of these steps, N LVs are specified (see steps 1–3). Moreover, the assumptions are satisfied. Similar to PFA, the factor loading, factor scores, and communality values can be calculated. These LVs can also be rotated via different kinds of rotation techniques.

Additional applied measures

The sets of indicators were compared by using Jaccard's distance, which is defined as follows:

$$d_J(A, B) = \frac{A \cap B}{A \cup B}, \tag{10}$$

where A, B are the sets of indicators.[13]

4. Results and discussion

4.1. Comparison of NDA with PCA and PFA on non-HDLSS data

The proposed NDA is compared to PFA and PCA on the CWTS'2020 database. Fig. 1(a) shows a scree plot of the indicators. The scree plot suggests that the first two components explain the most indicators; however, in the case of PCA, there are 6 components, while in the case of FA, there are 4 LVs.

The biplots and 3D plots show the correlations between the indicators and the components. The indicators are expected to be grouped and/or fitted into a latent variable. Common indicators correlate more than one LV, and therefore, they do not fit any LV. The biplot of PCA (Fig. 1(b)) shows that in the case of 2 LVs, 2 kinds of groups can be distinguished; however, these components are very mixed. For example, all the open-access (e.g., gold, bronze, hybrid and green) indicators, gender issues, collaborations (collab) and publication output indicators are diffused into these two components. In addition, there are several common indicators, which do not fit with any component (LV).

Similar to PCA, when specifying two latent factors with PFA (see Fig. 2(a)), the set of indicators is diffused within the two factors. The LVs are very difficult to interpret.

In the case of gender issues (PA_F, PA_F_MF, PA_M_MF)[14] Appendix A., the absolute indicators and relative indicators are mostly divided into separate LVs (see Fig. 2(b)). Nevertheless, the absolute and relative indicators are moderately correlated with each other; therefore, several indicators, such as proportional and absolute open-access indicators, belong to the same LV.

In contrast to PFA and PCA, the proposed NDA specifies the number of LVs, which is three in this case. LV_1 corresponds to relative and normalized values, while LV_2 corresponds to absolute values of the publication and collaboration performances of the universities. LV_3 specifies the authorship gender rate (see Fig. 3). The relative or proportional and the absolute indicators are clearly separated into different LVs.

4.2. Comparison of NDA with SPCA on HDLSS data

The real benefit of the proposed algorithm appears when the number of observations is lower than the number of indicators. Since the number of indicators in GovDB'20 is 2160 and the number of observations (i.e., countries) is only 138, PCA and PFA cannot be used. Therefore, in the analysis of the GovDB'20 dataset, NDA is compared to SPCA. Note that although KPCA is also a popular method for analyzing HDLSS datasets, due to the kernel trick it applies, interpretation of the latent variables (principal components) determined by this method is almost impossible.

In the case of SPCA, the number of extracted components (latent variables) was set to four as the input parameter. Fig. 4 shows the 3D representation of the variable loadings.

As Fig. 4 shows, due to the sparse structure of the resulting loadings, the original variables are separated well by the latent variables. Interpretation becomes difficult in the case of more than one hundred variables. Therefore, to interpret the LVs, text mining is applied to the variable names within each LV. The text-mining process involves lowercase transformation, special character removal, part-of-speech (POS) tagging, lemmatization, stop-word removal and tokenization. To visualize the most frequent terms per factor, word clouds are drawn.

In the applied database, each variable is assigned to a specific dataset. To achieve better interpretation, we also analyze the composition of the factors from the dataset perspective. For each term in the word clouds, the relative number of occurrences of the word within the variables of each dataset is calculated. Each term is assigned to the dataset in which its relative occurrence is maximal. Fig. 5 shows the most frequent words and the dataset contribution for each latent variable.

As Fig. 5(a) shows, LV_1 contains infrastructure and business-related variables, LV_2 corresponds to freedom and democracy indicators, and LV_3 is associated with trade and development. LV_4 is associated with indicators from mixed topics. Based on the dataset ratios (Fig. 5(b)), LV_2 and LV_3 have dominance that can be interpreted well in terms of contributing datasets, but LV_1 and LV_4 show small differences, as is also reflected by the wordclouds. In what follows, NDA is tested on the GovDB'20 dataset.

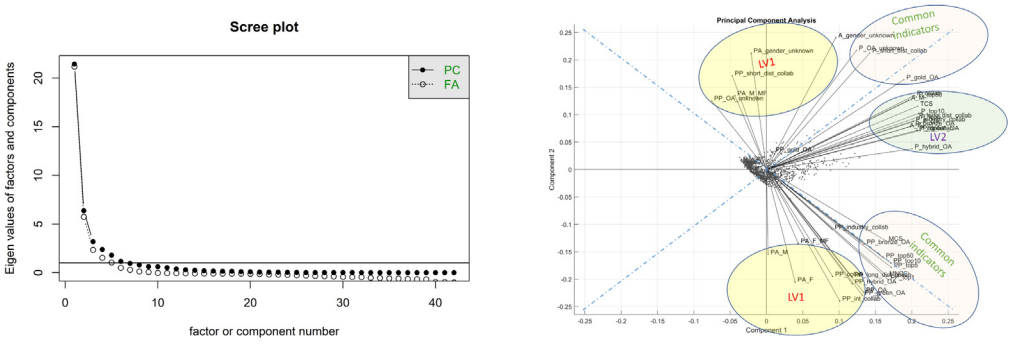
As a first step of NDA, the (square) correlation graph[15] is specified (step 1), where the minimal correlation is specified to be between 0.4 and 0.9 (see Fig. 6). The FA2 layout also shows the potential communities.

[12] $C_{\min} \leq 0.25$ is a constant.

[13] In this paper, $d_J(A, B) \in [0, 1]$ was used only for the comparison of the sets of indicators; however, this measure can also be used instead of the correlations if it is more interpretable than the correlations between indicators.

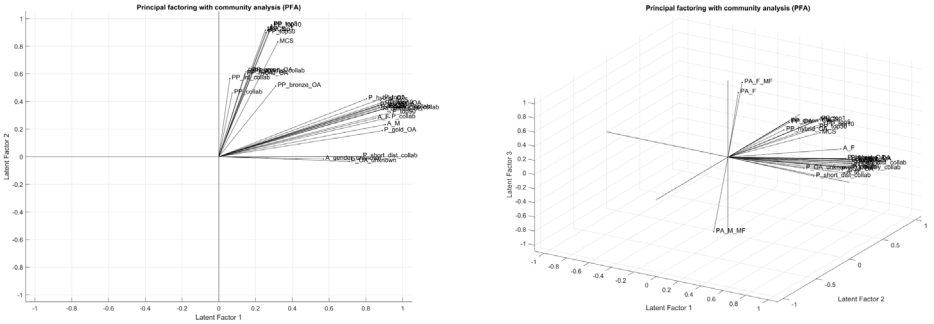
[14] The definitions of these variables can be found in.

[15] The correlation graph of Leiden's CWTS indicators can be found in Appendix C in Fig. 12.



(a) Scree plot of CWTS indicators (b) Biplot of PCA

Fig. 1. Scree plot and biplot of PCA.



(a) Biplot of PFA (b) 3D plot of PFA

Fig. 2. Indicators of PFA ($h_{\min} := 0.2, C_{\min} := 0.1$).

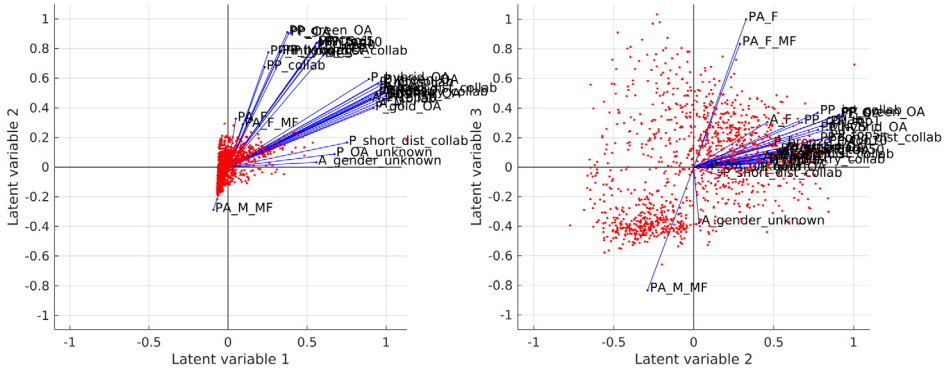


Fig. 3. Biplots of NDA ($h_{\min} := 0.2, C_{\min} := 0.1, c_{\min} := 0.05, n^c_{\min} := 2, r_{\min} := 0$).

Fig. 6 shows a comparison of correlation graphs with different thresholds ($r_{\min}$). These thresholds are applied only for the visualization because a higher threshold produces more isolated points in the correlation graph. In addition, the increase of the correlation threshold $r_{\min}$ can produce more communities. Since the distance correlation between two indicators is 0 if and only if they are independent of each other, we applied the DC; however, this required more than one hundred times the computational time required to calculate Pearson's or Spearman's correlation. Therefore, applying the Spearman correlation may be more practicable for a large-scale dataset. The Spearman correlation can also handle nonlinear and monotonous relationships between indicators. Fig. 7 shows the modules for the correlation graphs. For every correlation graph, both the Leiden and Louvain algorithms

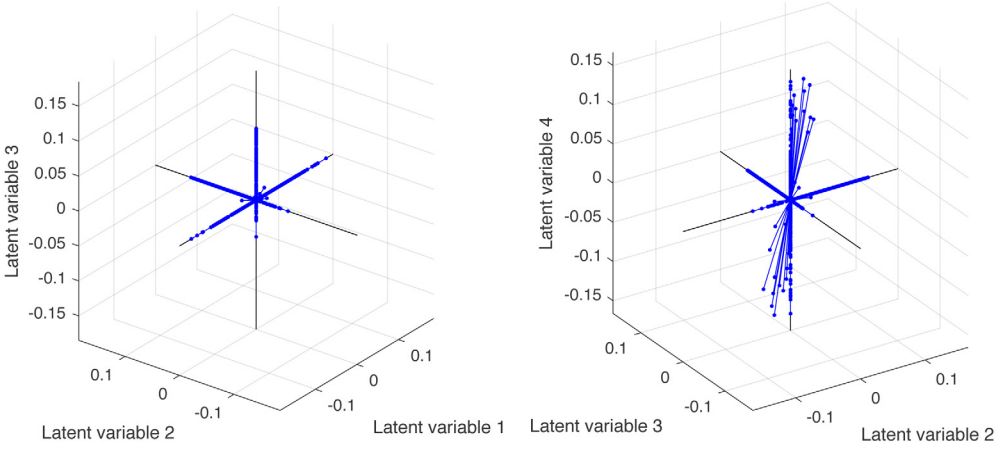


Fig. 4. 3D plots of the selected GovDB'20 variables given by the SPCA.

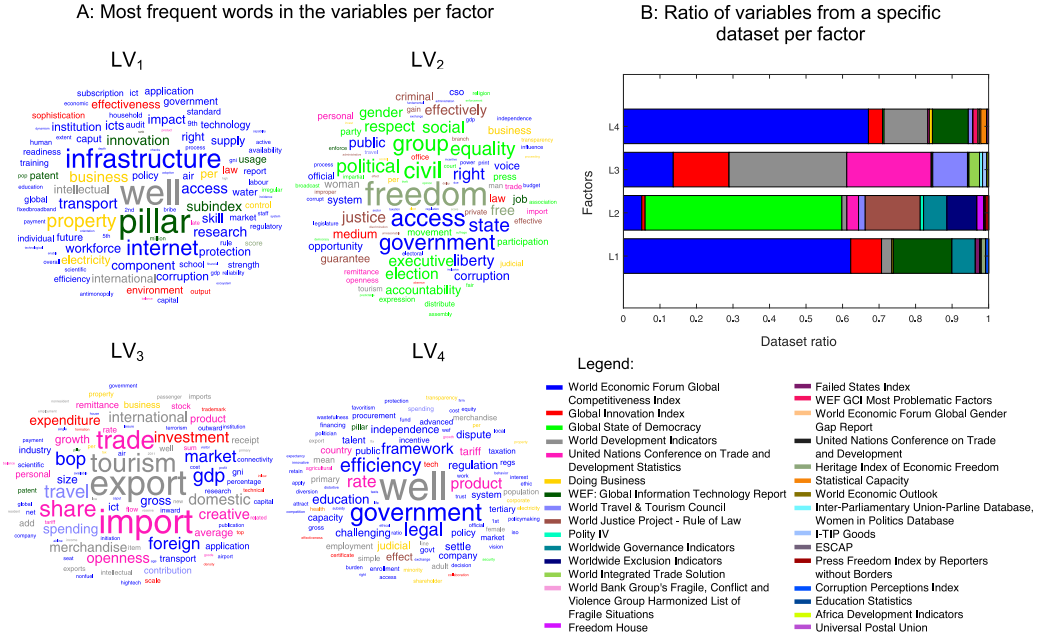


Fig. 5. Representation of the latent GovDB'20 variables given by the SPCA.

specify 4 communities. For a better interpretation, only those correlations that exceed the threshold ($r_{\min}$) are represented (see Fig. 7).

The results of Leiden's and Louvain's modularities are very similar (compare Fig. 7(a–b) and (c–d)). The mean Jaccard distance between the adequate Leiden and Louvain modules is 0.98, while the mean Jaccard distance of adequate modules on the Spearman and distance correlation graphs is still 0.85. Because of the beneficial properties of the distance correlation and Leiden's modularity algorithm, in this example, these algorithms are applied.

The remaining hyperparameters ($c_{\min}, h_{\min}, C_{\min}$) do not influence the number of communities (i.e., the number of latent variables), but they increase interpretability by satisfying Assumptions 1–2. Fig. 8 shows the effects of feature selection.

By increasing the minimum EVC ($c_{\min}$), peripheral indicators are excluded (see Fig. 8(a-e)). Nevertheless, this feature selection alone does not ensure better interpretability of the LVs. Table 1 shows that an increase in the minimal centrality value ($c_{\min}$ does not affect the minimal communality value ($\min(h)$). The low communality value indicates that there are indicators that are not correlated with any LV.

Table 2 shows the effect of increasing the minimal communality constraints. Increasing the minimal communality constraint ensures that all indicators correlate with at least one LV. Nevertheless, in this case, both core and peripheral indicators may be excluded (see Fig. 8(f-j)). Therefore, the minimal EVC is low (see Table 2).

An increase in both the minimal communality and minimal centrality constraint can ensure the satisfaction of Assumption 1. Hence, LVs are the core indicators, which are correlated to the adequate LV. The correlation of LVs can be high (see $\max_{ij} |r_{LV_{ij}}|$ columns in Tables 1–2). Therefore, the common communality constraint ($C_{\min}$) must be increased to improve the separation of

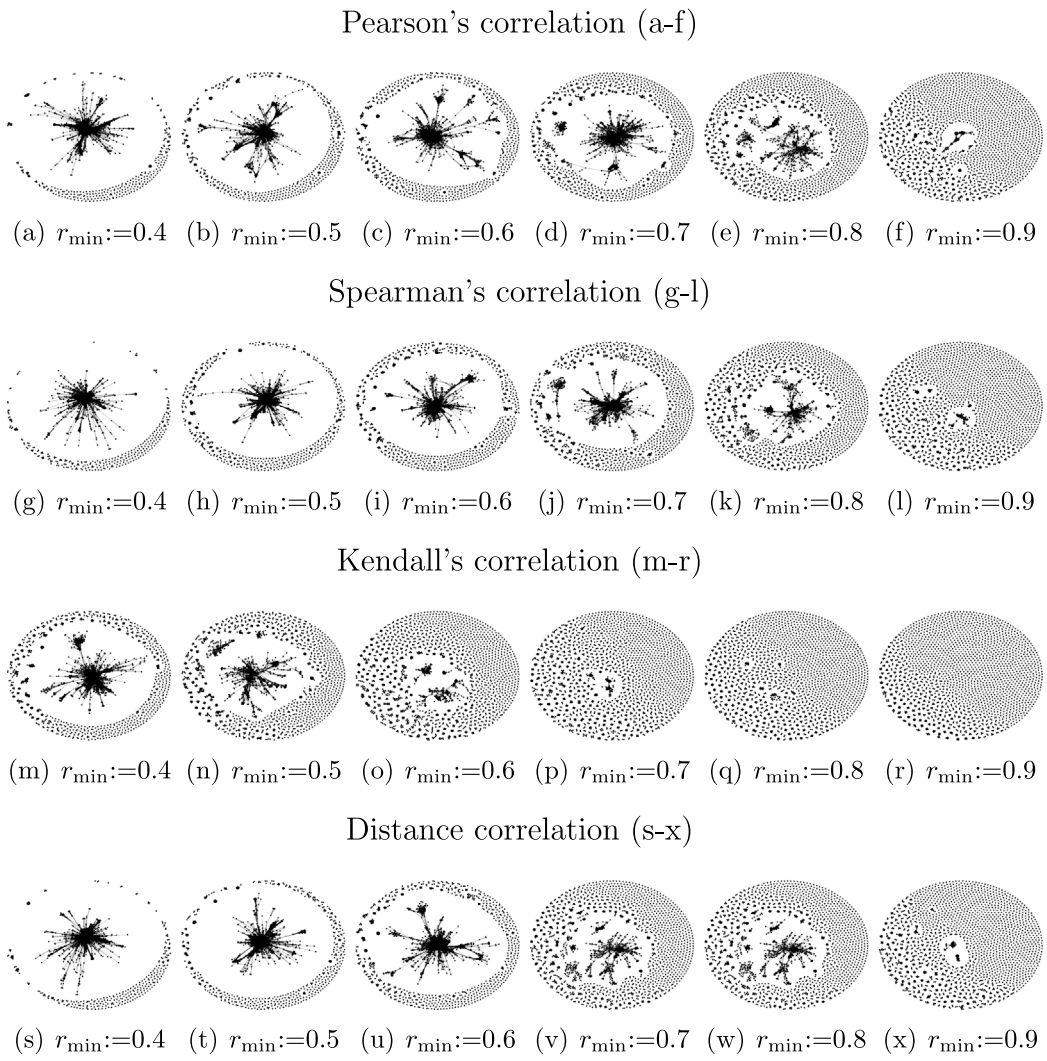


Fig. 6. Correlation graphs (FA2 layout).

Table 1
The effect of the centrality constraint ($c_{\min}$).

$c_{\min}$	$h_{\min}$	n_1	n_2	n_3	n_4	min(c)	min(h)	max$_{ij}$ $\lvert r_{LV_{ij}}\rvert$
0.00	0	592	680	328	558	0.0002	0.0016	0.9961
0.01	0	578	662	312	390	0.0099	0.0021	0.9981
0.02	0	529	558	266	250	0.0197	0.0015	0.9270
0.03	0	458	410	234	205	0.0298	0.0008	0.9984
0.04	0	351	257	197	187	0.0401	0.0010	0.9885
0.05	0	110	146	155	154	0.0501	0.0025	0.9980

Notations: $c_{\min}$: minimal centrality constraint; $h_{\min}$: minimal communality constraint; n_i: number of indicators in module i; min(c): minimal centrality value; min(h): minimal communality value; max$_{ij}$($r_{LV_{ij}}$): maximal absolute correlation value between latent variables.

Table 2
The effect of the communality constraint ($h_{\min}$).

$c_{\min}$	$h_{\min}$	n_1	n_2	n_3	n_4	min(c)	min(h)	max$_{ij}$ $\lvert r_{LV_{ij}}\rvert$
0	0.00	592	680	328	558	0.0002	0.0016	0.9961
0	0.05	283	307	187	241	0.0047	0.0501	0.9967
0	0.10	148	141	137	123	0.0116	0.1008	0.9964
0	0.15	72	61	101	77	0.0054	0.1702	0.9942
0	0.20	51	43	74	63	0.0035	0.1994	0.9938
0	0.25	32	33	53	55	0.0028	0.2468	0.9936

Notations: $c_{\min}$: minimal centrality constraint; $h_{\min}$: minimal communality constraint; n_i: number of indicators in module i; min(c): minimal centrality value; min(h): minimal communality value; max$_{ij}$($r_{LV_{ij}}$): maximal absolute correlation value between latent variables.

LVs. The orthogonal rotation methods can also be used to separate factors; however, if the unrotated LV matrix is highly correlated, the variables cannot be separated, and as a result, the LVs will be difficult to interpret. Therefore, all constraints, such as the minimal centrality constraint ($c_{\min}$), the minimal communality constraint ($h_{\min}$) and the common communality constraint ($C_{\min}$), should be specified to decrease the maximal correlation values between LVs max$_{ij}$ $\lvert r_{LV_{ij}}\rvert$ and increase the minimal communality value min(h) and minimal centrality value min(c). Fig. 9 shows a 3D plot of the selected variables, the minimal centrality (communality) of the variables and the maximal correlation between LVs.

The results show that all assumptions are satisfied. In addition, the indicators can be separated. However, Fig. 9 involves only four group indicators; the already introduced text-mining approach is applied (see also Fig. 5) to interpret the LVs. Fig. 10 presents the results of NDA without performing feature selection.

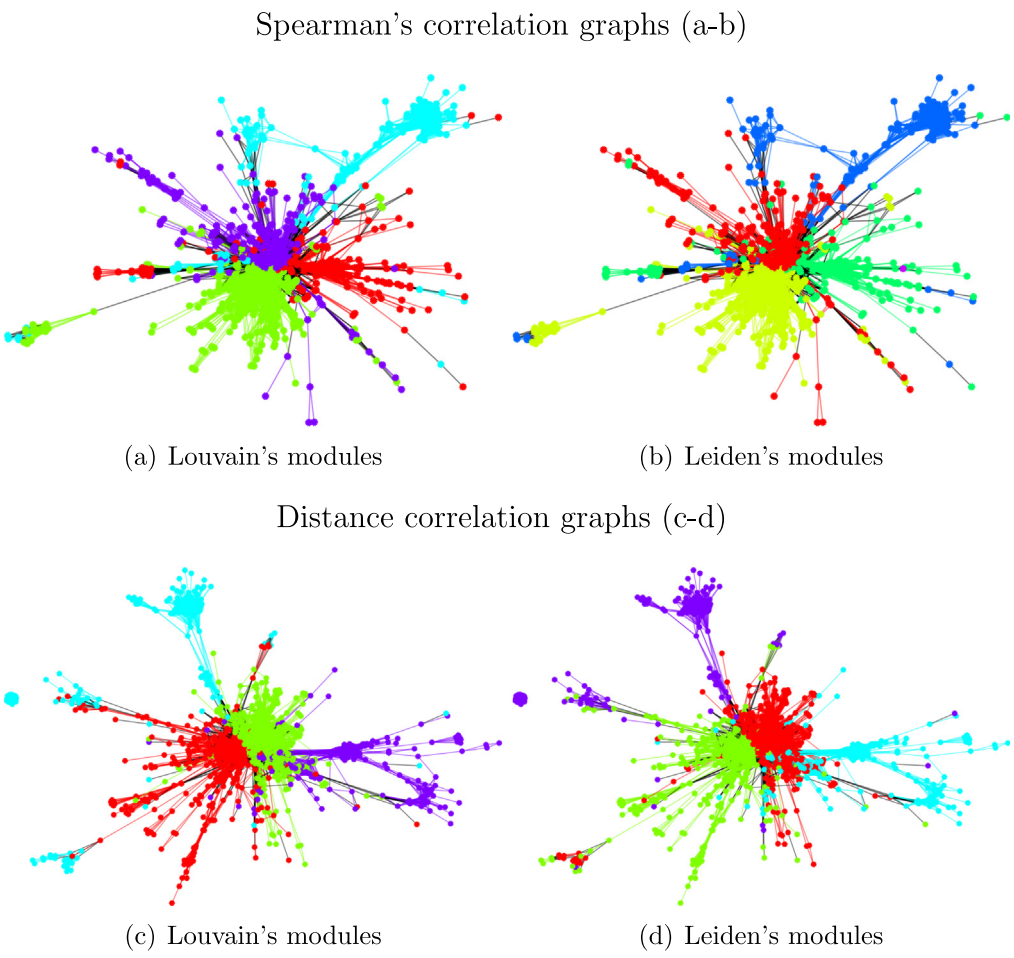


Fig. 7. Correlation graphs obtained with modularity-based community detection algorithms ($r_{\min} := 0.6$).

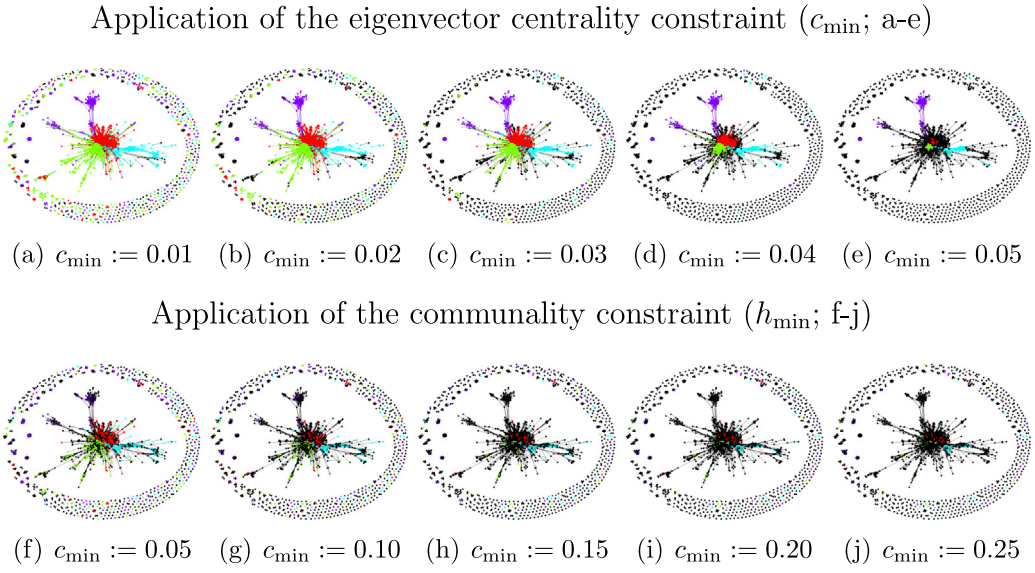


Fig. 8. Characteristics of feature selection.

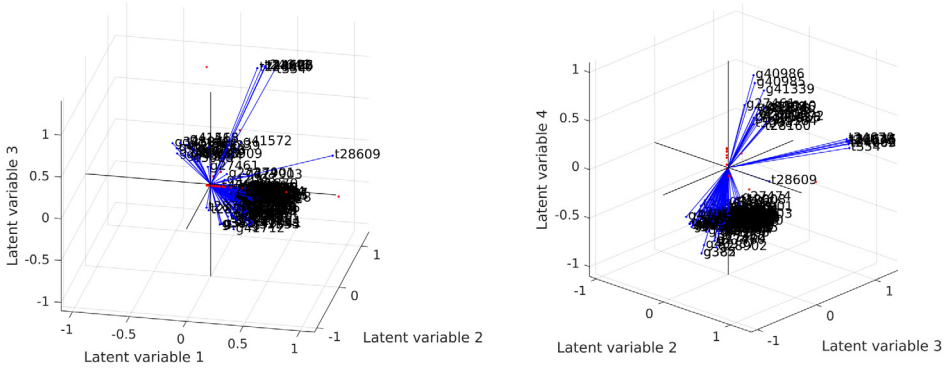


Fig. 9. 3D plots of selected variables, with $n_1 = 146$, $n_2 = 190$, $n_3 = 36$, $n_4 = 51$, $\min(c) = 0.1031$, $\min(h) = 0.0661$, $\max_{ij} |r_{LV_{ij}}| = 0.1205$.

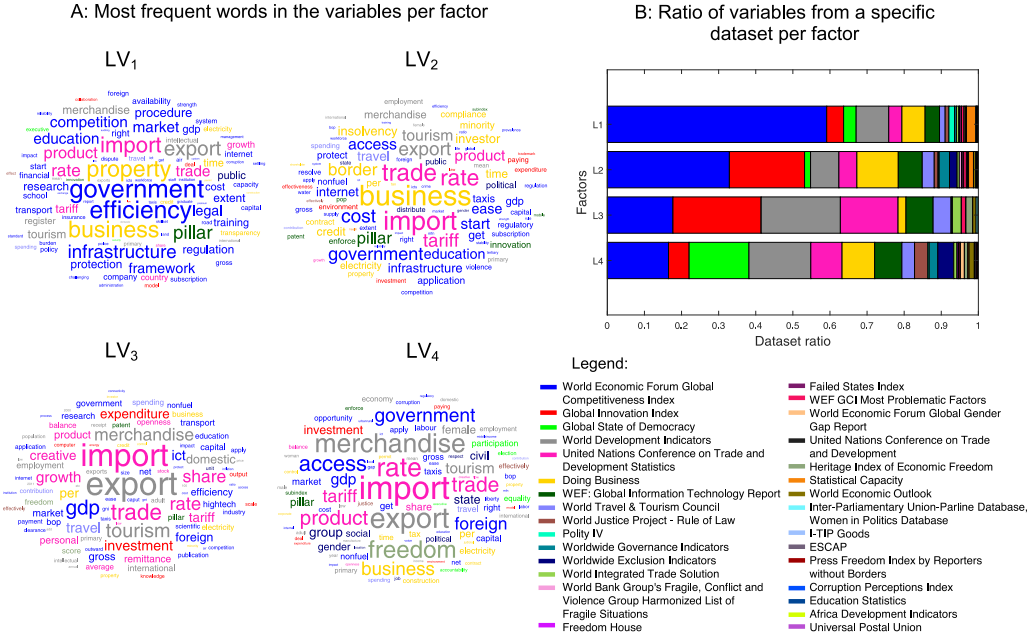


Fig. 10. Representation of the latent variables without selection.

Table 3
Correlation matrix of the latent variables (*LVs*)

	LV_1	LV_2	LV_3	LV_4
LV_1	1.0000	0.0024	−0.0236	−0.0548
LV_2	0.0024	1.0000	−0.0106	0.1107
LV_3	−0.0236	0.0106	1.0000	0.1205
LV_4	−0.0548	0.1107	0.1205	1.0000

Fig. 10 shows that the four groups are very mixed. Several words, such as import, trade, etc., are included in every group. Despite the variables being different, variable selection is required to filter out any irrelevant and common variables.

The variable selection saves the most relevant variables ($n_1 = 146, n_2 = 190, n_3 = 36, n_4 = 51$); nevertheless, it increases the communality and centrality values while keeping the LVs as uncorrelated as possible ($\min(c) = 0.1031$, $\min(h) = 0.0661$, $\max_{ij} |r_{LV_{ij}}| = 0.1205$). Table 3 shows the correlation factor matrix of the LVs.

Similar to Fig. 10, the results obtained after variable selection are shown in Fig. 11.

Fig. 11 helps us to interpret the LVs. Starting the interpretation from the last LV, LV_4 corresponds to the *freedom and democracy* indicators. Most of these indicators come from the Global State of Democracy database. The third LV (LV_3) corresponds to the global *trade and development* indicators, such as the import, export, GDP and trade indicators. Most of these variables come from the World Development Indicators and United Conference on Trade and Development statistics. The second LV (LV_2) corresponds mainly to indicators related to *business and investment*. Most of these indicators come from the World Economic Forum, Global Innovation Index and Doing Business datasets. The first LV (LV_1) corresponds mainly to the *research and economy* indicators. These indicators come mostly from the World Economic Forum dataset. Two out of the four resulting topics are also observable in the results of SPCA; however, the latent variables are easier to interpret, and NDA reveals two additional distinguishable topics, *research and economy* and *business and investment* (see Figs. 5 and 11).

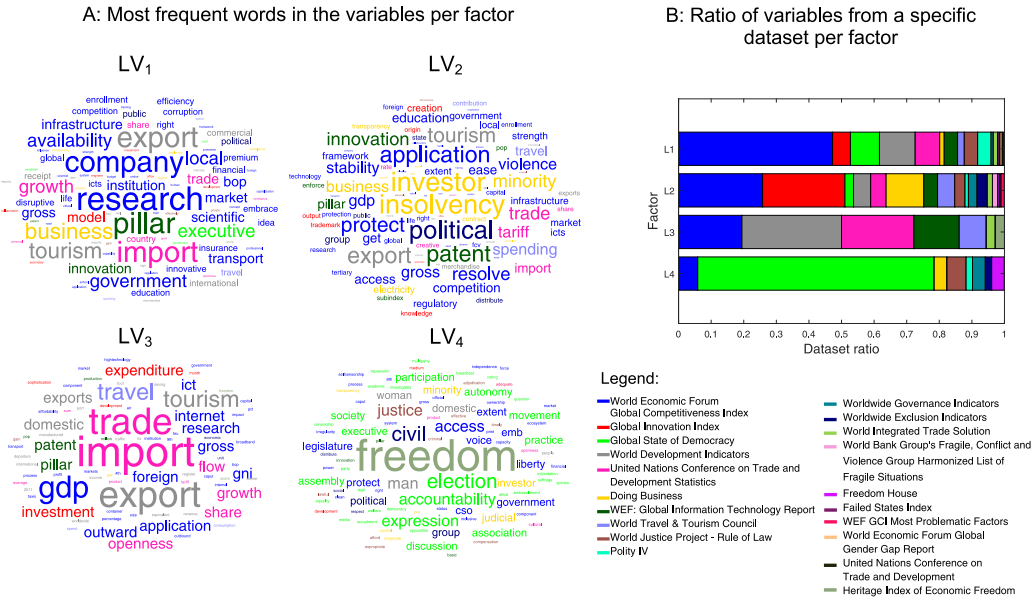


Fig. 11. Representation of the resulting latent variables after variable selection. LV_1 = Research & Economy, LV_2 = Business & Investment, LV_3 = Trade & Development, LV_4 = Freedom & Democracy.

Settings of NDA. The NDA is a nonparametric data reduction method. The term nonparametric means that the number of LVs is the result of dimensionality reduction; however, it can be controlled by several constraints, such as the minimal variables in a community and the minimal variable correlation. The increase in the minimal correlation ($r_{\min}$) separates the communities in the correlation graphs; therefore, it can increase the number of LVs (see Fig. 6). However, after community detection, none of the hyperparameters influence the number of LVs, only their contents via feature selection. The applied modularity-based community detection methods group highly correlated variables into a module, and within a module, an LV can be specified as a linear combination of the variables and their EVCs. If constraints are neglected (formally: $c_{\min} = h_{\min} = C_{\min} = 0$), feature selection is not provided. However, the increase in the centrality ($c_{\min}$) and communality constraints ($h_{\min}$) can reduce irrelevant indicators, while the increase in the common communality constraint ($C_{\min}$) can reduce the set of common indicators; in this way, it can reduce the correlations between LVs to promote the interpretability of LVs. After variable selection, independent LVs and their indicators can be specified even if there are few observations. As we introduced the usage of NDA, at first, it should be used without feature selection ($c_{\min} = h_{\min} = C_{\min} = 0$) in order to specify the number of communities. If the modules are stable, similarly to any clustering algorithm, the hyperparameters should be increased one by one to achieve the specified thresholds to satisfy Assumptions 1–2.

5. Summary and conclusions

This study proposes a nonparametric network-based dimensionality reduction method (NDA). The proposed NDA method is ideal for HDLSS datasets, where the number of variables is much larger than the number of observations, and furthermore its application shows several advantages compared to the existing methods.

First, it can be effectively applied to any datasets due to the built-in feature selection module. Second, NDA shows better interpretability then the results given by SPCA. Third, its application is easier since there is no need to specify the expected number of latent variables due to its nonparametric characteristics.

The NDA is implemented in both R and MATLAB. The authors will provide freely available packages for both R and MATLAB. In the next paper, we will show how to extend this algorithm to handle topic mining problems. In the case of topic mining when segmenting the so-called document term matrix, we have to face a similar problem. The number of topics also has to be predefined, and NDA provides the number of topics. In addition to feature selection, the frequency of common terms can be reduced. This method will be described in the next paper.

CRediT authorship contribution statement

Zsolt T. Kosztyán: Conception and design of study, Acquisition of data, Analysis and/or interpretation of data, Writing – original draft, Writing – review & editing. **Marcell T. Kurbucz:** Conception and design of study, Acquisition of data, Analysis and/or interpretation of data, Writing – original draft, Writing – review & editing. **Attila I. Katona:** Conception and design of study, Acquisition of data, Analysis and/or interpretation of data, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the TKP2020-NKA-10 project financed under the 2020-4.1.1-TKP2020 Thematic Excellence Programme by the National Research, Development and Innovation Fund of Hungary and by the Research Centre of the Faculty of Business and Economics (No. PE-GTK-GSKK A095000000-1) at the University of Pannonia (Veszprém, Hungary). All authors approved the version of the manuscript to be published.

Appendix A. Indicators of the CWTS Leiden Ranking 2020 database

See Table 4.

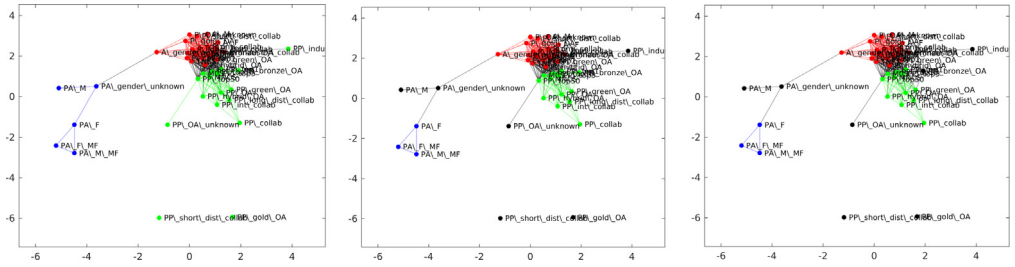
Table 4
Indicators of the CWTS Leiden Ranking 2020 database.

	Variable	Definition
Scientific impact	P	Total number of publications of a university.
	P(top 1%)	The number 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.
	P(top 5%)	The number of a university's publications that, compared with other publications in the same field and in the same year, belong to the top 5% most frequently cited.
	P(top 10%)	The number of a university's publications that, compared with other publications in the same field and in the same year, belong to the top 10% most frequently cited.
	P(top 50%)	The number of a university's publications that, compared with other publications in the same field and in the same year, belong to the top 50% most frequently cited.
	PP(top 1%)	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.
	PP(top 5%)	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 5% most frequently cited.
	PP(top 10%)	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 10% most frequently cited.
	PP(top 50%)	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 5% most frequently cited.
	TCS	The total number of citations of the publications of a university.
	MCS	The average number of citations of the publications of a university.
	TNCS	The total number of citations of the publications of a university, normalized for field and publication year.
	MNCS	The average number of citations of the publications of a university, normalized for field and publication year.
Collaboration	P(collab)	The number of a university's publications that have been coauthored with one or more other organizations.
	PP(collab)	The proportion of a university's publications that have been coauthored with one or more other organizations.
	P(int collab)	The number of a university's publications that have been coauthored by two or more countries.
	PP(int collab)	The proportion of a university's publications that have been coauthored by two or more countries.
	P(industry)	The number of a university's publications that have been coauthored with one or more industrial organizations.
	PP(industry)	The proportion of a university's publications that have been coauthored with one or more industrial organizations.
	P(100 km)	The number of a university's publications with a geographical collaboration distance of less than 100 km.
	PP(100 km)	The proportion of a university's publications with a geographical collaboration distance of less than 100 km.
	P(5000 km)	The number of a university's publications with a geographical collaboration distance of more than 5000 km.
	PP(5000 km)	The proportion of a university's publications with a geographical collaboration distance of more than 5000 km.
Open access	P(OA)	The number of open-access publications of a university.
	PP(OA)	The proportion of open-access publications of a university.
	P(gold OA)	The number of gold open-access publications of a university.
	PP(gold OA)	The proportion of gold open-access publications of a university.
	P(hybrid OA)	The number of hybrid open-access publications of a university.
	PP(hybrid OA)	The proportion of hybrid open-access publications of a university.
	P(bronze OA)	The number of bronze open-access publications of a university.
	PP(bronze OA)	The proportion of bronze open-access publications of a university.
	P(green OA)	The number of green open-access publications of a university.
	PP(green OA)	The proportion of green open-access publications of a university.
	P(unknown OA)	The number of a university's publications for which the open-access status is unknown.
	PP(unknown OA)	The proportion of a university's publications for which the open-access status is unknown.
Gender	A	The total number of authorships of a university.
	A(MF)	The number of male and female authorships of a university, that is, a university's number of authorships for which the gender is known.
	A(unknown)	The number of authorships of a university for which the gender is unknown.
	PA(unknown)	The number of authorships for which the gender is unknown as a proportion of a university's total number of authorships.
	A(M)	The number of male authorships of a university.
	PA(M)	The number of male authorships as a proportion of a university's total number of authorships.
	PA(M|MF)	The number of male authorships as a proportion of a university's number of male and female authorships.
	A(F)	The number of female authorships of a university.
	PA(F)	The number of female authorships as a proportion of a university's total number of authorships.
	PA(F|MF)	The number of female authorships as a proportion of a university's number of male and female authorships.

Table 5
GovDB'20 dataset.

Source dataset	Source platform	Number of variables
Africa Development Indicators	TCdata360	1
Corruption Perceptions Index	GovData360	2
Doing Business	GovData360, TCdata360	164, 5
Education Statistics	TCdata360	1
ESCAP	TCdata360	3
Failed States Index	GovData360	7
Freedom House	GovData360	4
Global Innovation Index	TCdata360	274
Global State of Democracy	GovData360	121
Heritage Index of Economic Freedom	TCdata360	13
Inter-Parliamentary Union - Parline Database, Women in Politics Database	GovData360	3
I-TIP Goods	TCdata360	4
Polity IV	GovData360	13
Poverty and Equity Data	GovData360	1
Press Freedom Index by Reporters without Borders	GovData360	2
Statistical Capacity	GovData360	26
United Nations Conference on Trade and Development	TCdata360	2
United Nations Conference on Trade and Development Statistics	TCdata360	153
Universal Postal Union	TCdata360	1
WEF GCI Most Problematic Factors	TCdata360	15
WEF: Global Information Technology Report	TCdata360	132
World Bank Group's Fragile, Conflict and Violence Group Harmonized List of Fragile Situations	TCdata360	4
World Development Indicators	TCdata360	132
World Economic Forum Global Competitiveness Index	GovData360, TCdata360	1, 267
World Economic Forum Global Gender Gap Report	GovData360, TCdata360	648, 76
World Economic Outlook	TCdata360	10
World Integrated Trade Solution	TCdata360	16
World Justice Project - Rule of Law	TCdata360	21
World Travel & Tourism Council	GovData360	33
Worldwide Exclusion Indicators	TCdata360	66
Worldwide Governance Indicators	GovData360	38
	GovData360	36

The whole dataset is publicly available at [50]. Details of the GovDB'20 variables employed in this paper can be found at data.mendeley.com/datasets/hzdnxph8vg/6 (retrieved: May 14, 2021).



(a) Modules of the correlation graph ($n_c = 3$)

(b) Core indicators of correlation graphs ($h_{\min} := 0.2$, $c_{\min} := 0.05$) (Assumption 1)

(c) Uncorrelated core indicators of correlation graphs ($C_{\min} := 0.1$) (Assumption 2)

Fig. 12. Modules of the correlation graph of the CWTS indicators and their modification after the feature selection ($r_{\min} = 0.5$).

Appendix B. GovDB'20 dataset

See Table 5.

Appendix C. Correlation graphs of CWTS indicators

See Fig. 12.

References

[1] Alexander N. Gorban, Nikolaos K. Kazantzis, Ioannis G. Kevrekidis, Hans Christian Öttinger, Constantinos Theodoropoulos, Model Reduction and Coarse-Graining Approaches for Multiscale Phenomena, Springer, 2006, http://dx.doi.org/10.1007/3-540-35888-9.

[2] Alexander N. Gorban, Valery A. Makarov, Ivan Y. Tyukin, High-dimensional brain in a high-dimensional world: Blessing of dimensionality, Entropy 22 (1) (2020) 82, http://dx.doi.org/10.3390/e22010082.

[3] Alexander N. Gorban, Balázs Kégl, Donald C. Wunsch, Andrei Y. Zinovyev, et al., Principal Manifolds for Data Visualization and Dimension Reduction, Vol. 58, Springer, 2008, http://dx.doi.org/10.1007/978-3-540-73750-6.

[4] Alexander N. Gorban, Andrei Zinovyev, Principal manifolds and graphs in practice: from molecular biology to dynamical systems, Int. J. Neural Syst. 20 (03) (2010) 219–232, http://dx.doi.org/10.1142/S0129065710002383.

[5] Andrzej Cichocki, Namgil Lee, Ivan Oseledets, Anh-Huy Phan, Qibin Zhao, Danilo P. Mandic, Tensor networks for dimensionality reduction and large-scale optimization: Part 1 low-rank tensor decompositions, Found. Trends® Mach. Learn. 9 (4–5) (2016) 249–429, http://dx.doi.org/10.1561/2200000059.

[6] Andrzej Cichocki, Anh-Huy Phan, Qibin Zhao, Namgil Lee, Ivan Oseledets, Masashi Sugiyama, Danilo P. Mandic, Tensor networks for dimensionality reduction and large-scale optimization: Part 2 applications and future perspectives, Found. Trends® Mach. Learn. (ISSN: 1935-8237) 9 (6) (2017) 431–673, http://dx.doi.org/10.1561/2200000067.

[7] Mohammad Sultan Mahmud, Xianghua Fu, Joshua Zhexue Huang, Md Abdul Masud, High-dimensional limited-sample biomedical data classification

using variational autoencoder, in: Australasian Conference on Data Mining, Springer, 2018, pp. 30–42, http://dx.doi.org/10.1007/978-981-13-6661-1_3.

[8] Mohammad Sultan Mahmud, Xianghua Fu, Unsupervised classification of high-dimension and low-sample data with variational autoencoder based dimensionality reduction, in: 2019 IEEE 4th International Conference on Advanced Robotics and Mechatronics (ICARM), IEEE, 2019, pp. 498–503, http://dx.doi.org/10.1109/ICARM.2019.8834333.

[9] Mohammad Sultan Mahmud, Joshua Zhexue Huang, Xianghua Fu, Rukhsana Ruby, Kaishun Wu, Unsupervised adaptation for high-dimensional with limited-sample data classification using variational autoencoder, Comput. Inform. 40 (1) (2021) 1–28, http://dx.doi.org/10.31577/cai_2021_1_1.

[10] Marcel Dettling, Peter Bühlmann, Boosting for tumor classification with gene expression data, Bioinformatics 19 (9) (2003) 1061–1069.

[11] Le Zhang, Ponnuthurai Nagaratnam Suganthan, Random forests with ensemble of feature spaces, Pattern Recognit. 47 (10) (2014) 3429–3437, http://dx.doi.org/10.1016/j.patcog.2014.04.001.

[12] Shadi Abpeykar, Mehdi Ghatee, Neural trees with peer-to-peer and server-to-client knowledge transferring models for high-dimensional data classification, Expert Syst. Appl. 137 (2019) 281–291, http://dx.doi.org/10.1016/j.eswa.2019.07.003.

[13] Shadi Abpeykar, Mehdi Ghatee, An ensemble of RBF neural networks in decision tree structure with knowledge transferring to accelerate multi-classification, Neural Comput. Appl. 31 (11) (2019) 7131–7151, http://dx.doi.org/10.1007/s00521-018-3543-9.

[14] Shadi Abpeykar, Mehdi Ghatee, Hadi Zare, Ensemble decision forest of RBF networks via hybrid feature clustering approach for high-dimensional data classification, Comput. Statist. Data Anal. 131 (2019) 12–36, http://dx.doi.org/10.1016/j.csda.2018.08.015.

[15] Nitin Khosla, Dimensionality Reduction Using Factor Analysis, Griffith University, Australia, 2004, http://dx.doi.org/10.25904/1912/3890.

[16] M. Usman Ali, Shahzad Ahmed, Javed Ferzund, Atif Mehmood, Abbas Rehman, Using PCA and factor analysis for dimensionality reduction of bio-informatics data, 2017, arXiv preprint arXiv:1707.07189.

[17] I.T. Jolliffe, Springer-Verlag, Principal Component Analysis, in: Springer Series in Statistics, Springer, ISBN: 9780387954424, 2002, URL https://books.google.hu/books?id=_olByCrhjwIC.

[18] Hervé Abdi, Lynne J. Williams, Principal component analysis, WIREs Comput. Statist. 2 (4) (2010) 433–459, http://dx.doi.org/10.1002/wics.101, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/wics.101.

[19] Gianmarco Aversano, Zhiyi Li, Olivier Gicquel, Alessandro Parente, Model reduction by PCA and kriging, in: International Conference of Computational Methods in Sciences and Engineering, 2018.

[20] Yugo Nakayama, Kazuyoshi Yata, Makoto Aoshima, Clustering by principal component analysis with Gaussian kernel in high-dimension, low-sample-size settings, J. Multivariate Anal. (2021) 104779, http://dx.doi.org/10.1016/j.jmva.2021.104779.

[21] Chih-Fong Tsai, Ya-Ting Sung, Ensemble feature selection in high dimension, low sample size datasets: Parallel and serial combination approaches, Knowl.-Based Syst. 203 (2020) 106097, http://dx.doi.org/10.1016/j.knosys.2020.106097.

[22] Aman Gupta, Haohan Wang, Madhavi Ganapathiraju, Learning structure in gene expression data using deep architectures, with an application to gene clustering, in: 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2015, pp. 1328–1335, http://dx.doi.org/10.1109/BIBM.2015.7359871.

[23] Bernhard Schölkopf, Alexander Smola, Klaus-Robert Müller, Kernel principal component analysis, in: International Conference on Artificial Neural Networks, Springer, 1997, pp. 583–588, http://dx.doi.org/10.1007/BFb0020217.

[24] Bernhard Schölkopf, Alexander Smola, Klaus-Robert Müller, Nonlinear component analysis as a kernel eigenvalue problem, Neural Comput. 10 (5) (1998) 1299–1319, http://dx.doi.org/10.1162/089976698300017467.

[25] Zhenqiu Liu, Dechang Chen, Halima Bensmail, Gene expression data classification with kernel principal component analysis, J. Biomed. Biotechnol. 2005 (2) (2005) 155, http://dx.doi.org/10.1155/JBB.2005.155.

[26] Ferran Reverter, Esteban Vegas, Pedro Sánchez, Mining gene expression profiles: an integrated implementation of kernel principal component analysis and singular value decomposition, Genom., Proteom. Bioinform. 8 (3) (2010) 200–210, http://dx.doi.org/10.1016/S1672-0229(10)60022-8.

[27] Dan Shen, Haipeng Shen, James Stephen Marron, Consistency of sparse PCA in high dimension, low sample size contexts, J. Multivariate Anal. 115 (2013) 317–333, http://dx.doi.org/10.1016/j.jmva.2012.10.007.

[28] Hui Zou, Trevor Hastie, Robert Tibshirani, Sparse principal component analysis, J. Comput. Graph. Statist. 15 (2) (2006) 265–286, http://dx.doi.org/10.1198/106186006X113430.

[29] Karl Sjöstrand, Mikkel B. Stegmann, Rasmus Larsen, Sparse principal component analysis in medical shape modeling, in: Medical Imaging 2006: Image Processing, Vol. 6144, International Society for Optics and Photonics, 2006, 61444X, http://dx.doi.org/10.1007/s10994-021-06025-3.

[30] Haiyan Jiang, Haoyi Xiong, Dongrui Wu, Ji Liu, Dejing Dou, AgFlow: fast model selection of penalized PCA via implicit regularization effects of gradient flow, Mach. Learn. 110 (8) (2021) 2131–2150, http://dx.doi.org/10.1007/s10994-021-06025-3.

[31] An Gie Yong, Sean Pearce, et al., A beginner's guide to factor analysis: Focusing on exploratory factor analysis, Tutor. Quant. Methods Psychol. 9 (2) (2013) 79–94, http://dx.doi.org/10.20982/tqmp.09.2.p079.

[32] Leandre R. Fabrigar, Duane T. Wegener, Exploratory Factor Analysis, Oxford University Press, 2011.

[33] Rudolf J. Rummel, Applied Factor Analysis, Northwestern University Press, 1988.

[34] Hee-Ju Kim, Common factor analysis versus principal component analysis: Choice for symptom cluster research, Asian Nurs. Res. (ISSN: 1976-1317) 2 (1) (2008) 17–24, http://dx.doi.org/10.1016/S1976-1317(08)60025-0, URL https://www.sciencedirect.com/science/article/pii/S1976131708600250.

[35] Hervé Abdi, Factor rotations in factor analyses, in: Encyclopedia for Research Methods for the Social Sciences, Sage, Thousand Oaks, CA, 2003, pp. 792–795.

[36] Leandre R. Fabrigar, Duane T. Wegener, Robert C. MacCallum, Erin J. Strahan, Evaluating the use of exploratory factor analysis in psychological research., Psychol. Methods 4 (3) (1999) 272, http://dx.doi.org/10.1037/1082-989X.4.3.272.

[37] Robert C. MacCallum, Keith F. Widaman, Shaobo Zhang, Sehee Hong, Sample size in factor analysis., Psychol. Methods 4 (1) (1999) 84, http://dx.doi.org/10.1037/1082-989X.4.1.84.

[38] Keith E. Muller, Yueh-Yun Chi, Jeongyoun Ahn, J.S. Marron, Limitations of high dimension, low sample size principal components for Gaussian data, 2008, Under Revision for Resubmission.

[39] Sungkyu Jung, J. Stephen Marron, PCA consistency in high dimension, low sample size context, Ann. Statist. 37 (6B) (2009) 4104–4130, http://dx.doi.org/10.1214/09-AOS709.

[40] Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann, Mathieu Bastian, ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software, PLoS One 9 (6) (2014) 1–12, http://dx.doi.org/10.1371/journal.pone.0098679.

[41] Michael T. Brannick, Paul E. Spector, Estimation problems in the block-diagonal model of the multitrait-multimethod matrix, Appl. Psychol. Meas. 14 (4) (1990) 325–339, http://dx.doi.org/10.1177/014662169001400401.

[42] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre, Fast unfolding of communities in large networks, J. Stat. Mech. Theory Exp. 2008 (10) (2008) P10008, http://dx.doi.org/10.1088/1742-5468/2008/10/p10008.

[43] V.A. Traag, L. Waltman, N.J. van Eck, From louvain to leiden: guaranteeing well-connected communities, Sci. Rep. (ISSN: 2045-2322) 9 (1) (2019) 5233, http://dx.doi.org/10.1038/s41598-019-41695-z.

[44] Gábor J. Székely, Maria L. Rizzo, Brownian distance covariance, Ann. Appl. Stat. 3 (4) (2009) 1236–1265, http://dx.doi.org/10.1214/09-AOAS312.

[45] Marcell Tamás Kurbucz, A joint dataset of official COVID-19 reports and the governance, trade and competitiveness indicators of world bank group platforms, Data Brief (ISSN: 2352-3409) 31 (2020) 105881, http://dx.doi.org/10.1016/j.dib.2020.105881, URL https://www.sciencedirect.com/science/article/pii/S2352340920307757.

[46] Gábor J. Székely, Maria L. Rizzo, The distance correlation t-test of independence in high dimension, J. Multivariate Anal. (ISSN: 0047-259X) 117 (2013) 193–213, http://dx.doi.org/10.1016/j.jmva.2013.02.012, URL https://www.sciencedirect.com/science/article/pii/S0047259X13000262.

[47] Mark E.J. Newman, Modularity and community structure in networks, Proc. Natl. Acad. Sci. 103 (23) (2006) 8577–8582, http://dx.doi.org/10.1073/pnas.0601602103.

[48] Britta Ruhnau, Eigenvector-centrality — a node-centrality? Social Networks (ISSN: 0378-8733) 22 (4) (2000) 357–365, http://dx.doi.org/10.1016/S0378-8733(00)00031-9, URL https://www.sciencedirect.com/science/article/pii/S0378873300000319.

[49] Andreas Noack, Modularity clustering is force-directed layout, Phys. Rev. E 79 (2009) 026102, http://dx.doi.org/10.1103/PhysRevE.79.026102, URL https://link.aps.org/doi/10.1103/PhysRevE.79.026102.

[50] Marcell Tamás Kurbucz, A joint dataset of official COVID-19 reports and the governance, trade and competitiveness indicators of world bank group platforms, 2020, URL https://data.mendeley.com/datasets/hzdnxph8vg/6.