



## Research article

## pyDRMetrics - A Python toolkit for dimensionality reduction quality assessment

Yinsheng Zhang<sup>a,c,\*</sup>, Qian Shang<sup>b,c</sup>, Guoming Zhang<sup>d,e,\*\*</sup><sup>a</sup> School of Management and E-Business, Zhejiang Gongshang University, Hangzhou 310018, China<sup>b</sup> School of Management, Hangzhou Dianzi University, Hangzhou 310018, China<sup>c</sup> School of Information Sciences, University of Illinois at Urbana Champaign, Champaign, IL 61820-6211, USA<sup>d</sup> Pediatric Retinal Surgery Department, Shenzhen Eye Hospital, Shenzhen 518040, China<sup>e</sup> Shenzhen Key Ophthalmic Laboratory, The Second Affiliated Hospital of Jinan University, Shenzhen 518040, China

## ARTICLE INFO

## Keywords:

Dimensionality reduction

Reconstruction error

Distance matrix

Co-ranking matrix

Co-k-nearest neighbor

## ABSTRACT

High-dimensional data are pervasive in this bigdata era. To avoid the curse of the dimensionality problem, various dimensionality reduction (DR) algorithms have been proposed. To facilitate systematic DR quality comparison and assessment, this paper reviews related metrics and develops an open-source Python package pyDRMetrics. Supported metrics include reconstruction error, distance matrix, residual variance, ranking matrix, co-ranking matrix, trustworthiness, continuity, co-k-nearest neighbor size, LCMC (local continuity meta criterion), and rank-based local/global properties. pyDRMetrics provides a native Python class and a web-oriented API. A case study of mass spectra is conducted to demonstrate the package functions. A web GUI wrapper is also published to support user-friendly B/S applications.

## 1. Introduction

High dimensionality is a pervasive phenomenon in various domains. Below are some examples of very high-dimensional data. (1) In healthcare, the complete patient profile can have thousands of features, including basic demographic data, laboratory test results, genetic background, medical imaging results, allergy, symptoms, past diseases, etc. (2) Financial data can also be high dimensional. If we want to analyze the stock market, thousands of stocks listed on the open market are available. (3) In computer vision, a digital image is treated as a vector of pixels. A  $224 \times 224$  RGB color image can be flattened as a 150,528 ( $224 \times 224 \times 3$ ) long vector. (4) In geoscience, geographical data can have hundreds of features, such as longitude, latitude, altitude, temperature, air pressure, wind, rainfall, humidity, season, sunlight, diurnal amplitude, soil properties, remote sensing, and other spatio-temporal data. (5) In life science, microarray gene expression data is also typical high-dimensional data. (6) In text mining applications, an article can be converted to a word-frequency vector (e.g., bag-of-words or N-Gram) containing thousands of terms. (7) In biochemical research, spectroscopic profiling data also have high dimensionality. For example, a typical Raman spectrum has

several thousand dimensions (wavenumbers), and a mass spectrum has tens of thousands of dimensions (m/z).

These high-dimensional data bring unique challenges to researchers. According to the theory of the “curse of dimensionality”, as the number of dimensions increases, the amount of data needed to generalize a machine learning model grows exponentially. It also results in more storage space and longer model training time. To address this issue, various dimensionality reduction (DR) algorithms [1] have been proposed over the years. Figure 1 shows a taxonomy view of the commonly used DR algorithms.

The benefits of DR include the following. (1) Reduce the computational cost of machine learning and inference. (2) Reduce the risk of overfitting. (3) Data visualization. After DR to three or fewer dimensions, we can use the scatter plot to visualize samples. (4) Better data interpretation. Fewer features after DR may better reveal the data distribution structure. (5) Improve SNR (signal-noise ratio). Some algorithms, such as PCA, have the effect of keeping only the high-variance components and filtering out the trivial “noisy” components.

Among the many DR algorithms, how to select the most appropriate one for the current dataset is a challenging task. This paper reviews DR quality metrics and implements an open-sourced Python toolkit.

\* Corresponding author.

\*\* Corresponding author.

E-mail addresses: [zhangys@illinois.edu](mailto:zhangys@illinois.edu) (Y. Zhang), [13823509060@163.com](mailto:13823509060@163.com) (G. Zhang).

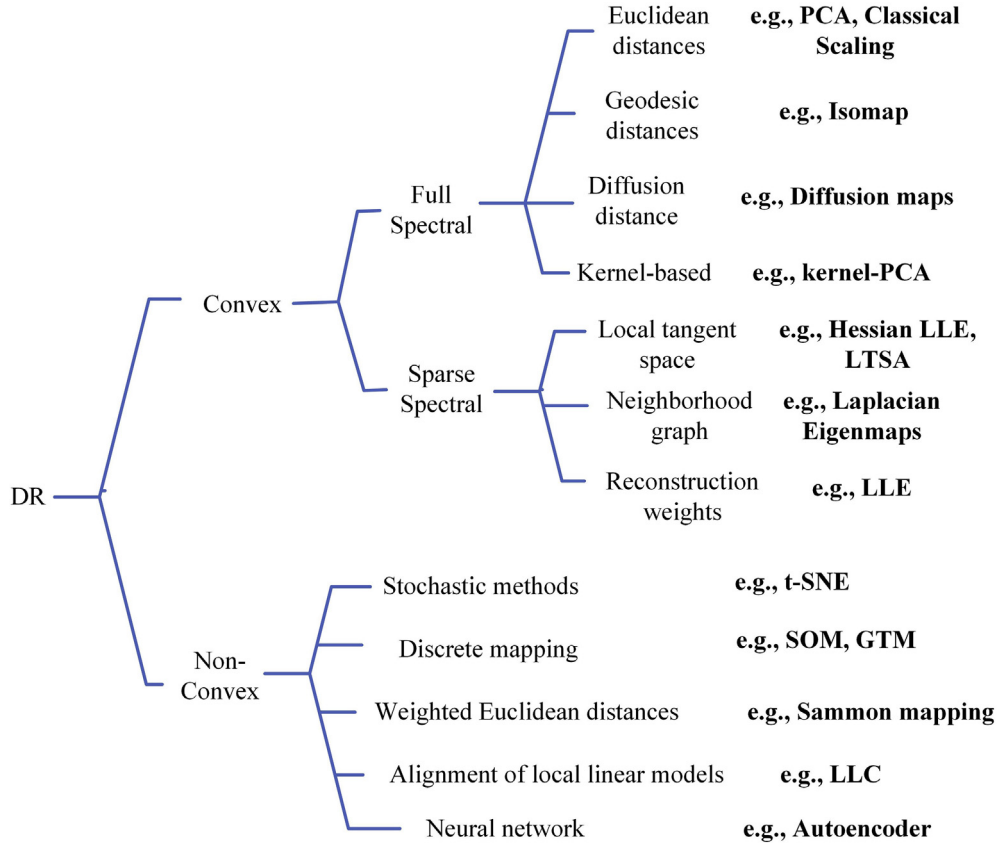


Figure 1. Taxonomy of dimensionality reduction algorithms. Image adapted from Van Der Maaten [1].

Table 1. Symbols and terms.

Symbol/Term	Explanation
DR	Dimensionality Reduction
$X$	The data before DR. An $m$ -by- $n$ matrix.
$x_i$	The $i$ -th sample. The $i$ -th row vector of $X$ .
$Z$	Data after DR. An $m$ -by- $k$ matrix.
$X_r$	Reconstructed data from $Z$ . An $m$ -by- $n$ matrix.
$m$	The number of samples/data points.
$n$	Dimension number before DR. The Dimensionality of the original space. Column number of $X$ .
$k$	In the context of DR, $k$ is the dimension number after DR. The Dimensionality of the DR's latent space. Column number of $Z$ . In the context of the rank-based metrics, such as $Q_{NN}(k)$ , $k$ means the rank or the $k$ -th nearest neighbor.
MSE	Mean Squared Error
rMSE	relative Mean Squared Error
$\  \cdot \ _F$	The Frobenius norm of a matrix
$\# \{ \}$	The cardinality of a set. The number of elements in a set.
$e$	An $m$ -dimensional column vector of ones.
$D$	The distance matrix. An $m$ -by- $m$ matrix
$r$	Pearson correlation coefficient
$V_r$	Residual variance
$R$	The ranking matrix. An $m$ -by- $m$ matrix.
$Q$	The co-ranking matrix. An $m$ -by- $m$ matrix.
$T(k)$	Trustworthiness
$C(k)$	Continuity
$Q_{NN}(k)$	co- $k$ -nearest neighbor size
AUC	Area Under Curve
LCMC	Local Continuity Meta Criterion
$k_{max}$	The maximum cutoff point of LCMC
$Q_{local}$	Local property metric
$Q_{global}$	Global property metric

## 2. Symbols and glossary

Table 1 lists all the math symbols and terms used in this study.

## 3. Metrics

Due to different design principles, each DR algorithm tries to minimize a different objective function. For example, PCA (principal component analysis) maximizes the explained variance, while t-SNE (t-distributed stochastic neighbor embedding) [2] minimizes the KL (Kullback-Leibler)-divergence between the original and the low-dimensional data. Therefore, it is unreasonable to use algorithm-specific metrics for DR quality comparison.

The basic principles for choosing DR metrics are as follows. (1) The metrics should be generic and algorithm-independent, (2) The metrics better be diversified, i.e., each evaluates the DR from a different perspective.

In the following manuscript, we will introduce a set of candidate metrics suitable for DR quality assessment. The related concepts and theories for each metric are also provided.

### 3.1. Reconstruction error

From the perspective of signal processing, DR is a process of “compression” or “encoding”. An ideal DR has a high compression ratio with low information loss. The reconstruction error can be used to measure how much information is lost due to DR.

Suppose  $X$  is the original data matrix.  $X_r$  is the reconstructed data matrix. The row number  $m$  is the total sample count. The column number  $n$  is the number of features or dimensions. The reconstruction error of DR can be defined by MSE (mean squared error).

$$MSE = \frac{\|X - X_r\|_F^2}{mn}, \quad (1)$$

Because MSE is an absolute error depending on a specific dataset, when comparing DR on different datasets, relative reconstruction error (rMSE) is preferred. rMSE is defined as:

$$rMSE = \frac{\|X - X_r\|_F^2}{\|X\|_F^2}, \quad (2)$$

In the above equations,  $\|\cdot\|_F$  is the Frobenius norm of a matrix.

One restriction to use the reconstruction error metric is that the DR algorithm must be “bi-directional” or “reversible”. In other words, the algorithm should provide both forward and backward mappings. However, not all DR algorithms meet this requirement. Typical examples include t-SNE [2] and random projection (RP) [3]. In the *sci-kit-learn* library, only those classes with an “*inverse\_transform*” method are reversible, such as PCA and NMF (non-negative matrix factorization) [4].

### 3.2. Distance matrix

From the perspective of the pair-wise property, the distance matrix is another metric for evaluating DR. For an  $m$ -by- $n$  data matrix  $X$ , the distance matrix  $D$  is an  $m$ -by- $m$  diagonal matrix. Each element  $D_{ij}$  is the distance/dissimilarity between the  $i$ -th and  $j$ -th samples.

$$D_{ij} = \|x_i - x_j\|, \quad (3)$$

In the above equation,  $\|x_i - x_j\|$  can be the Euclidean distance or the RBF (radial basis function) similarity. Essentially, the distance matrix can be seen as a kernel function that maps  $X$  in  $n$ -dimensional space to  $D$  in  $m$ -dimensional space.

A good DR algorithm should preserve the pair-wise property. In other words, the distance matrices before and after DR (denoted as  $D$  and  $D'$ ) should be similar. Luckily, this can be guaranteed by the Johnson-

Lindenstrauss lemma [5], i.e., an  $m$  point set in Euclidean space can be embedded in  $O(\log n/\epsilon^2)$  dimensions without distorting the pair-wise distances by a factor of no more than  $(1 \pm \epsilon)$ , for any  $0 < \epsilon < 1$ .

To compare  $D$  and  $D'$ , we provide two methods. (1) Visualize the matrices as heatmaps. The two matrices should display a similar visual pattern. (2) Calculate their residual variance:  $Vr = 1 - r^2(D, D')$ .  $r$  can be either Pearson or Spearman correlation coefficient.

### 3.3. Ranking matrix

The ranking matrix  $R$  is derived from the distance matrix  $D$ . It uses ranks/orders instead of distances to measure pair-wise properties. Each element  $R_{ij}$  in  $R$  is defined as:

$$R_{ij} = \#\{k : D_{ik} < D_{ij} \text{ or } (D_{ik} = D_{ij} \text{ and } k < j)\}, \quad (4)$$

The symbol  $\#$  means the cardinality of a set or the number of elements in the set.

$R_{ij}$  means  $x_j$  is the  $R_{ij}$ -th nearest point of  $x_i$ .

After DR, for the distance matrix  $D'$ , we can get its ranking matrix  $R'$  in the same way:

$$R'_{ij} = \#\{k : D'_{ik} < D'_{ij} \text{ or } (D'_{ik} = D'_{ij} \text{ and } k < j)\}, \quad (5)$$

The distance matrix  $D$  and the ranking matrix  $R$  are closely related. (1) Distance and ranking are both non-parametric statistics, i.e., they don't need to make distribution assumptions. (2) From the perspective of machine learning, both  $D$  and  $R$  can be seen as a kind of kernel operation. The widely used Gaussian kernel measures pair-wise “similarities”, while  $D$  and  $R$  measure pair-wise “distance” and “ranking”.

The ranking matrix  $R$  has the following properties:

$$0 \leq R_{ij} \leq m - 1, \quad (6)$$

$$\sum_{0 \leq i < j < m} (R_{ij}) = e^T R e = m \times C_m^2 = m^2(m-1)/2, \quad (7)$$

$\sum_{0 \leq i < j < m} (R_{ij})$  is the summation of all elements in the matrix  $R$ .  $e$  is an  $m$ -dimensional column vector of ones.

### 3.4. Co-ranking matrix

The co-ranking matrix [6]  $Q$  is derived from two ranking matrices, e.g.,  $R$  and  $R'$ . Its element  $Q_{kl}$  is defined as:

$$Q_{kl} = \#\{(i, j) : R_{ij} = k \text{ and } R'_{ij} = l\}, \quad (8)$$

The co-ranking matrix  $Q$  provides valuable information about a DR transform.  $Q_{kl}$  counts how many samples of rank  $k$  become rank  $l$ . The off-diagonal elements of  $Q$  correspond to errors due to DR. The co-ranking matrix  $Q$  has the following properties:

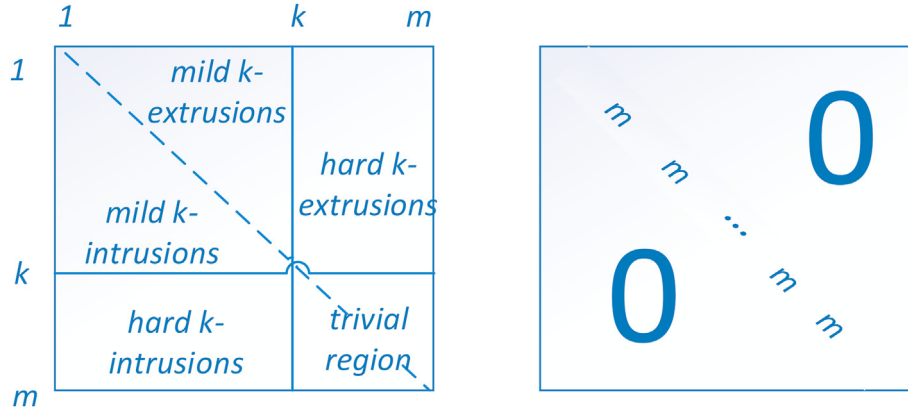
$$0 \leq Q_{ij} \leq m, \quad (9)$$

$$\forall i \in [0, m), \sum_{0 \leq j < m} (Q_{ij}) = m, \quad (10)$$

$$\forall j \in [0, m), \sum_{0 \leq i < m} (Q_{ij}) = m, \quad (11)$$

$$\sum_{0 \leq i < j < m} (Q_{ij}) = e^T Q e = m^2, \quad (12)$$

An ideal co-ranking matrix is diagonal, i.e., all diagonal elements equal to  $m$ , while all non-diagonal elements are zeros. This means no samples have their ranks changed after DR (Figure 2b). For elements in the lower triangle, the row index  $k$  is smaller than the column index  $l$ . If non-zero elements are observed in this region, it means the DR pulls far-away points close. These points are called “intrusions”. Likewise, if non-



**Figure 2.** (a) An illustration of the co-ranking matrix. The intrusions are located in the low triangle region, which means DR pulls far-away points closer. The extrusions are in the upper triangle, which means DR pushes near points apart. The right-bottom area is a trivial region for large ranks, which are much less important than local relations. (b) An ideal co-ranking matrix is diagonal. All diagonal elements are  $m$ , while all non-diagonal elements are zeros.

zero elements are found in the upper triangle, it means the DR pushes near points away. Such points are “extrusions”. In most cases, the preservation of local relationships is much more important than distant ones. Therefore, the rank errors for large ranks are deemed trivial. As shown in Figure 2a, the right-bottom area is the “trivial region”.

### 3.5. Trustworthiness and continuity

Trustworthiness and continuity [7] are two rank-based metrics. Trustworthiness is related to the error induced by hard intrusions. It equals the inverse of weighted hard- $k$ -intrusions:

$$T(k) = 1 - \frac{1}{mk(m-k)} \sum_{(i,j) \in \mathbf{L}\mathbf{L}_k} Q_{ij}(i-k) = 1 - \frac{1}{mk(m-k)} \sum_{i=k}^m \sum_{j=1}^k Q_{ij}(i-k), \quad (13)$$

LL means the lower-left block, i.e., the “hard  $k$ -intrusion” region in Figure 2a.  $(i-k)$  in the equation is the rank error, used as a penalty weight.  $i$  is the actual rank in the original data. If  $i$  is far from the  $k$ -nearest neighbor after DR, the penalty will be large, which means a “false” close pair is generated and the DR is “untrustworthy”. Likewise, continuity is defined as:

$$C(k) = 1 - \frac{1}{mk(m-k)} \sum_{(i,j) \in \mathbf{U}\mathbf{R}_k} Q_{ij}(j-k) = 1 - \frac{1}{mk(m-k)} \sum_{i=1}^k \sum_{j=k}^m Q_{ij}(j-k), \quad (14)$$

UR means the upper-right block, i.e., the “hard  $k$ -extrusion” region in Figure 2a.  $(j-k)$  is the rank error, used as a penalty weight.  $j$  is the actual rank after DR. If  $j$  becomes far away from the  $k$ -nearest neighbor before DR, the penalty will be large, which means these points become “discontinued” or “disconnected” after DR.

Both trustworthiness and continuity range from 0 to 1. In extremely bad cases, when the pair-wise ranks are reversed, there will be lots of hard intrusions and extrusions, and  $T(k)$  and  $C(k)$  will be near 0.

The AUCs (area under the curve) of  $T(k)$  and  $C(k)$  are also available, but they are less used than the AUC of  $Q_{NN}(k)$ , which will be introduced in the next section.

### 3.6. Co- $k$ -nearest neighbor size

Similar to trustworthiness and continuity, the “co- $k$ -nearest neighbor size” [6] is also derived from the co-ranking matrix, denoted as  $Q_{NN}(k)$ .

$$Q_{NN}(k) = \frac{1}{km} \sum_{i=1}^k \sum_{j=1}^k Q_{ij}, \quad (15)$$

$Q_{NN}(k)$  is the sum of all elements in the upper-left  $k$ -by- $k$  sub-matrix of  $Q$ . It counts how many points are in both  $k$ -nearest neighbors before and after the DR. The purpose of dividing  $(km)$  is to normalize the value to range  $[0,1]$ . The symbol “ $Q_{NN}$ ” (the “ $NN$ ” means nearest-neighbor) in this paper is the same as “ $Q_{NX}$ ” (the “ $NX$ ” stands for intrusions and extrusions) in [6]. In [6],  $Q_{NX}(k) = \frac{1}{km} \sum_{(i,j) \in \mathbf{D}_k} Q_{ij} + \frac{1}{km} \sum_{(i,j) \in \mathbf{U}\mathbf{T}_k} Q_{ij} + \frac{1}{km} \sum_{(i,j) \in \mathbf{L}\mathbf{T}_k} Q_{ij}$ .  $\mathbf{D}$  is the main diagonal.  $\mathbf{U}\mathbf{T}$  and  $\mathbf{L}\mathbf{T}$  are the upper and lower triangles, corresponding to the “mild  $k$ -extrusion” and “mild  $k$ -intrusion” regions in Figure 2a. According to their equations, “ $Q_{NN}$ ” and “ $Q_{NX}$ ” are the same.

$Q_{NN}(k)$  is an array, other than a single number. We can use AUC (area under the curve) as a single metric irrelevant to  $k$ .

$$AUC = \frac{1}{m} \sum_{k=1}^m Q_{NN}(k), \quad (16)$$

### 3.7. Local continuity meta criterion

As an improvement on  $Q_{NN}$ , the adjusted LCMC (Local Continuity Meta Criterion) [8] for measuring how well the  $k$ -nearest neighbors agree is defined as follows:

$$LCMC(k) = Q_{NN}(k) - \frac{k}{m-1}, \quad (17)$$

LCMC is  $Q_{NN}$  with the baseline  $\left(\frac{k}{m-1}\right)$  removal. LCMC favors locality more than  $Q_{NN}$ . A large  $k$  tends to have a bigger penalty. LCMC has a well-defined maximum point  $k_{max}$ :

$$k_{max} = \arg\max_k LCMC(k), \quad (18)$$

### 3.8. The local and global properties

Given the above  $k_{max}$ , two metrics are designed [9]:

$$Q_{local} = \frac{1}{k_{max}} \sum_{k=1}^{k_{max}} Q_{NN}(k), \quad (19)$$

$$Q_{global} = \frac{1}{m - k_{max}} \sum_{k=k_{max}}^{m-1} Q_{NN}(k), \quad (20)$$

$Q_{local}$  and  $Q_{global}$  correspond to local and global properties respectively. In most cases, the locality is usually preferred over the global property, as people are more sensitive to nearer neighbors. For example, after DR, for a specific data point, its nearest (1<sup>st</sup>) neighbor that becomes the third (3<sup>rd</sup>) is usually considered much more important than a 101<sup>st</sup>

**Table 2.** DR Quality Metrics and their Explanations.

Metric	Math Equation	Explanation	Comment
Reconstruction Error*	$MSE = \frac{\ X - X_r\ _F^2}{mn}$	Reconstruction error, measured by the MSE between $X$ and $X_r$	Requires the reconstructed data $X_r$ , i.e., the DR algorithm must be reversible or have an inverse transform.
Relative Reconstruction Error*	$rMSE = \frac{\ X - X_r\ _F^2}{\ X\ _F^2}$	Relative reconstruction error, measured by the relative MSE between $X$ and $X_r$	
Distance matrix	$D_{ij} = \ x_i - x_j\ $	Measures the pair-wise distance property. The distance matrices before and after DR should be similar.	The distance can be Euclidean, or the RBF (radial basis function) similarity
Residual variance*	$Vr = 1 - r^2(D, D')$	Residual variance of the distance matrices before and after DR.	$r$ is the Pearson or Spearman correlation coefficient
Ranking matrix	$R_{ij} = \#\{k : D_{ik} < D_{ij} \text{ or } (D_{ik} = D_{ij} \text{ and } k < j)\}$	Contains the ranking information. The ranking matrices before and after DR should be similar.	$R_{ij}$ means $x_i$ is the $R_{ij}$ -th nearest neighbor of $x_j$ .
Co-ranking matrix	$Q_{kl} = \#\{(i, j) : R_{ij} = k \text{ and } R'_{ij} = l\}$	Measures how sample ranks change after the DR.	$Q_{kl}$ counts how many samples of rank $k$ become rank $l$ . An ideal $Q$ is diagonal.
Trustworthiness	$T(k) = 1 - \frac{1}{mk(m-k)} \sum_{i=k}^m \sum_{j=1}^k Q_{ij}(i-k)$	Measures error of hard intrusions. $(i-k)$ is the weighted rank error	Ranges from 0 to 1.
Continuity	$C(k) = 1 - \frac{1}{mk(m-k)} \sum_{i=1}^k \sum_{j=k}^m Q_{ij}(j-k)$	Measures error of hard extrusions. $(j-k)$ is the weighted rank error	Ranges from 0 to 1.
Co- $k$ -nearest neighbor size	$Q_{NN}(k) = \frac{1}{km} \sum_{i=1}^k \sum_{j=1}^k Q_{ij}$	Count how many points are in both $k$ -nearest neighbors before and after the DR.	Divide by $(km)$ to normalize the value to range $[0, 1]$
The area under the curve*	$AUC = \frac{1}{m} \sum_{k=1}^m Q_{NN}(k)$	The area under the $Q_{NN}(k)$ curve.	Ranges from 0.5 to 1.
Local Continuity Meta Criterion	$LCMC(k) = Q_{NN}(k) - \frac{k}{m-1}$	LCMC is $Q_{NN}$ with baseline removal.	LCMC favors locality more than $Q_{NN}$ . A large $k$ has a bigger penalty than a smaller one.
Local property metric*	$Q_{local} = \frac{1}{k_{max}} \sum_{k=1}^{k_{max}} Q_{NN}(k)$	Measures local property (small $ks$ ). Usually favored over $Q_{global}$ .	$k_{max}$ is the maximum cutoff point of LCMC: $k_{max} = \arg\max_k LCMC(k)$
Global property metric*	$Q_{global} = \frac{1}{m - k_{max}} \sum_{k=k_{max}}^{m-1} Q_{NN}(k)$	Measures global property (big $ks$ ). Use when $Q_{local}$ is a tie.	

\* These items are single numeric metrics. Others are arrays or matrices.

neighbor that becomes the 103<sup>rd</sup>. The original authors also recommend using  $Q_{local}$  over  $Q_{global}$ .  $Q_{global}$  is often used when  $Q_{local}$  is at a tie.

### 3.9. The final set of metrics

After reviewing the above metrics, the final set of DR quality evaluation metrics is shown in Table 2. As each metric measures the DR quality from a different perspective, we recommend users to use them to form an integrated view. In the following manuscript, we will also study the inter-relationship or consistency between these metrics, i.e., to see whether they increase/decrease consistently with different DR levels.

Under certain circumstances, users may prefer certain metrics over the others, depending on various factors, such as the data generation process, pre-processing procedures, statistical properties of the dataset, and concrete application settings. For example, if the dataset manifests a “swiss-roll” style, local property metrics are usually preferable over global properties. For applications where data quality is important, the reconstruction error metrics may be preferred over others.

## 4. Implementation

Until now, ready-to-use toolkits for DR quality evaluation are still scarce. One available toolkit is the “*dimRed*” package in the *R* platform [10], which provides popular DR algorithm implementations and co-ranking based metrics. For Python, there is a “coranking” package (<https://coranking.readthedocs.io/>). “coranking” is light-weighted and only provides three metrics, i.e., trustworthiness, continuity, and LCMC.

This paper develops a “pyDRMetrics” package for the Python platform. This package implements the comprehensive set of DR quality metrics. Table 3 lists the core API definition.

## 5. Case study and web application

### 5.1. Background

In recent years, our team has been using spectroscopic profiling technology (e.g., Raman spectroscopy and mass spectroscopy) to identify various biomarkers in biomedical applications [11, 12, 13, 14]. The biomarkers are molecules generated by normal biological processes or the body's responses to pathogens. The detection of biomarkers is essential for the diagnosis and evaluation of relevant diseases. When combined with machine learning and statistical models, spectroscopic profiling provides a promising data-driven approach for biochemical analysis.

However, one major problem faced by the spectroscopic data analysis is high dimensionality. For example, a typical Raman spectrum has several thousand dimensions (*wavenumbers*), and a time-of-flight mass spectrum has tens of thousands of dimensions (*m/z*).

### 5.2. Dataset

The dataset used in this case study is “OvarianCancer-NCI-PBSII-061902” [14], which contains SELDI-TOF-MS (surface-enhanced laser desorption and ionization time-of-flight mass spectroscopy) data of 253 blood serum samples. SELDI is a subtype of MALDI (matrix-assisted laser desorption/ionization). It preselects the serum proteins by binding them to a treated metal surface. Figure 3 illustrates the design of TOF-MS. In the dataset, each of the 15154 features represents the quantity (abundance) of a particular chemical particle of different *m/z* (mass-to-charge ratio). Figure 4 shows the averaged waveform of the collected mass spectra.

**Table 3.** API definition.

Class DRMetrics(builtins.object)	Define a set of dimensionality reduction metrics.
Properties / Fields	
X	Data before DR. m-by-n matrix. m is the sample size. n is the dimension/feature number.
Z	Data after DR. m-by-k matrix. Typically, $k \ll n$
Xr	Reconstructed Data. m-by-n matrix. Optional parameter. If a DR algorithm has no inverse transform. Pass None.
D	Distance matrix of X
Dz	Distance matrix of Z
Vr	The residual variance between D and Dz. The default version uses Pearson's r to calculate the residual variance. Use Vrs for the Spearman's r version.
mse	Reconstruction error. MSE of X and Xr
rmse	Relative reconstruction error. Relative MSE of X and Xr
R	Ranking matrix of X
Rz	Ranking matrix of Z
Q	Co-ranking matrix between R and Rz
T	Trustworthiness. An array. There is also a single-valued AUC_T that measures the area under the T(trustworthiness) curve.
C	Continuity. An array. There is also a single-valued AUC_C that measures the area under the C(continuity) curve.
QNN	Co-k-nearest neighbor size. An array.
AUC	The area under the QNN curve.
LCMC	Local Continuity Meta Criterion. An array.
Qlocal	Local property metric
Qglobal	Global property metric
Member Methods	
__init__(self, X, Z, Xr=None)	Constructor. X is the original data. Z is the data after DR. Xr is the reconstructed data.
test(cls, csv, k = 3, dr = 'PCA')	A constructor overload that facilitates testing common DR algorithms. csv is the data file. dr is used to specify the algorithm, e.g. "PCA", "NMF", "RP", "TSNE", etc.
plot_coranking_matrix(self)	Visualize the co-ranking matrix between R and Rz as heatmaps.
plot_distance_matrix(self)	Visualize the distance matrices before and after DR as heatmaps, i.e., D and Dz.
plot_ranking_matrix(self)	Visualize the ranking matrices before and after DR as heatmaps, i.e., R and Rz.
visualize_reconstruction(self)	Plot the original data and the reconstruction data side by side. Show 3 random samples.
report(self)	Print out a summary report, with inline plots.
get_json(self)	Generate a JSON-format dictionary object containing all DR quality metrics. Only numeric metrics are returned.
get_html(self)	Generate an HTML segment that can be embedded in web pages. Plotted images are embedded as base64 strings to avoid referencing external files.
Sample Code	
<pre> from pyDRMetrics import * from sklearn.decomposition import PCA %matplotlib inline  K = 3 pca = PCA(n_components = K) # keep the first K components pca.fit(X) Z = pca.transform(X) plotComponents3D(Z, y, labels) Xr = pca.inverse_transform(Z) print('explained variance ratio:', pca.explained_variance_ratio_[0:5])  drm = DRMetrics(X, Z, Xr) # construct a DRMetrics object print("Qlocal = ", drm.Qlocal) # get Qlocal drm.report() # print out the summary  from IPython.display import display, HTML display(HTML(drm.get_html())) # render the returned html string. You can embed this HTML segment in a web page. </pre>	

### 5.3. Result

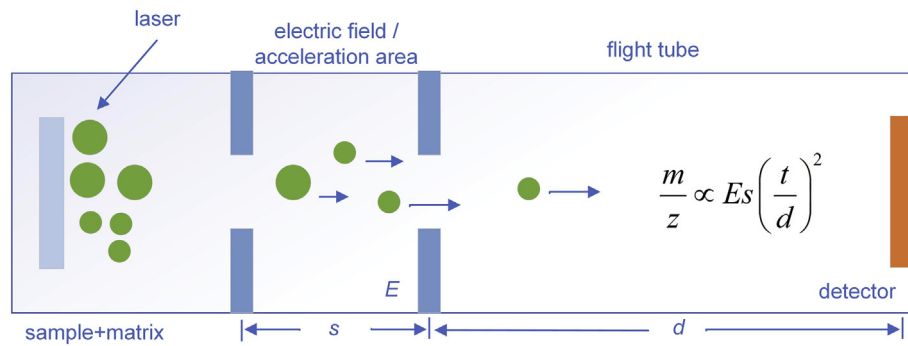
#### 1) DR using PCA

In this case study, PCA is used to reduce the dimensions of the dataset. Figure 5 shows the explained variances of different ks (number of

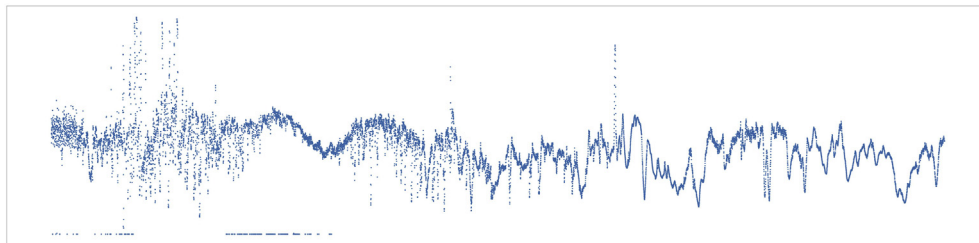
components). With the "elbow" method, we choose  $k = 5$ . After  $k = 5$ , the explained variances of each PC are less than 3%. The first 5 PCs hold 78.7% of the total original information.

Therefore, the original data X is a 253-by-15154 matrix, and after PCA, the data Z becomes a 253-by-5 matrix. Xr is the reconstructed

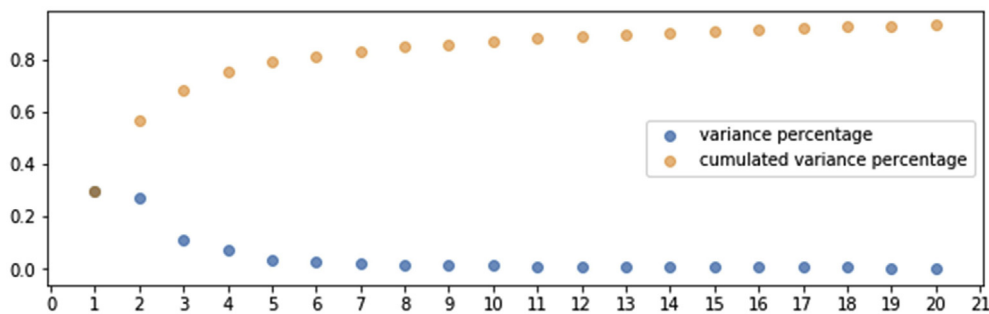




**Figure 3.** Scheme of the TOF MS (time-of-flight mass spectroscopy).  $m/z$  = mass-to-charge ratio.  $t$  = drift time. The detector collects the arrival time of charged particles (e.g., ions), and we can then calculate the  $m/z$  (mass-to-charge ratio) by the drift time. The signal intensity received by the detector at a specific drift time corresponds to the abundance/quantity of a specific particle.



**Figure 4.** The averaged waveform of the SELDI-TOF-MS dataset.



**Figure 5.** The explained variances of different numbers of components.

matrix from  $Z$ . These three matrices are used to initialize/construct the *DRMetrics* object in the *pyDRMetrics* package.

The DR metrics returned by the object are listed in Table 4. From the results, we can see that PCA achieves a decent quality by keeping only 5 features from the original 15154 ones (compression ratio is  $5/15154 = 0.00033$ ). The relative reconstruction error is 3.55%. The distance matrices and ranking matrices before and after PCA are almost the same. Residual variance is 0.0286 (ideal value is 0). The trustworthiness and continuity are very close to the ideal case, i.e., horizontal lines at 1. The co-ranking matrix  $Q$  is concentrated alongside the diagonal line (an ideal  $Q$  is a diagonal matrix). AUC of  $Q_{NN}$  is 0.932 (ideal value is 1).  $Q_{local} = 0.717$  (ideal value is 1).  $Q_{global} = 0.954$  (ideal value is 1).

## 2) Dynamic properties

To study the dynamic properties of DR metrics, PCA at different  $k$  values (from 1 to 20) is conducted. As shown in Figure 6, when  $k$  increases, (1) reconstruction error decreases as more variance/information is preserved. (2) The residual variance between  $D$  and  $D'$  decreases, as information loss is reduced. (3) AUC increases. (4, 5)  $Q_{local}$  and  $Q_{global}$

increase because both the local and global information is better preserved.

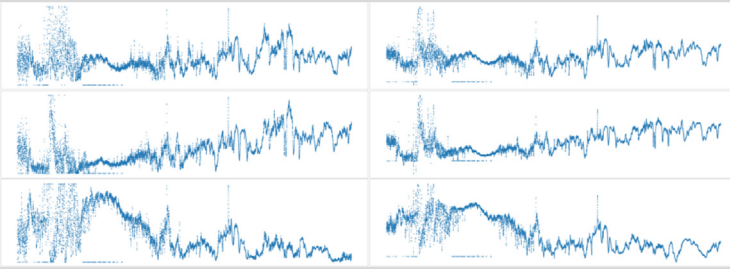
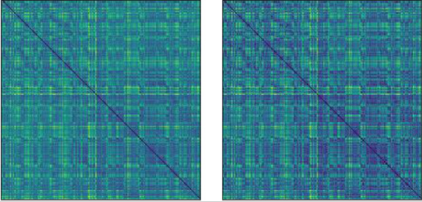
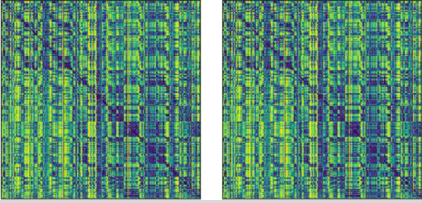

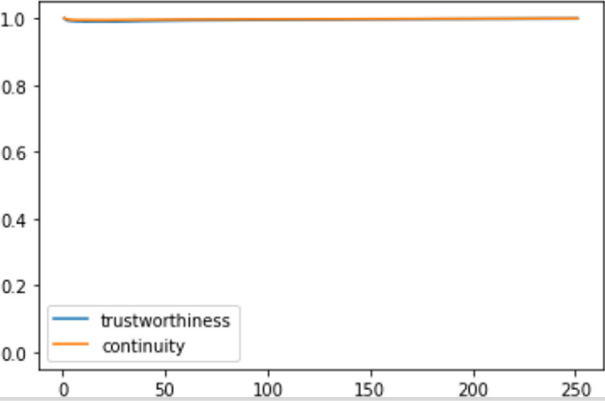
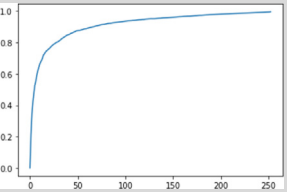
These curves also show a high degree of consistency with the explained variance curve (Figure 5). They all have a “turning point” or “elbow” near  $k = 5$ .

We also investigate how certain metrics change in two extreme cases. 1)  $k = n$ , or a “perfect” DR. In this case, all the  $n$  dimensions are kept and there is no information loss. As shown in Table 5, it has a perfect diagonal co-ranking matrix. The  $Q_{NN}$  curve becomes a horizontal line located at 1.0, and the *LCMC* curve is a straight line from (1, 1) to ( $m$ , 0). 2)  $k = 0$ . In this case, all the original points are collapsed into one single point. A single point is seen as a zero-dimensional space, as it has zero degrees of freedom or movement direction. The  $Q_{NN}$  curve becomes a line from (1, 0) to ( $m$ , 1). The *LCMC* curve is a horizontal line located at 0.

## 5.4. Web application

To better serve users, we also developed a user-friendly web GUI. The GUI is a wrapper of *pyDRMetrics*. It supports two modes. 1) The basic mode supports several built-in algorithms, including PCA, NMF, MDS, t-

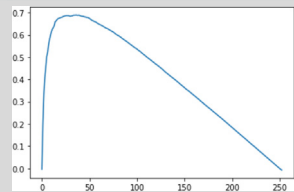
**Table 4.** DR Quality Metrics of PCA ( $k = 5$ ) returned by pyDRMetrics

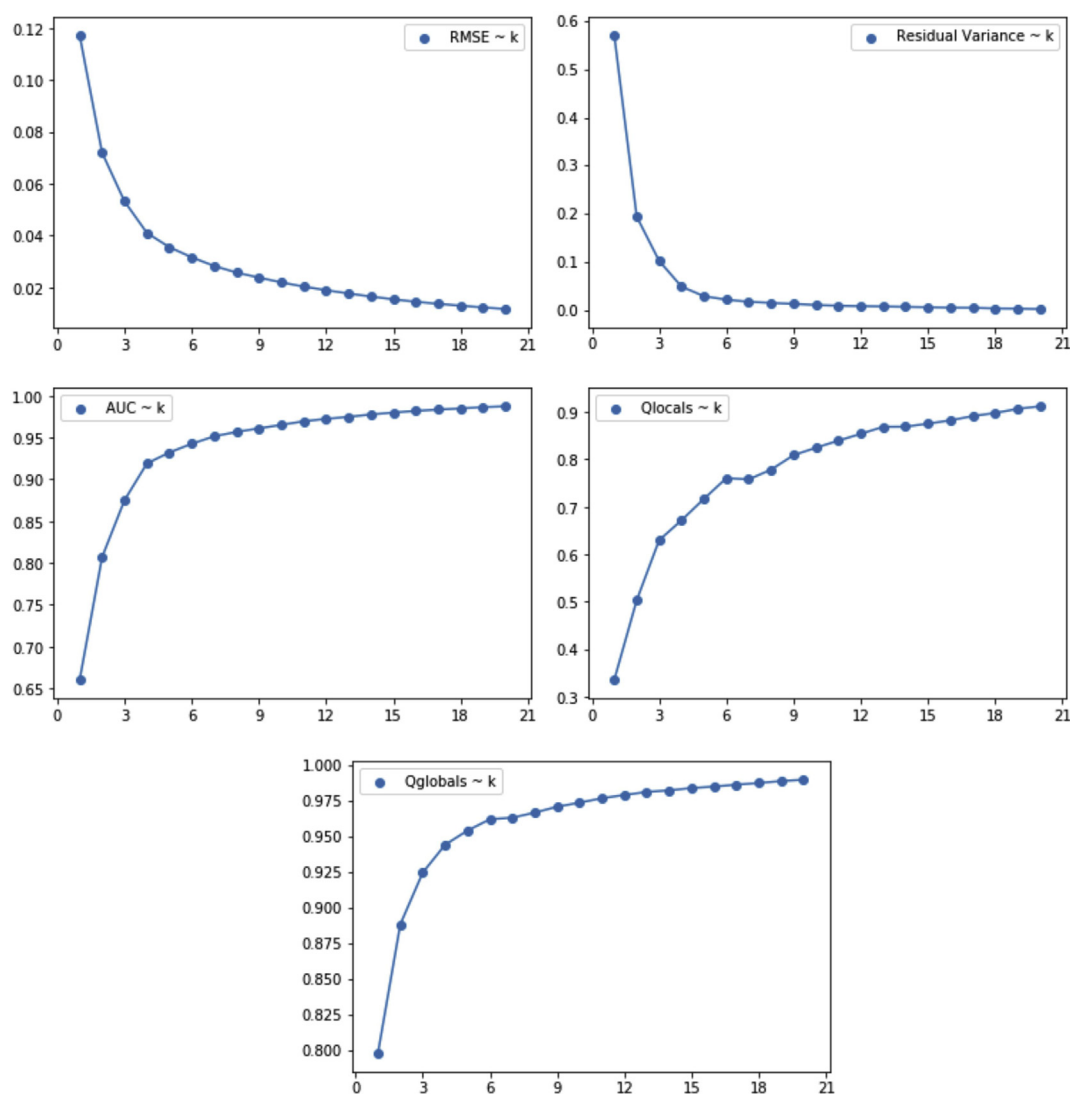
Metric	DRMetrics Field/Method	Returned value
Reconstruction Error	obj.mse	0.00683
Relative Reconstruction Error	obj.rmse	0.0355
Show random samples before and after DR, side by side	obj.visualize_reconstruction()	
Distance matrices before and after DR	obj.D, obj.Dz, obj.plot_distance_matrix()	
Residual variance	obj.Vr, obj.Vrs	0.029, 0.028
Ranking matrices before and after DR	obj.R, obj.Rz, obj.plot_ranking_matrix()	
Co-ranking matrix	obj.Q, obj.plot_coranking_matrix()	
Trustworthiness, Continuity	obj.T, obj.C	
	obj.AUC_T, obj.AUC_C	0.996, 0.998
Co-k-nearest neighbor size	obj.QNN	
The area under the curve of $Q_{NN}$	obj.AUC	0.932

(continued on next page)



Table 4 (continued)

Metric	DRMetrics Field/Method	Returned value
Local Continuity Meta Criterion	obj.LCMC	
Maximum cutoff of LCMC	obj.kmax	23
Local property	obj.Qlocal	0.717
Global property	obj.Qglobal	0.954



**Figure 6.** The curves of numeric metrics against different  $k$  values. (1) Relative reconstruction error (RMSE) curve. (2) The curve of the residual variance between distance matrices. (3) The curve of  $Q_{NN}$  AUC (area under the curve). (4)  $Q_{local}$  curve. (5)  $Q_{global}$  curve.

SNE, etc. The GUI can be accessed at <http://spacs.brahma.pub/research/DR>. 2) The other mode is the extended mode for testing non-built-in (e.g., the more recent UMAP [15]) or user-defined algorithms. In this mode,

the DR implementation details are left to the user. Users need to upload three CSV files, corresponding to  $X$ ,  $Z$ , and  $X_r$  (optional) respectively (see Figure 7).

Table 5. DR metrics for two special cases.

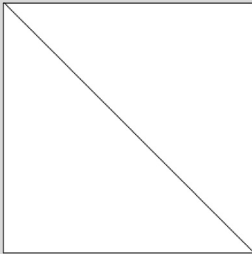
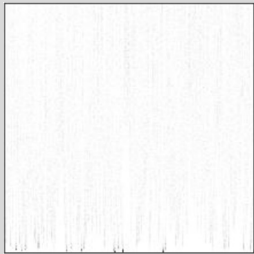
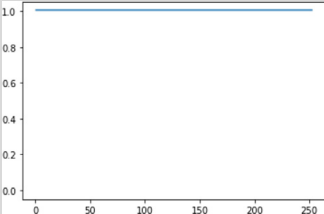
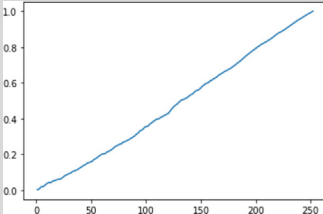
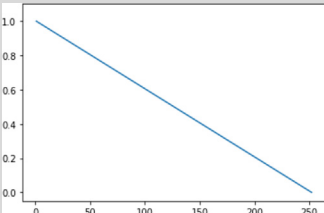
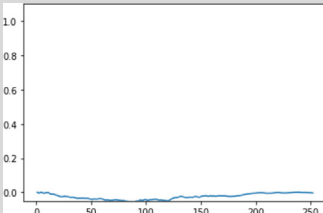
Metric	$k = n$	$k = 0$
Co-ranking matrix (Q)		
Co-k-nearest neighbor size ( $Q_{NN} \sim k$ )		
Local Continuity Meta Criterion (LCMC ~ k)		



Figure 7. The web-based GUI powered by pyDRMetrics. The GUI has a basic mode that supports testing built-in public algorithms and an extended mode for testing user-defined algorithms. The GUI is published at <http://spacs.brahma.pub/research/DR>.

6. Conclusion

Dimensionality reduction is a widely used technique in handling high dimensional data. How to compare the existing DR algorithms and

choose the most appropriate one is a challenging task. To provide a technical solution to this issue, this paper reviews a set of metrics for evaluating DR qualities and develops the pyDRMetrics toolkit. pyDRMetrics is the first Python-oriented package that provides comprehensive

DR quality evaluation. A web application based on pyDRMetrics is also developed for users' convenience. The latest package version supports 6 built-in algorithms, i.e., PCA, NMF, RP, VQ, MDS, and t-SNE. It also allows users to test customized algorithms. Until now, this package has been deployed in the Rapid Spectroscopic Profiling and Bigdata Research Institute (Hangzhou 310018, China) and has been positively received by peer researchers.

## Declarations

### Author contribution statement

YS Zhang: Conceived and designed the experiments; Performed the experiments; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Qian Shang: Performed the experiments; Analyzed and interpreted the data; Wrote the paper.

Guoming Zhang: Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

### Funding statement

This work is supported by the National Natural Science Foundation of China under Grant 61806177, Shenzhen Science and Technology Innovation Commission basic discipline layout project under Grant JCYJ20170817112542555, Shenzhen-Hong Kong Co-financing Project under Grant SGDX20190920110403741, Shenzhen Key Medical Discipline Construction Fund under Grant SZXK038, and China Scholarship Council under Grant 201808330609.

### Data availability statement

Data associated with this study has been deposited at <https://doi.org/10.17632/jbjd5fmggh.1>.

### Declaration of interests statement

The authors declare no conflict of interest.

## Additional information

No additional information is available for this paper.

## References

- [1] L. van der Maaten, E. Postma, J. van den Herik, Dimensionality Reduction: A Comparative Review, (n.d.) 36.
- [2] L. van der Maaten, G. Hinton, Visualizing data using t-SNE, *J. Mach. Learn. Res.* 9 (2008) 2579–2605.
- [3] E. Bingham, H. Mannila, Random projection in dimensionality reduction: applications to image and text data, in: *Proc. Seventh ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. - KDD 01*, ACM Press, San Francisco, California, 2001, pp. 245–250.
- [4] D.D. Lee, H.S. Seung, Learning the parts of objects by non-negative matrix factorization, *Nature* 401 (1999) 788–791.
- [5] S. Dasgupta, A. Gupta, An elementary proof of the Johnson-Lindenstrauss Lemma, (n.d.) 6.
- [6] J.A. Lee, M. Verleysen, Quality assessment of dimensionality reduction: rank-based criteria, *Neurocomputing* 72 (2009) 1431–1443.
- [7] J. Venna, S. Kaski, Local multidimensional scaling, *Neural Network.* 19 (2006) 889–899.
- [8] L. Chen, A. Buja, Local multidimensional scaling for nonlinear dimension reduction, graph drawing, and proximity analysis, *J. Am. Stat. Assoc.* 104 (2009) 209–219.
- [9] J.A. Lee, M. Verleysen, Scale-independent quality criteria for dimensionality reduction, *Pattern Recogn. Lett.* 31 (2010) 2248–2257.
- [10] G. Kraemer, M. Reichstein, M.D. Mahecha, dimRed and coRanking - unifying dimensionality reduction in R, *R J* 10 (2018) 342.
- [11] R. Bakry, M. Rainer, C.W. Huck, G.K. Bonn, Protein profiling for cancer biomarker discovery using matrix-assisted laser desorption/ionization time-of-flight mass spectrometry and infrared imaging: a review, *Anal. Chim. Acta* 690 (2011) 26–34.
- [12] M. Paraskeva, K.M. Ashton, H.F. Stringfellow, N.J. Wood, P.J. Keating, A.W. Rowbottom, P.L. Martin-Hirsch, F.L. Martin, Raman spectroscopic techniques to detect ovarian cancer biomarkers in blood plasma, *Talanta* 189 (2018) 281–288.
- [13] S. Long, Q. Qin, Y. Wang, Y. Yang, Y. Wang, A. Deng, L. Qiao, B. Liu, Nanoporous silica coupled MALDI-TOF MS detection of Bence-Jones proteins in human urine for diagnosis of multiple myeloma, *Talanta* 200 (2019) 288–292.
- [14] E.F.P. Iii, A.M. Ardekani, B.A. Hitt, P.J. Levine, V.A. Fusaro, S.M. Steinberg, G.B. Mills, C. Simone, D.A. Fishman, E.C. Kohn, L.A. Liotta, Use of proteomic patterns in serum to identify ovarian cancer, *Lancet* 359 (2002) 6.
- [15] L. McInnes, J. Healy, J. Melville, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, 2020. ArXiv180203426 Cs Stat. <http://arxiv.org/abs/1802.03426>. (Accessed 2 December 2020).