Clustering Evaluation in High-Dimensional Data

Nenad Tomašev and Miloš Radovanović

Abstract Clustering evaluation plays an important role in unsupervised learning systems, as it is often necessary to automatically quantify the quality of generated cluster configurations. This is especially useful for comparing the performance of different clustering algorithms as well as determining the optimal number of clusters in clustering algorithms that do not estimate it internally. Many clustering quality indexes have been proposed over the years and different indexes are used in different contexts. There is no unifying protocol for clustering evaluation, so it is often unclear which quality index to use in which case. In this chapter, we review the existing clustering quality measures and evaluate them in the challenging context of high-dimensional data clustering. High-dimensional data is sparse and distances tend to concentrate, possibly affecting the applicability of various clustering quality indexes. We analyze the stability and discriminative power of a set of standard clustering quality measures with increasing data dimensionality. Our evaluation shows that the curse of dimensionality affects different clustering quality indexes in different ways and that some are to be preferred when determining clustering quality in many dimensions.

1 Introduction

Unsupervised learning arises frequently in practical machine learning applications and clustering is considered to be one of the most important unsupervised learning tasks. Dividing the data up into clusters helps with uncovering the hidden structure in the data and is a useful step in many data processing pipelines.

N. Tomašev (✉)
Artificial Intelligence Laboratory and Jožef Stefan International Postgraduate School,
Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
e-mail: nenad.tomasev@gmail.com

M. Radovanović
Faculty of Sciences, Department of Mathematics and Informatics,
University of Novi Sad, Trg Dositeja Obradovića 4, 21000 Novi Sad, Serbia
e-mail: radacha@dmi.uns.ac.rs

M.E. Celebi, K. Aydin (eds.), *Unsupervised Learning Algorithms*,
DOI 10.1007/978-3-319-24211-8_4

Clustering is a difficult problem and it is often not clear how to reach an optimal cluster configuration for non-trivial datasets. This problem extends not only to the choice of the clustering algorithm itself, but also to choosing among the generated cluster configurations in case of non-deterministic clustering methods.

In order to be able to compare different cluster configurations and establish preference among them, various clustering quality indexes have been introduced. Each quality index quantifies how well a configuration conforms to some desirable properties.

Internal clustering quality indexes usually measure compactness and separation between the clusters, while the *external* indexes are based on the ground truth about the optimal partition of the data. *Relative* quality indexes are primarily used for choosing the best and most stable output among the repetition of a single clustering method with varying parameters.

The main problem with clustering evaluation is that there is no single preferred clustering quality criterion and it is often not clear which criterion to use for a given application. Different quality indexes should be preferred in different contexts.

High intrinsic data dimensionality is known to be challenging for many standard machine learning methods, including the unsupervised learning approaches. Surprisingly, it has been given comparatively little attention in the context of clustering evaluation.

This chapter gives an overview of the existing approaches for clustering quality evaluation and discusses the implications of high intrinsic data dimensionality for their selection and applicability.

2 Basic Notation

This section introduces the basic notation that will be used throughout the text.

Even though clustering is usually applied to unlabeled data, ground truth information might be available, so the presented definitions include this possibility. Alternatively, these labels might also represent categories in the data that do not correspond to the underlying clusters.

Let $T = \{(x_1, y_1), (x_2, y_2) \ldots (x_N, y_N)\}$ represent the training data given over $X \times Y$, where X is the feature space and Y the finite label space.

A significant portion of the analysis presented in Sect. 7.3 has to do with assessing the influence of prominent neighbor points on clustering quality estimation. Therefore, we introduce the basic notation for k-nearest neighbor sets as follows.

The k-neighborhood of x_i will be denoted by $D_k(x_i) = \{(x_{i1}, y_{i1}), (x_{i2}, y_{i2}) \ldots (x_{ik}, y_{ik})\}$. Any $x \in D_k(x_i)$ is a neighbor of x_i and x_i is a reverse neighbor of any $x \in D_k(x_i)$. An occurrence of an element in some $D_k(x_i)$ is referred to as k-occurrence. $N_k(x)$ denotes the k-occurrence frequency of x. In the supervised case, the occurrences can be further partitioned based on the labels of the reverse neighbor points.

In the context of clustering, a hard partition of the data is defined as a collection $\{C_1, C_2 \dots C_K\}$ of non-overlapping data subsets where K is the number of clusters. In general, the clusters are required to cover the data, so $\bigcup_{i=1}^{K} C_i = T$. In some cases when outlier removal is done along with the clustering, this condition can be relaxed.

Alternative definitions exist for fuzzy or soft partitioning of the data, as well as overlapping partitioning. The overlap usually arises in subspace clustering methods that allow for the same point to belong to different clusters in different projections [1].

Since each data point is assigned to exactly one cluster under the examined framework, denote the cluster label of x_i by $\hat{y}_i$. Given the ground truth and a clustering of the data, it is possible to define the following quantities:

TP: True positive count, as $\text{TP} = |x_i, x_j : y_i = y_j \wedge \hat{y}_i = \hat{y}_j|$.
FP: False positive count, as $\text{FP} = |x_i, x_j : y_i \neq y_j \wedge \hat{y}_i = \hat{y}_j|$.
TN: True negative count, as $\text{TN} = |x_i, x_j : y_i \neq y_j \wedge \hat{y}_i \neq \hat{y}_j|$.
FN: False negative count, as $\text{FN} = |x_i, x_j : y_i = y_j \wedge \hat{y}_i \neq \hat{y}_j|$.

Many clustering methods and evaluation techniques take into account the *centroid* of each cluster which is obtained by averaging across all dimensions. The centroid of cluster i will be denoted by $\bar{x_i}$. Similarly, the centroid (mean) of the entire sample will be denoted by $\bar{x}$.

3 Problems in Analyzing High-Dimensional Data

The term "Curse of Dimensionality" was first coined by Bellman [8] to denote difficulties that arise when working with high-dimensional data. An increase in data dimensionality induces an increase of the containing volume, which in turn leads to sparsity. Sparse data is difficult to handle, since it becomes very hard to obtain reliable density estimates. The amount of data required to derive statistically sound estimates rises exponentially with the number of dimensions. Even the big datasets for large-scale industrial applications do not usually contain enough data to entirely overcome these issues.

The curse of dimensionality poses new challenges for many types of machine learning and pattern recognition methods, including similarity search [20, 21, 90], classification [73], kernel methods [9, 27], privacy-preserving data preprocessing [4, 5], artificial neural networks [87], and clustering [63].

Some types of learning methods are affected by high embedding data dimensionality, while others only in case of high intrinsic data dimensionality. There are certain data domains, like time-series data, where a high embedding data dimensionality does not necessarily imply a high intrinsic dimensionality, due to correlation between the measurements.

The manifold assumption is often used in practical machine learning approaches, where data is assumed to lie on a lower-dimensional manifold in a high-dimensional

space. Restricting the analysis to such manifolds via projections can improve the performance of many methods [74, 92]. The intrinsic dimensionality of the data can be estimated in various ways [17, 29, 37, 62].

There are two phenomena in particular that might be highly relevant for analyzing the behavior of clustering evaluation in intrinsically high-dimensional data: distance concentration and hubness. They are discussed in Sects. 3.1 and 3.2, respectively.

3.1 Distance Concentration

Many clustering evaluation measures take into account either the average or the extremal distances between points that belong to certain clusters in the partitioning.

The concentration of distances in intrinsically high-dimensional data is a highly negative aspect of the curse of dimensionality and a very counterintuitive property of the data [34, 61]. In many dimensions, the relative contrast between the distances calculated on pairs of examples drawn from the same distribution is known to decrease. This makes it hard to distinguish between close and distant points, which is essential in many practical applications.

Let d_M be the maximal observed distance from a fixed query point to other points in the sample and d_m the minimal observed distance. The relative contrast (RC) is then defined as $\rho_d^n = \frac{d_M - d_m}{d_m}$. This definition can be extended to cover all pairs of points. With increasing data dimensionality, relative contrast vanishes, $\lim_{d\to\infty} \rho_d^n = 0$. Distance concentration usually stems from the fact that the expected value for the distance increases, while the variance remains constant, or shrinks.

Distance concentration has an impact on instance-based learning and k-nearest neighbor methods in particular. The difference between nearest and farthest neighbors can appear to vanish in intrinsically high-dimensional data, therefore the very notion of nearest neighbors in many dimensions has been questioned [11, 26, 39]. However, data that follows multiple distributions is less prone to the negative effects of distance concentration, so the nearest neighbor methods remain potentially useful in high dimensions, and can even benefit from it [10, 94].

Establishing the exact conditions for stability of distance functions in high-dimensional data is non-trivial [43, 49]. Many standard metrics are known to be affected by severe distance concentration. Fractional distances are somewhat less susceptible to concentration, though it still occurs [34]. Redesigning metrics for high-dimensional data analysis might lead to certain improvements [3]. Secondary distances have shown promising results in practical applications, including the shared-neighbor distances [42, 46, 91], local scaling, NICDM, and global scaling (mutual proximity) [71].

Not all learning methods are equally susceptible to the distance concentration phenomenon [48]. As for clustering in particular, many quality indexes that are used for clustering evaluation rely on explicit or implicit contrasts over radii or

distance sums. While it might be non-trivial to express their dependency on data dimensionality explicitly in closed-form rules, it is reasonable to assume that they would be at least somewhat affected by distance concentration when it occurs.

3.2 *Hubness: The Long Tail of Relevance and the Central Tendencies of Hubs*

Hubness is a common property of intrinsically high-dimensional data that has recently been shown to play an important role in clustering.

It denotes a tendency of the data to give rise to hubs in the k-nearest neighbor graph as exceedingly frequent and influential nearest neighbors. With increasing dimensionality, the resulting distribution of the neighbor occurrence frequency becomes increasingly skewed. An illustrative example is shown in Fig. 1.

The neighbor occurrence frequency N_k is by convention used to quantify the hubness degree of individual data points in the sample. The hubness of the entire sample is then defined as the skewness of the neighbor occurrence frequency distribution, as per the following formula:

$$SN_k(x) = \frac{m_3(N_k(x))}{m_2^{3/2}(N_k(x))} = \frac{1/N \sum_{i=1}^{N}(N_k(x_i) - k)^3}{(1/N \sum_{i=1}^{N}(N_k(x_i) - k)^2)^{3/2}} \quad (1)$$

Substantial hubness has been shown to arise in most data domains of interest for practical information systems, including text, images, audio signals, and time series.

Hubness is often detrimental to data analysis and can interfere with many standard types of machine learning methods [64–66]. However, it turns out that it is possible to design hubness-aware learning methods that improve on the performance of standard algorithms simply by taking the neighbor occurrence

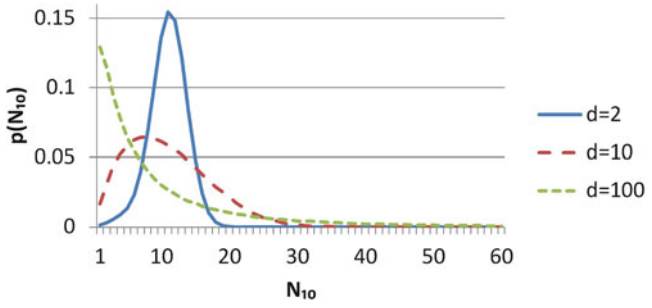

Fig. 1 The change in the distribution shape of 10-occurrence frequency (N_{10}) in i.i.d. Gaussian data with increasing dimensionality when using the Euclidean distance, averaged over 50 randomly generated data sets [78]

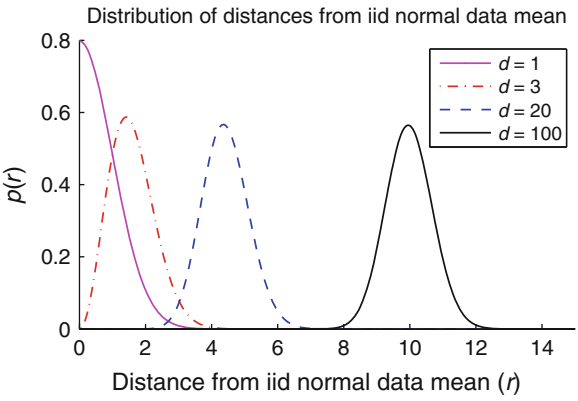

Fig. 2 Probability density function of observing a point at distance r from the mean of a multivariate d-dimensional normal distribution, for $d = 1,\ 3,\ 20,\ 100$ [65]

frequency distribution explicitly into account by learning training data neighbor occurrence models. Hubness-aware methods have been proposed for data reduction [15], ranking [80], representation learning [82], metric learning [71, 72, 76, 79], classification [77, 78, 81], outlier detection [67], and clustering [83].

Points in certain regions of the data space are more likely to become hubs than others, under standard distance measures. This can be related to the distance concentration. The concentration causes points to lie approximately on hyperspheres around cluster means. Non-negligible variance ensures that some points remain closer to the means than others [34, 35] and these points tend to be closer on average to the remaining points in the sample as well. While this might hold for any particular dimensionality, the phenomenon is amplified in high-dimensional data and the central points have a much higher probability of being included in k-nearest neighbor sets of other points in the data. This line of reasoning suggests that hubness might be closely linked to centrality in such cases.

As an illustration, consider Fig. 2. It depicts the distribution of Euclidean distances of all points to the true data mean, for i.i.d. Gaussian data, for different d values. The distribution of distances is the Chi distribution with d degrees of freedom [65]. As previously mentioned, the variance of distance distributions is asymptotically constant with respect to increasing d, unlike the means that asymptotically behave like $\sqrt{d}$ [34, 65]. Their ratio tends to 0 as $d \to \infty$, which reaffirms the hyper-sphere view of high-dimensional data.

Figure 3 further illustrates why hubs are relevant for clustering and why hubness can be used as a measure of local cluster centrality in high-dimensional data. It shows the interaction between norm, hubness, and density estimates in synthetic Gaussian data.

Density estimates represent a reasonable way to estimate cluster centrality and interior regions close to compact cluster centers tend to have a higher density than other regions of the data space, as shown in Fig. 3a. However, this natural interpretation of density no longer holds in high-dimensional data (Fig. 3b) and it is no longer possible to rely on density as the primary indicator of centrality.

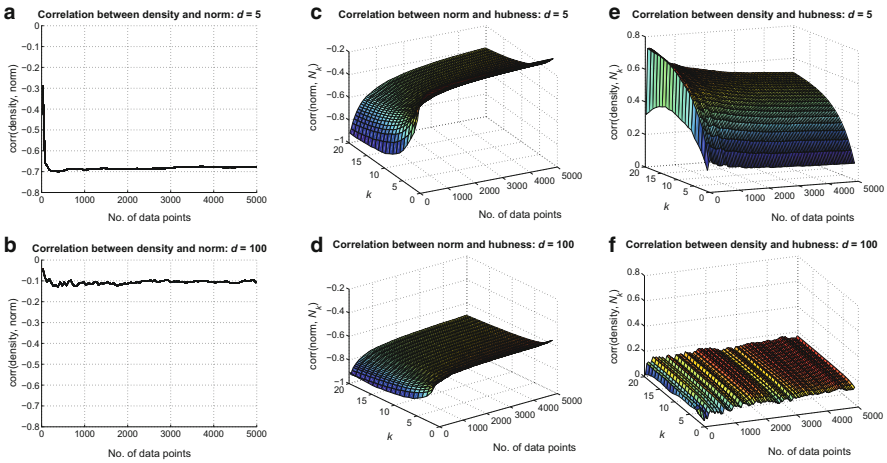

Fig. 3 Interaction between norm, hubness, and density in the simulated setting, in low- and high-dimensional scenarios. (**a**) Correlation between density and norm for $d = 5$. (**b**) Correlation between density and norm for $d = 100$. (**c**) Correlation between norm and hubness for $d = 5$. (**d**) Correlation between norm and hubness for $d = 100$. (**e**) Correlation between density and hubness for $d = 5$. (**f**) Correlation between density and hubness for $d = 100$ [83]

Unlike density, hubness exhibits the opposite trend. It is not highly correlated with centrality in low-dimensional data (Fig. 3c), but an increase in dimensionality gives rise to a very strong correlation between the two (Fig. 3d), so it becomes possible to use hubness as a centrality indicator.

Figure 4 compares the major cluster hubs to medoids and other points in the same simulated setting, based on their pairwise distances and distances to the distribution mean. It can be seen in Fig. 4a, b that in high dimensions the hub is equally informative about the location of the cluster center as the medoid, while in low dimensions the hub and medoid are unrelated. However, this does not imply that hubs and medoids then correspond to the same data points, as Fig. 4c, d shows that the distances from hubs to medoids remain non-negligible. This is also indicated in Fig. 4e, f that shows the ratio between hub to medoid distance and average pairwise distance. In addition, Fig. 4f suggests that in high dimensions hubs and medoids become relatively closer to each other.

Due to its properties, hubness can be exploited for high-dimensional data clustering [84].

In this chapter we have demonstrated that it can also be used as a tool to analyze the behavior of certain clustering quality indexes.

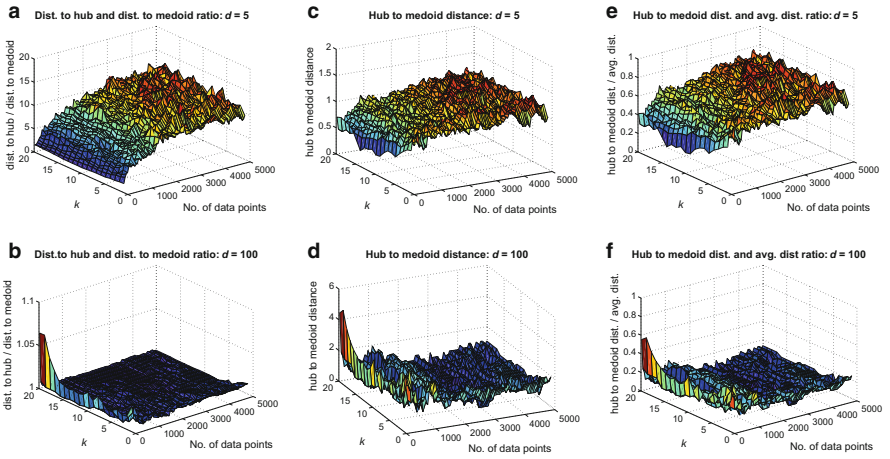

Fig. 4 Interaction between hubs, medoids, and other points in the simulated setting, expressed through distances, in low- and high-dimensional scenarios. (**a**) Ratio between distance to hub and distance to medoid for $d = 5$. (**b**) Ratio between distance to hub and distance to medoid for $d = 100$. (**c**) Hub to medoid distance for $d = 5$. (**d**) Hub to medoid distance for $d = 100$. (**e**) Ratio between hub to medoid distance and average pairwise distance for $d = 5$. (**f**) Ratio between hub to medoid distance and average pairwise distance for $d = 100$ [83]

4 Clustering Techniques for High-Dimensional Data

Clustering high-dimensional data can be challenging and many different techniques have been proposed to this end.

The manifold assumption is often used for subspace clustering of text documents [47, 50], images, and data streams [2, 59]. It is not always necessary to explicitly form a mapping to a feature subspace. The process can be simulated by iterative or non-iterative feature weighting that affects the influence of different features on the proximity measures [6, 19, 53]. Increasing the weights of low-variance dimensions is a common approach [12].

Standard clustering methods designed for low-dimensional data do not perform well in many dimensions and model-based approaches are often over-parametrized [14]. However, they can be applied in conjunction with subspace methods, in order to perform partitioning in lower-dimensional feature subspaces [47]. Different types of hybrid approaches can be used in subspace clustering methods, including density-based techniques, K-means and its extensions and decision trees [51].

Despite that, subspace clustering methods should not be viewed simply as a variation of standard methods with a different notion of similarity that is based on a particular subspace, as different clusters might lie in different subspaces of the data.

Common alternatives to subspace clustering include approximate expectation maximization (EM) [23], spectral clustering [89], shared-neighbor methods [46, 91, 93] and relevant set correlation [41, 88], and clustering ensembles [31, 32].

Hubness-based clustering has recently been proposed for high-dimensional clustering problems [83, 84] and has been successfully applied in some domains like document clustering [40].

Careful data preprocessing and feature selection in particular are often an integral part of the clustering process [13, 28].

5 Clustering Quality Indexes: An Overview

Clustering evaluation is a complex task and many different approaches have been proposed over the years. Most approaches incorporate measures of compactness and separation between the proposed clusters. In this chapter we examine 18 different clustering quality indexes.

Silhouette index is a very well-known clustering evaluation approach that introduces clustering quality scores for each individual point and calculates the final quality index as an average of the point-wise quality estimates [69]. Each point-wise estimate for a point $x_p \in C_i$ is derived from two quantities: $a_{i,p}$ and $b_{i,p}$ which correspond to the average distance to other points within the same cluster and the minimal average distance to points from a different cluster, respectively. Formally, $a_{i,p} = \frac{1}{|C_i|-1} \sum_{x_q \in C_i, q \neq p} \parallel x_q - x_p \parallel$ and $b_{i,p} = \min_{j \in \{1 \ldots K\}, i \neq j} \frac{1}{|C_j|} \sum_{x_q \in C_j} \parallel x_q - x_p \parallel$.

$$\mathrm{SIL}(x_p) = \frac{a_{i,p} - b_{i,p}}{\max a_{i,p}, b_{i,p}} \tag{2}$$

$$\mathrm{SIL} = \frac{1}{N} \sum_{p=1}^{N} \mathrm{SIL}(x_p) \tag{3}$$

The time complexity of the standard Silhouette criterion is $O(dN^2)$, making it difficult to scale to large datasets.

Simplified Silhouette index is an approximation of the original Silhouette coefficient that computes the intra-cluster and inter-cluster distances as distances to the respective cluster centroids [85], thus achieving a significant speed-up and an overall time complexity of $O(dNK)$.

Dunn index is a simple index that has been frequently used in the past. It is defined as a ratio between the minimal between-cluster distance and the maximal cluster diameter [25].

$$\mathrm{DN} = \min_{i,j\in\{1\ldots K\}, i\neq j} \left(\frac{\min_{x_p\in C_i} \min_{x_q\in C_j} \parallel x_p - x_q \parallel}{\max_{l\in\{1\ldots K\}} \max_{x_p,x_q\in C_l} \parallel x_p - x_q \parallel} \right) \tag{4}$$

The value of the index increases both when the separation between the clusters is improved or when the within-cluster dispersion decreases. The cluster diameter and the between-cluster distance can be defined in more than one way, leading to possible generalizations of the original Dunn index.

Davies–Bouldin index is also based on a ratio between the intra-cluster and inter-cluster distances [22]. It is defined as follows:

$$\mathrm{DB} = \frac{1}{K} \sum_{i=1}^{K} \max_{j\neq i} \left(\frac{\frac{1}{|C_i|} \sum_{x_p\in C_i} \parallel x_p - \bar{x_i} \parallel + \frac{1}{|C_j|} \sum_{x_q\in C_j} \parallel x_q - \bar{x_j} \parallel}{\parallel \bar{x_i} - \bar{x_j} \parallel} \right) \tag{5}$$

The time complexity of computing the Davies–Bouldin index is $O(d(K^2 + N))$. Smaller values of Davies–Bouldin index correspond to better cluster configurations.

Isolation index is defined simply as an average proportion of neighbors in the data that agree with the query point in terms of their cluster label [60]. Assume that $\delta_{p,k} = \frac{|x_q\in D_k(x_p):(\nexists C_i: x_p, x_q\in C_i)|}{k}$ is the local neighborhood disagreement ratio. The isolation index is defined as follows:

$$\mathrm{IS} = \frac{1}{N} \sum_{p=1}^{N} (1 - \delta_{p,k}) \tag{6}$$

A weighted version of the isolation index was also later proposed [36].

***C*-index** is derived from the intra-cluster distances and their extremal values [45]. Let $I_{p,q}$ be the indicator function that equals 1 when p and q belong to the same cluster and 0 otherwise. The factor θ that the index is based on is then defined as $\theta = \sum_{p,q\in\{1\ldots N\}, p\neq q} I_{p,q} \parallel x_p - x_q \parallel$. The final C-index is merely the normalized version of the θ factor, taking into account the maximal and minimal values that θ could possibly take on the data.

$$\mathrm{CInd} = \frac{\theta - \min\theta}{\max\theta - \min\theta} \tag{7}$$

$C\sqrt{K}$ index takes individual features into account and measures their corresponding contributions to the inter-cluster and intra-cluster distances [68]. Let $SST_l = \sum_{p=1}^{N} \parallel x_p^l - \bar{x}^l \parallel^2$ be the contribution of the feature l to the average total divergence from the overall data mean $\bar{x}$. The contribution of l to the between-cluster distances is then given as $SSB_l = SST_l - \sum_{i=1}^{K} \sum_{x_p\in C_i} (x_p^l - \bar{x_i}^l)^2$. The $C\sqrt{K}$ index is calculated from SST_l and SSB_l as follows:

$$\mathrm{C}\sqrt{\mathrm{K}}\mathrm{Ind} = \frac{1}{d \cdot \sqrt{K}} \sum_{l=1}^{d} \sqrt{\frac{SSB_l}{SST_l}} \tag{8}$$

The division by $\sqrt{K}$ is used to counterbalance the expected increase of the $\sqrt{\frac{SSB_l}{SST_l}}$ ratio with increasing K. This is necessary in order to prevent the index from being biased toward preferring configurations that contain more clusters by design.

Calinski–Harabasz index is defined as a variance ratio between inter-cluster and intra-cluster dispersions [16]. Let V_B and V_W be the inter-cluster and intra-cluster dispersion matrices.

$$V_B = \sum_{i=1}^{K} |C_i|(\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^T \tag{9}$$

$$V_W = \sum_{i=1}^{K} \sum_{x_p \in C_i} (x_p - \bar{x}_i)(x_p - \bar{x}_i)^T \tag{10}$$

The Calinski–Harabasz index is then defined as:

$$\mathrm{CH} = \frac{\mathrm{trace}(V_B)}{\mathrm{trace}(V_W)} \cdot \frac{N-K}{K-1} \tag{11}$$

Compact and well-separated cluster configurations are expected to have high inter-cluster variance and a comparatively low intra-cluster variance, leading to the high values of the Calinski–Harabasz index. The traces of the inter-cluster and intra-cluster dispersion matrices can be calculated efficiently in $O(dN)$ time.

Fowlkes–Mallows index is defined as the square root of the product of precision and recall given the ground truth of the data [33]. In particular, if $\mathrm{prec} = \frac{\mathrm{TP}}{\mathrm{TP+FP}}$ and $\mathrm{recall} = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$ then the Fowlkes–Mallows index can be calculated as $\mathrm{FM} = \sqrt{\mathrm{prec} \cdot \mathrm{recall}}$. A higher Fowlkes–Mallows index corresponds to a better match between the two compared partitions of the data. It can also be used to compare two different generated clustering outputs and has been used for this in case of hierarchical clustering.

Goodman–Kruskal index is derived from the concepts of concordant and discordant distance pairs [7]. A pair of distances is said to be concordant in case the distance between the objects from the same cluster is lower than the distance between the objects belonging to different clusters. Denote the total number of concordant pairs in the data w.r.t. the partitioning induced by the clustering under evaluation as S_+ and the number of discordant distance pairs as S_-. Goodman–Kruskal is then defined simply as a ratio involving these two quantities.

$$\mathrm{GK} = \frac{S_+ - S_-}{S_+ + S_-} \tag{12}$$

G_+ **index** is another simple index derived from the concepts of concordance and discordance among pairs of distances in the data. Unlike the Goodman–Kruskal index, it only takes into account the discordant counts. Let $t = \frac{N(N-1)}{2}$ be the number of distances defined by N data points. The G_+ index is given as the count of discordant distance pairs normalized by the total number of distance comparisons, namely $G_+ = \frac{2S_-}{t(t-1)}$. The higher values of the G_+ index correspond to a lower clustering quality, unlike in most other indexes. This is why we have used $\bar{G}_+ = 1 - G_+$ in our experiments instead.

Hubert's Γ **statistic** can be used to calculate the correlation between two data matrices. It is often used for calculating the correlation between the original and projected data, though it is also possible to apply it for quantifying the correlation between the two distance matrices, which makes it possible to apply for evaluating clustering quality. Namely, let Ψ_T be the distance matrix on the training data. We define the post-clustering centered distance matrix $\bar{\Psi}_T$ to be the distance matrix where $d(x_p, x_q)$ has been replaced by $d(\bar{x}_{y_p}, \bar{x}_{y_q})$ which is the distance between the corresponding centroids. Hubert's statistic is defined as follows:

$$\Gamma_N = \frac{\sum_{p,q=1}^{N} \Psi_T(p,q) \cdot \bar{\Psi}_T(p,q)}{\sqrt{\sum_{p,q=1}^{N} (\Psi_T(p,q) - \mu_{\Psi_T})^2 \cdot \sum_{p,q=1}^{N} (\bar{\Psi}_T(p,q) - \mu_{\bar{\Psi}_T})^2}} = \frac{\sigma_{\Psi_T, \bar{\Psi}_T}}{\sqrt{\sigma^2_{\Psi_T} \cdot \sigma^2_{\bar{\Psi}_T}}} \tag{13}$$

This is the normalized version of Γ and it is used in most practical cases. It is also possible to use the non-normalized alternative that is defined simply as the average cross-matrix element product, namely $\Gamma = \frac{1}{N} \sum_{p,q=1}^{N} \Psi_T(p,q) \cdot \bar{\Psi}_T(p,q)$.

McClain–Rao index represents the quotient between the mean intra-cluster and inter-cluster distances. Again, let us denote for simplicity by $I_{p,q}$ the indicator function that equals 1 when p and q belong to the same cluster and 0 otherwise. Let b_d and w_d represent the number of inter-cluster and intra-cluster pairs, so that $b_d = \sum_{i=1}^{K} |C_i|(N - |C_i|)$ and $w_d = \sum_{i=1}^{K} \frac{|C_i|(|C_i|-1)}{2}$.

$$\text{MCR} = \frac{\sum_{p,q \in \{1 \ldots N\}, p \neq q} I_{p,q} \parallel x_p - x_q \parallel}{\sum_{p,q \in \{1 \ldots N\}, p \neq q} (1 - I_{p,q}) \parallel x_p - x_q \parallel} \cdot \frac{b_d}{w_d} \tag{14}$$

PBM index is given as a normalized squared ratio between inter-cluster and intra-cluster distances, where they are calculated w.r.t. cluster centroids.

$$\text{PBM} = \left(\frac{\sum_{p=1}^{N} \parallel x_p - \bar{x} \parallel}{\sum_{x_q \in C_j} \parallel x_q - \bar{x}_j \parallel} \cdot \frac{\max_{i,j \in \{1 \ldots K\}, i \neq j} \parallel \bar{x}_i - \bar{x}_j \parallel}{K} \right)^2 \tag{15}$$

Point-biserial index is a maximization criterion based on inter-cluster and intra-cluster distances [55]. Assume that $t = \frac{N(N-1)}{2}$ corresponds to the number of pairs of points in the data. Let b_d and w_d represent the number of inter-cluster and intra-cluster pairs. Furthermore, let d_w represent the average intra-cluster

distance and d_b represent the average inter-cluster distance. Denote by σ_d the standard distance deviation.

$$\text{PBS} = \frac{(d_b - d_w) \cdot \frac{\sqrt{w_d \cdot b_d}}{t}}{\sigma_d} \tag{16}$$

RS index denotes the ratio between the sum of squared distances between the clusters and the total squared distance sum, where distances are calculated between the points and the respective centroids.

$$\text{RS} = \frac{\sum_{p=1}^{N} \parallel x_p - \bar{x} \parallel^2 - \sum_{i=1}^{K} \sum_{x_p \in C_i} \parallel x_p - \bar{x_i} \parallel^2}{\sum_{p=1}^{N} \parallel x_p - \bar{x} \parallel^2} \tag{17}$$

Rand index is an index that is based on simple pairwise label comparisons with either the ground truth or alternative cluster configurations. Let a be the number of pairs of points in the same cluster with the same label, b the number of pairs of points in same cluster with different labels, c the number of pairs of points in different clusters with the same label, and d the number of pairs of points in different clusters with different labels. Rand index of clustering quality is then defined as the following ratio:

$$\text{RAND} = \frac{a + d}{a + b + c + d} \tag{18}$$

However, the Rand index has known weaknesses, such as bias toward partitions with larger numbers of clusters. An improved version of the Rand index was proposed by [44], referred to as the **adjusted Rand index (ARI)**, and is considered to be one of the most successful cluster validation indices [70]. ARI can be computed as:

$$\text{ARI} = \frac{\binom{N}{2}(a + d) - [(a + b)(a + c) + (c + d)(b + d)]}{\binom{N}{2}^2 - [(a + b)(a + c) + (c + d)(b + d)]} \tag{19}$$

SD index is a combination of two factors: scatter and separation. They will be denoted by F_S and F_D, respectively. The total SD index is defined as $\text{SD} = \alpha F_S + F_D$. The scatter component of the SD index is derived from the global and within-cluster variance vectors over different data features. Denote by V_T and V_{C_i} such d-dimensional feature variance vectors, for $i \in \{1 \ldots K\}$. The scatter component is defined as the ratio between the average within-cluster variance vector norm and the global variance vector norm, $F_S = \frac{\sum_{i=1}^{K} \|V_{C_i}\|}{K \cdot \|V_T\|}$. On the other hand, separation is defined via between-centroid distances, $F_D = \frac{\max_{i \neq j} \|\bar{x_i} - \bar{x_j}\|}{\min_{i \neq j} \|\bar{x_i} - \bar{x_j}\|} \cdot \sum_{i=1}^{K} \frac{1}{\sum_{j=1, j \neq i}^{K} \|\bar{x_i} - \bar{x_j}\|}$.

τ index represents the correlation between the distance matrix of the data and a binary matrix corresponding to whether pairs of points belong to the same cluster or not. It can be written down in terms of concordance and discordance rates. Again, let $t = \frac{N(N-1)}{2}$ be the number of distances defined by N data points. Similarly, let b_d and w_d represent the number of inter-cluster and intra-cluster pairs, respectively. Let $t_{bw} = \binom{b_d}{2} + \binom{w_d}{2}$ be the total number of distance pairs that can not be considered discordant since they belong to the same distance type.

$$\tau = \frac{S_+ - S_-}{\left(\frac{t(t-1)}{2} - t_{bw}\right)\frac{t(t-1)}{2}} \tag{20}$$

This is by no means an exhaustive list of clustering quality measures. Many more exist and continue to be developed.

6 Clustering Quality Indexes: Existing Surveys

Several studies were conducted in order to elucidate the overall applicability of various clustering quality indexes. Different studies have focused on different aspects of the clustering problem and different properties that good quality measures are expected to have. Detailed surveys of the employed methods discuss the motivation behind the different approaches that can be taken [38].

In one such study [52], the authors have identified five primary criteria to evaluate the clustering quality indexes by, as follows: non-monotonicity, robustness to noise, proper handling of cluster configurations with highly varying cluster density, handling of subclusters, and performance under skewed distributions. The authors have compared 11 different internal clustering quality measures according to the proposed criteria on a series of carefully crafted examples and have ranked the examined indexes accordingly.

Standard clustering evaluation approaches that were tailored for partitional clustering methods [18] are not universal and not well suited for other types of tasks like the evaluation of nested clustering structures, for which they need to be extended and adapted [24].

An inherent problem in many existing studies lies in the fact that one has to know the ground truth of all the clustering problems used to evaluate the indexes on the level of detail required for properly estimating the degree to which they satisfy each of the proposed benchmarking criteria. This is why most examples in such studies are taken to be 2-dimensional [54]. Indeed, it is easy to construct 2-dimensional cluster configurations of arbitrary shape, compactness, and density. More importantly, it is possible to visually inspect the data and confirm the correctness of the assumptions.

In contrast, high-dimensional data is rarely used for such testing and we believe that this was a major flaw in the experimental design of many approaches that aim to establish which clustering quality indexes are to be preferred in practical

applications. This is not to say that the conclusions reached in those studies are any less valuable nor that they cannot be generalized. However, high-dimensional data is often counterintuitive and is known to exhibit different properties than low-dimensional data.

Since most practical applications involve clustering intrinsically high-dimensional data, it is necessary to carefully examine the behavior of clustering quality indexes with increasing data dimensionality in order to better understand their role in high-dimensional clustering evaluation.

Existing studies on the feasibility of clustering evaluation in high-dimensional clustering tasks have mostly been focused on subspace clustering [56, 57] as it is a popular clustering approach in such cases.

This study does not aim to replicate the types of experiments or results that are already available in the literature. Instead, it aims to complement them in order to provide insight into the overall behavior of clustering quality indexes in many dimensions.

It should be noted that there are experimental approaches that can be applied both in the low-dimensional and the high-dimensional case. Most notably, the clustering quality indexes can be compared based on their ability to select the optimal cluster configuration among a series of generated cluster configurations for each dataset, while varying the number of clusters [85]. The main problem with this approach lies in determining a priori what constitutes the optimal clustering configuration for a given dataset. This is often taken to correspond to the number of classes in the data, which is problematic. Even in cases when the ground truth information is available, preferring a clustering configuration produced by a clustering algorithm that has the same number of clusters as the number of clusters in the data is not necessarily optimal. Namely, those proposed clusters need not match the ground truth clusters exactly. In case of mismatch, another configuration might provide a more natural partitioning of the data and this is usually difficult to estimate unless it is possible to visualize the clustering outcomes and inspect the results.

Supervised evaluation can be used as an alternative [30], when labels are available, though many real-world datasets violate the cluster assumption in such way that the classes do not correspond to clusters nor does the data cluster well. In such cases it is not advisable to use the average cluster homogeneity of the selected configurations as the primary evaluation measure.

7 Clustering Evaluation in Many Dimensions

In order to understand the potential problems that might arise when applying different clustering quality criteria to many-dimensional datasets, it is necessary to consider the basic underlying question of what it means to cluster well in terms of index scores.

Most clustering quality indexes are relative and can be used either as an objective function to optimize or as a criterion to make comparisons between different cluster

configurations on the same feature representation of the same dataset, under a chosen distance measure. Ideally, it would be very useful if an index could be applied to perform meaningful comparisons across different feature representations and metric selections as well, or across datasets.

Unlike supervised measures like accuracy or the F_1 score, clustering quality indexes are sometimes non-trivial to interpret. For instance, it is not immediately clear what a Silhouette index of 0.62 means. It is also unclear whether it means the same thing in 2 dimensions as it does in 100 dimensions, assuming the same underlying data distribution.

Different clustering quality indexes capture similar but slightly different aspects of clustering quality. This is why ensembles of clustering quality indexes are sometimes used instead of single quality measures [58, 86]. The implicit assumption of the ensemble approach is that the ensemble constituents exhibit equal sensitivity to the varying conditions in the data. If this assumption is violated, a different index combination would be required for each dataset, which would be highly impractical.

Additionally, if an index is to be used for cluster configuration selection over different sampled feature subspaces on a dataset, then the stability w.r.t. dimensionality and representation is a strict requirement for the comparisons to make sense.

The experiments presented here are aimed at clarifying the issues regarding the sensitivity of clustering quality indexes to data dimensionality, in order to determine whether it is possible to safely use certain indexes for cross-dimensional comparisons and evaluation.

7.1 Experimental Protocol

In order to evaluate the performance and robustness of different cluster configuration quality estimation indexes, a series of intrinsically high-dimensional multi-cluster datasets was generated. Each synthetic dataset consisted of about 10,000 points of i.i.d. Gaussian data based on a diagonal covariance matrix. In turn, ten datasets were generated for each dimensionality, experimental context, and tested cluster number. Tests were run for data with $C \in \{2, 3, 5, 10, 20\}$ in order to see how the number of clusters correlates with index performance.

Two different experimental contexts were examined: well-defined and separated clusters and clusters with significant overlap between them. It was assumed that different indexes might exhibit different behavior in the two cases.

The degree of overlap between the clusters in the overlapping case was controlled via a constrained iterative procedure for generating feature-specific distribution means. Since the diagonal covariance matrices were used, the diagonal entry $\sigma^l_{C_i}$ corresponding to the dispersion of l-th feature in cluster C_i and the mean $\mu^l_{C_i}$ were used to determine where $\mu^l_{C_j}$ should be placed for $j \neq i$ and vice versa. The procedure was executed as follows, for any given dimension (feature).

A permutation $\{i_1, i_2 \ldots i_C\}$ of cluster indexes was randomly selected. The mean and standard deviation of the first Gaussian, $\mu^l_{C_{i_0}}$ and $\sigma^l_{C_{i_0}}$ were selected randomly from a fixed range. For each $p \in \{2 \ldots C\}$, cluster C_{i_p} got randomly paired with a cluster C_{i_q}, for $1 \leq q < p$. The mean and standard deviation for feature l in cluster q were then generated by the following rules.

$$\mu^l_{C_{i_q}} = \mu^l_{C_{i_p}} \pm \alpha\sigma^l_{C_{i_p}} \tag{21}$$

$$\sigma^l_{C_{i_q}} = \beta(1 + N_{0,1})\sigma^l_{C_{i_p}} \tag{22}$$

In the experiments presented in this chapter, the parameter values of $\alpha = 0.5$ and $\beta = 0.75$ were used.

On each dataset in both experimental contexts, the indexes were run both on ground truth cluster configurations as well as a series of cluster configurations produced by repeated runs of K-means clustering. The clustering was repeated 10 times for each dataset. These results will be presented separately. Euclidean distance was used in all cases.

The experiments were based on the clustering quality index implementations in Hub Miner (https://github.com/datapoet/hubminer) [75], within the learning. unsupervised.evaluation.quality package. The experimental framework for tests with increasing dimensionality on synthetic Gaussian data was implemented in the QualityIndexEvalInHighDim class in the experimental sub-package.

All clustering quality indexes in Hub Miner are implemented so as to correspond to a maximization problem. Most existing indexes follow this approach. For those indexes that naturally correspond to minimization problems like Davies Bouldin or G_+, an inverse or a complement was computed and reported instead, in order to present the results in a more consistent way.

7.2 Sensitivity to Increasing Dimensionality

Since the synthetic datasets in all dimensionalities were generated from the same distribution type, differing only in the number of dimensions, the underlying assumption was that the robust clustering quality indexes would yield similar quality estimates in all cases, on average. In contrast, the indexes that are sensitive to data dimensionality would either display different average estimates or would become more or less stable in terms of the variance of quality predictions.

7.2.1 Sensitivity of the Average Quality Assessment

The average clustering quality estimation by different indexes was calculated for each generated dataset and experimental context. These have been subsequently averaged over collections of datasets in order to calculate the average estimates

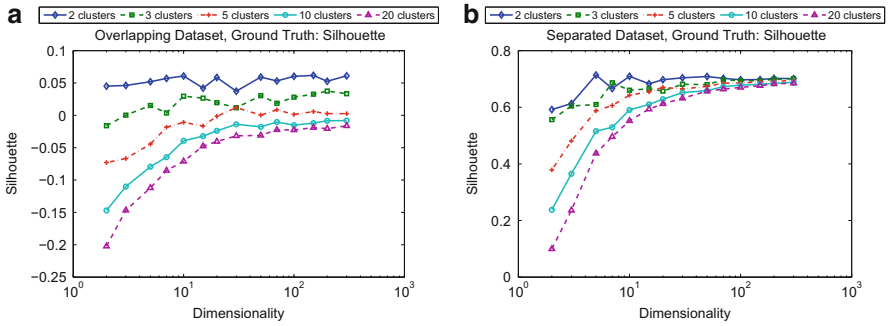

Fig. 5 Average clustering quality index values for the Silhouette index with increasing dimensionality, when evaluated on ground truth cluster labels. (**a**) Overlapping clusters. (**b**) Well-separated clusters

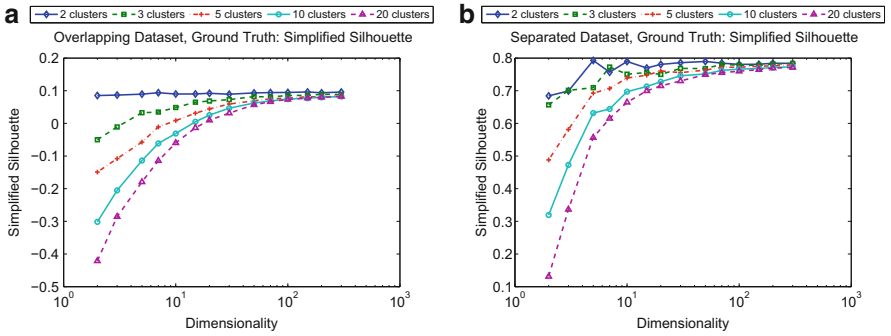

Fig. 6 Average clustering quality index values for the simplified Silhouette index with increasing dimensionality, when evaluated on ground truth cluster labels. (**a**) Overlapping clusters. (**b**) Well-separated clusters

associated with the used Gaussian mixtures for each tested data dimensionality. The results are shown in Figs. 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, and 21.

Some indexes seem to be robust to increasing dimensionality w.r.t. the average cluster configuration quality estimates and these include: isolation index, C-index, $C\sqrt{K}$-index, Calinski–Harabasz, $\bar{G}_+$ index, Goodman–Kruskal, RS index, and τ-index. In these cases, the cluster configuration quality estimates remain similar when the dimensionality is increased. It should be noted that, though there are no emerging trends in these indexes, some of them do exhibit substantial variance in their output.

In other cases, it is possible to observe changes in the clustering quality estimates and there can be substantial differences between the output of a same index for different dimensionalities of the data, both in case of clearly separated clusters as well as overlapping clusters with the same controlled degree of overlap.

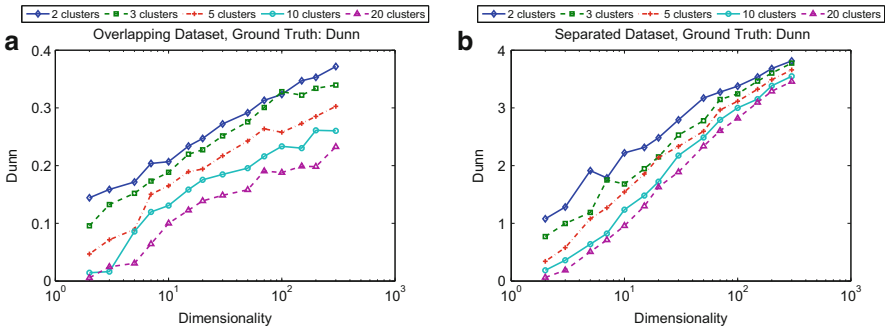

Fig. 7 Average clustering quality index values for the Dunn index with increasing dimensionality, when evaluated on ground truth cluster labels. (**a**) Overlapping clusters. (**b**) Well-separated clusters

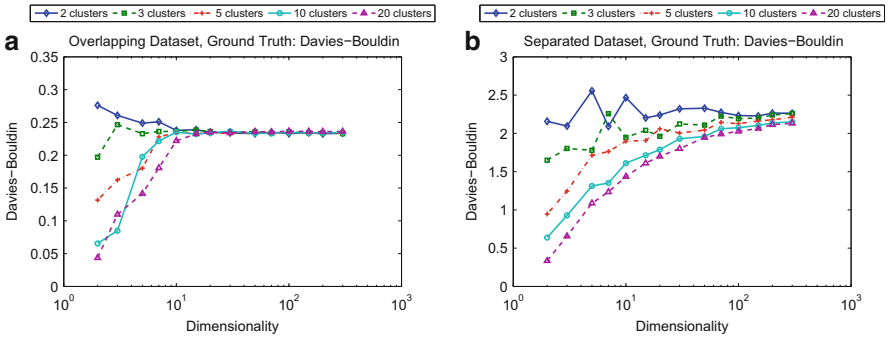

Fig. 8 Average clustering quality index values for the inverted Davies–Bouldin index with increasing dimensionality, when evaluated on ground truth cluster labels. (**a**) Overlapping clusters. (**b**) Well-separated clusters

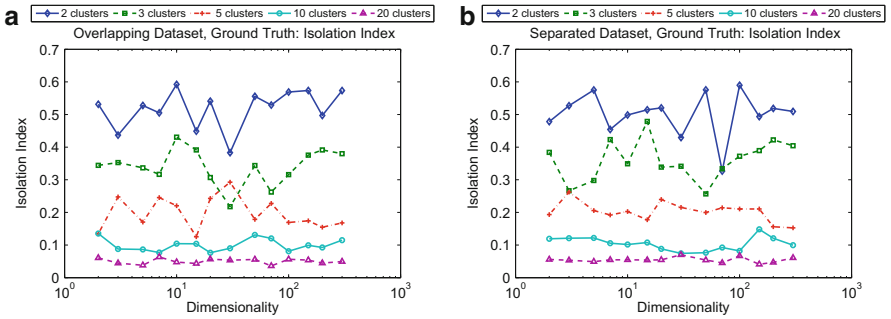

Fig. 9 Average clustering quality index values for the isolation index with increasing dimensionality, when evaluated on ground truth cluster labels. (**a**) Overlapping clusters. (**b**) Well-separated clusters

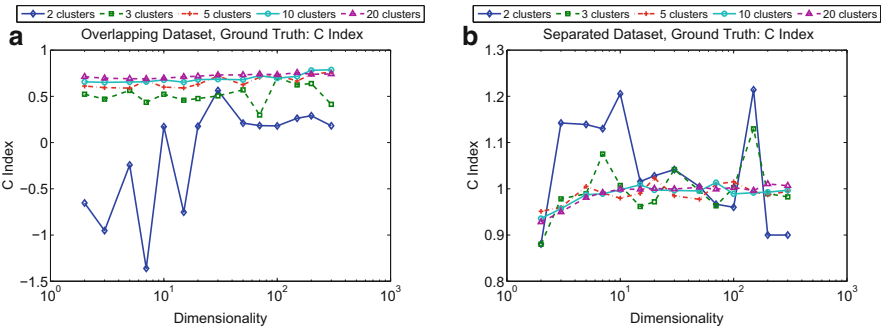

Fig. 10 Average clustering quality index values for the C index with increasing dimensionality, when evaluated on ground truth cluster labels. (**a**) Overlapping clusters. (**b**) Well-separated clusters

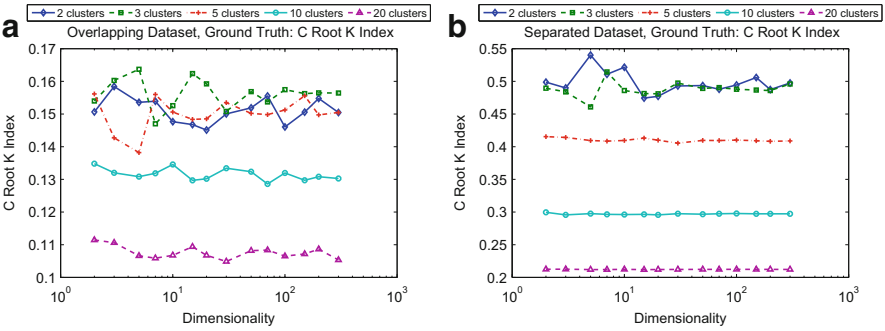

Fig. 11 Average clustering quality index values for the $C\sqrt{K}$ index with increasing dimensionality, when evaluated on ground truth cluster labels. (**a**) Overlapping clusters. (**b**) Well-separated clusters

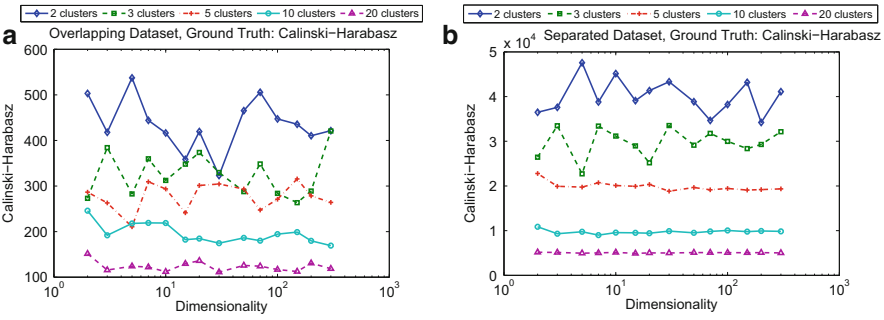

Fig. 12 Average clustering quality index values for the Calinski–Harabasz index with increasing dimensionality, when evaluated on ground truth cluster labels. (**a**) Overlapping clusters. (**b**) Well-separated clusters

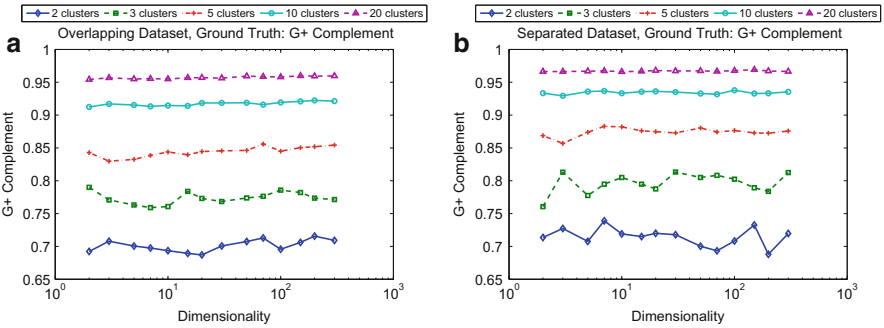

Fig. 13 Average clustering quality index values for the $\bar{G}_+$ index with increasing dimensionality, when evaluated on ground truth cluster labels. (**a**) Overlapping clusters. (**b**) Well-separated clusters

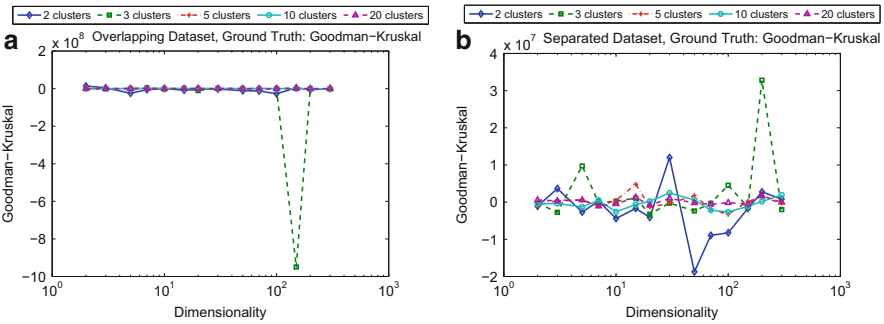

Fig. 14 Average clustering quality index values for the Goodman–Kruskal index with increasing dimensionality, when evaluated on ground truth cluster labels. (**a**) Overlapping clusters. (**b**) Well-separated clusters

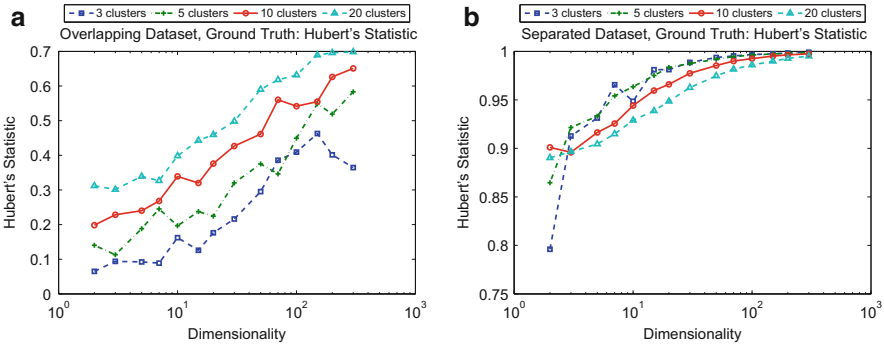

Fig. 15 Average clustering quality index values for the Hubert's statistic with increasing dimensionality, when evaluated on ground truth cluster labels. (**a**) Overlapping clusters. (**b**) Well-separated clusters

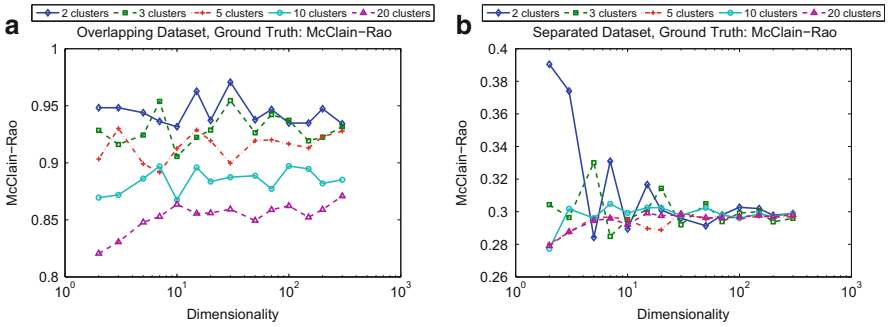

Fig. 16 Average clustering quality index values for the McClain–Rao index with increasing dimensionality, when evaluated on ground truth cluster labels. (**a**) Overlapping clusters. (**b**) Well-separated clusters

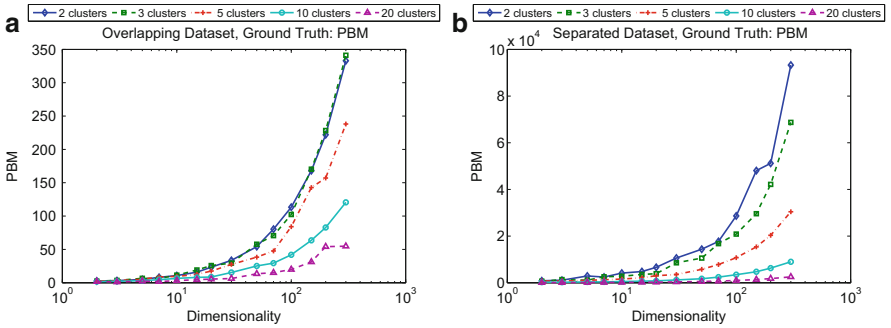

Fig. 17 Average clustering quality index values for the PBM index with increasing dimensionality, when evaluated on ground truth cluster labels. (**a**) Overlapping clusters. (**b**) Well-separated clusters

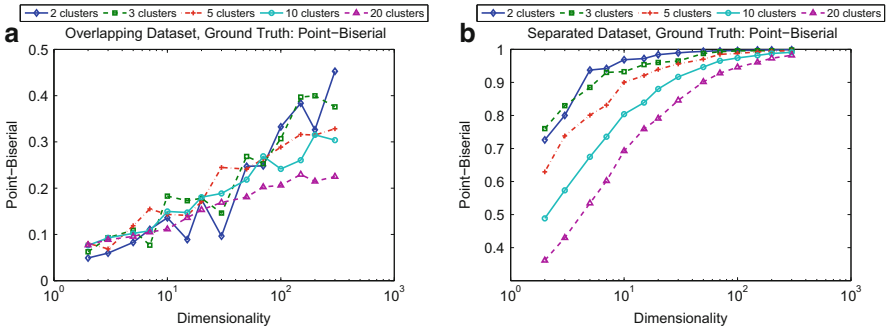

Fig. 18 Average clustering quality index values for the point-biserial index with increasing dimensionality, when evaluated on ground truth cluster labels. (**a**) Overlapping clusters. (**b**) Well-separated clusters

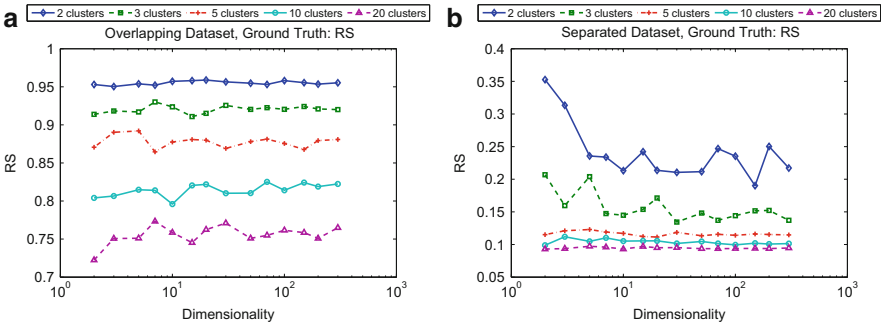

Fig. 19 Average clustering quality index values for the RS index with increasing dimensionality, when evaluated on ground truth cluster labels. (**a**) Overlapping clusters. (**b**) Well-separated clusters

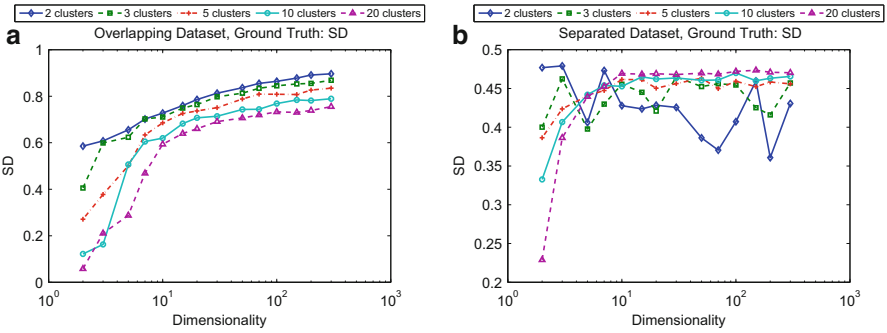

Fig. 20 Average clustering quality index values for the SD index with increasing dimensionality, when evaluated on ground truth cluster labels. (**a**) Overlapping clusters. (**b**) Well-separated clusters

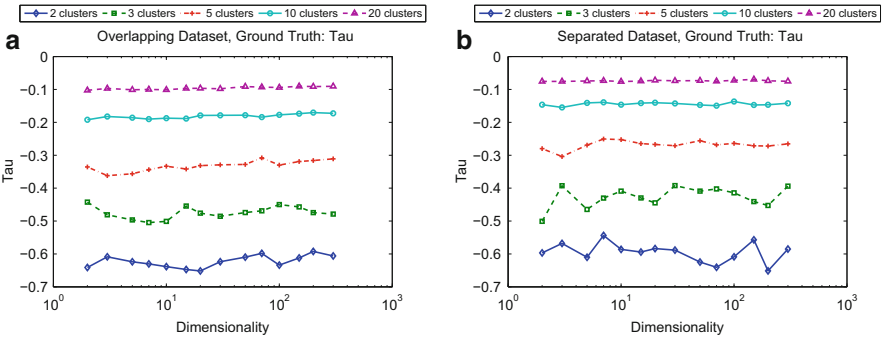

Fig. 21 Average clustering quality index values for the τ-index with increasing dimensionality, when evaluated on ground truth cluster labels. (**a**) Overlapping clusters. (**b**) Well-separated clusters

The estimated clustering quality increases with increasing dimensionality for the Silhouette index, the simplified Silhouette index, Dunn index, inverted Davies–Bouldin index, Hubert's statistic, PBM, and point-biserial index. These

indexes seem to be prone to preferring high-dimensional configurations and interpreting them as better. While it might be the case that introducing more dimensions produces additional separation in commonly used distance measures, this makes it hard to compare the quality of clustering across dimensions and across different feature subsets of the data.

Being able to reliably compare the clustering quality across dimensions is especially important in certain subspace clustering applications where the dimensionality of the observed subspaces varies. It is also of importance in applications where multiple feature representations and feature subsets of the same underlying data are being used in the experiments. In contrast, robustness to changes in dimensionality has no impact on the simple clustering use cases when the dimensionality is fixed and no subspaces are observed. Such use cases are more common in low-dimensional clustering tasks.

There are also notable differences in terms of the preferred number of clusters. The following indexes consistently prefer fewer clusters in the data, regardless of the dimensionality: Silhouette index, simplified Silhouette index, Dunn, Davies Bouldin, isolation index, Calinski–Harabasz index, $\bar{G}_+$ index, PBM index, and RS index. This tendency is also present in $C\sqrt{K}$-index, though to a somewhat lower degree, at least when it comes to lower cluster numbers. It is interesting to see that there are cases where this tendency is present only in case of evaluating clearly separated cluster configurations (point-biserial index, Fig. 18), as well as cases where the preference is present only for overlapping cluster configurations (McClain–Rao index, Fig. 16). Additionally, there is the surprising case of Hubert's statistic, which seems to exhibit a preference for fewer clusters in clearly separated cases and a preference for more clusters in the case of overlapping distributions.

The tendencies of the clustering quality estimates to either increase or remain approximately constant w.r.t. dimensionality do not depend on the number of clusters in the data, which shows that they are an intrinsic property of some of the compared clustering quality indexes.

Furthermore, dependencies on representational dimensionality do not merely affect the cluster configuration quality estimation for ground truth clustering configurations, but for the output of the K-means algorithm as well, which is important in practical terms. Figures 22 and 23 show the behavior of the Fowlkes–Mallows and adjusted Rand indexes with increasing data dimensionality. Both indexes prefer high-dimensional cluster configurations generated by K-means on the synthetic Gaussian data, meaning that K-means was actually more successful in finding the designated true clustering in high dimensions.

Surprisingly, it seems that the tendencies that were observed on ground truth cluster configurations do not necessarily generalize to the tendencies of the same indexes when evaluating the cluster configurations produced by the repeated K-means clustering. While in some cases the dependency curves might take the same shape, it is not always the case. Many indexes that seem not to depend on data dimensionality when it comes to ground truth estimates exhibit a strong dependency on data dimensionality in terms of the K-means quality estimates.

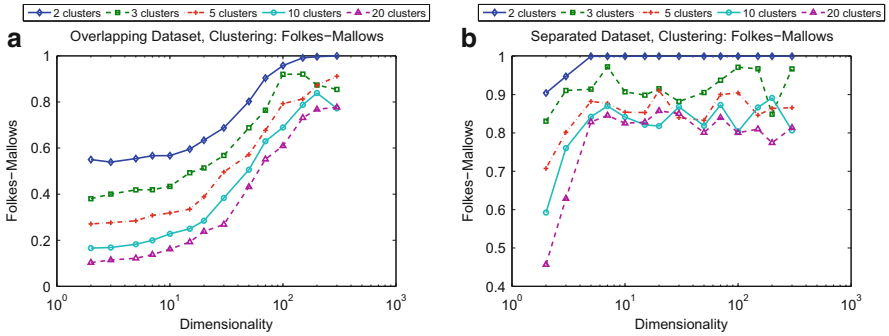

Fig. 22 Average clustering quality index values for the Fowlkes–Mallows index with increasing dimensionality, when evaluated on K-means clustering results. (**a**) Overlapping clusters. (**b**) Well-separated clusters

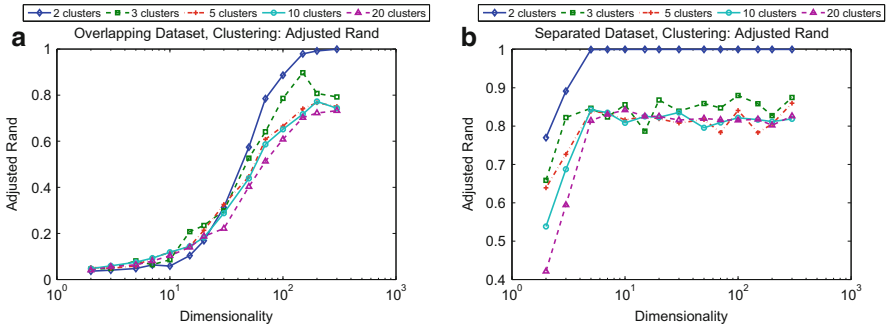

Fig. 23 Average clustering quality index values for the adjusted Rand index with increasing dimensionality, when evaluated on K-means clustering results. (**a**) Overlapping clusters. (**b**) Well-separated clusters

For instance, Fig. 24 gives an example of the behavior of the Calinski–Harabasz index and the $C\sqrt{K}$-index when evaluating the configurations produced by K-means clustering on data of increasing dimensionality. In both cases, the average index values clearly decrease with data dimensionality, unlike what was previously observed for the ground truth configurations. Apparently, changing properties of distance distributions as dimensionality increases (i.e., increasing expected values, stable variances) cause these indexes to produce values that are not comparable across dimensionalities. For the $C\sqrt{K}$-index it seems that normalization by $d\sqrt{K}$ may not be the most appropriate choice to counter the effects of dimensionality on computed contributions of features to the average total divergence from the overall data mean, and to between-cluster distances (concretely, division by d may be too strong, since increasing expected values already cancel out, at least partially, in the ratio of SSB and SST).

Yet, not all indexes exhibit trends that are either constant or monotonically increasing/decreasing. The behavior of several indexes with increasing data dimen-

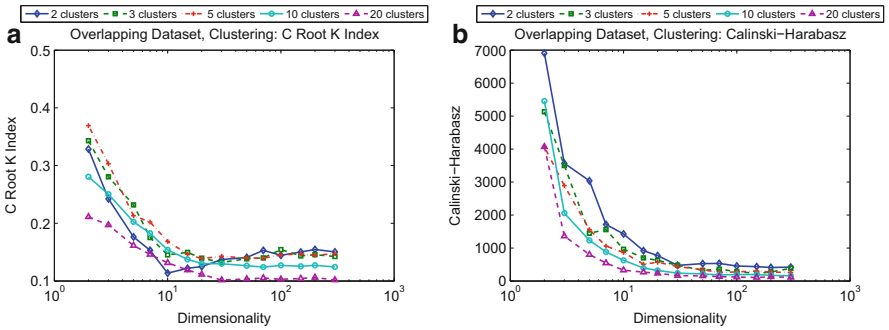

Fig. 24 Average clustering quality index values for the $C\sqrt{K}$ and Calinski–Harabasz index with increasing dimensionality, when evaluated on K-means clustering results. These trends clearly differ from the trends that the same indexes exhibit on ground truth configurations. (**a**) $C\sqrt{K}$, Overlapping clusters. (**b**) Calinski–Harabasz, Overlapping clusters

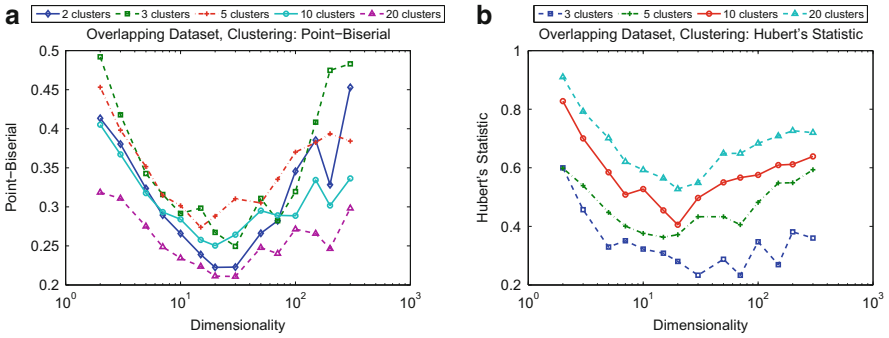

Fig. 25 Average clustering quality index values for the point-biserial index and Hubert's statistic with increasing dimensionality, when evaluated on K-means clustering results. (**a**) Point-biserial, Overlapping clusters. (**b**) Hubert's-Statistic, Overlapping clusters

sionality can be captured by U-shaped curves, as is the case with Hubert's statistic and the point-biserial index. This is shown in Fig. 25. In those cases the average estimates first decrease with increasing dimensionality and then start increasing after reaching a local minimum. This sort of behavior is highly surprising and also suggests that some indexes might have either ideal operating points in terms of the underlying data dimensionality or ranges of dimensionality values that should be avoided.

Overall, for indexes which implicitly or explicitly rely on some combination of notions of within- and between-cluster distances it can be seen that changes in distance distributions which result from increase of dimensionality affect the produced trends in scores in various ways, implying that some indexes adjust to these changes better than others.

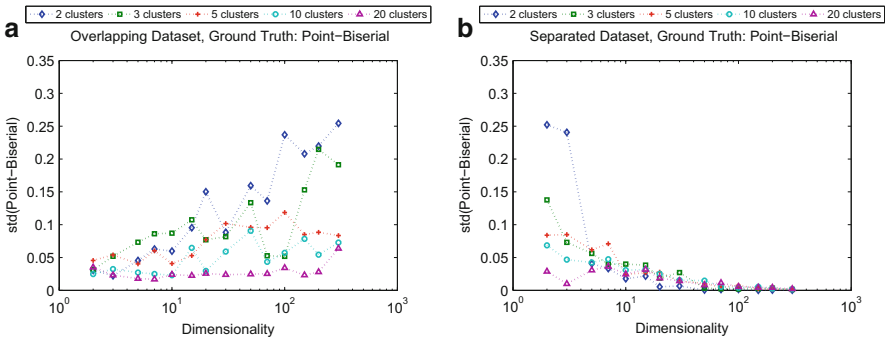

Fig. 26 Standard deviation of the clustering quality index values for the point-biserial index with increasing dimensionality, when evaluated on ground truth cluster labels. (**a**) Overlapping clusters. (**b**) Well-separated clusters

7.2.2 Stability in Quality Assessment

Increasing data dimensionality also impacts the performance of the clustering quality indexes in terms of their stability and associated predictive variance. Similarly to the average index performance, different indexes are influenced in different ways.

Consider, for example, the point-biserial index. The standard deviation of the index values with increasing data dimensionality is given in Fig. 26, both for the overlapping and the clearly separated context. In case of overlapping clusters, the variance of the point-biserial index increases for configurations with a low number of clusters as the dimensionality is increased. In case of separated clusters, the variance decreases with increasing dimensionality.

This is not the case for other quality measures. In case of PBM (Fig. 27), the variance increases both for separated and overlapping cluster configurations. For $\bar{G}_+$ index, the variance does not increase or decrease with dimensionality, but there is a consistent difference between the stability of the estimation when applied to configurations with different numbers of clusters. The stability of the $\bar{G}_+$ index (Fig. 28) decreases for more complex cluster configurations and the standard deviation is the lowest in case of 2 clusters.

It is also interesting to note that the stability of different quality indexes also depends on whether it is determined based on the ground truth configuration quality estimates or the clustering configuration estimates. These trends need to be analyzed separately. The differences are shown in Figs. 29, 30, and 31, for the simplified Silhouette index, Dunn index, and isolation index.

Stability is important in clustering evaluation in order to allow for meaningful interpretations of the results. If the estimates for different repeated samples drawn from the same underlying distribution vary greatly, interpretation of the estimated quality becomes difficult. Stable indexes are to be preferred when used for external comparisons, especially when the number of data samples is low.

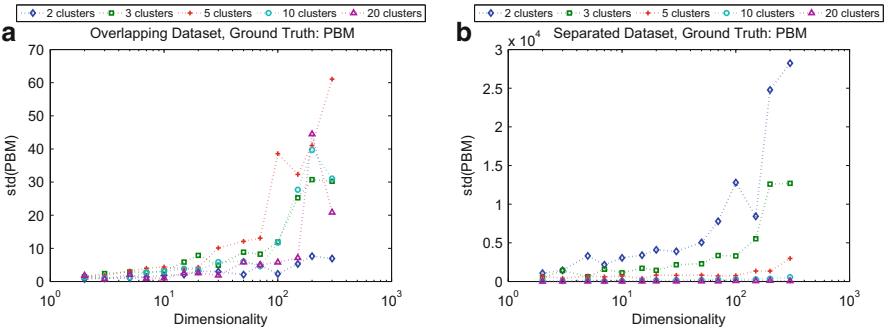

Fig. 27 Standard deviation of the clustering quality index values for the PBM index with increasing dimensionality, when evaluated on ground truth cluster labels. (**a**) Overlapping clusters. (**b**) Well-separated clusters

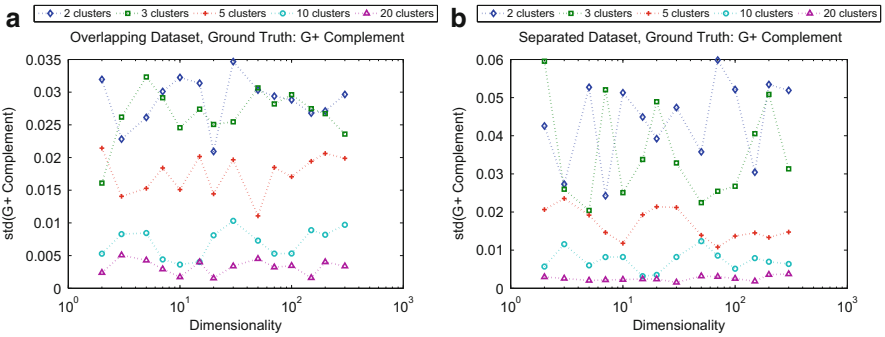

Fig. 28 Standard deviation of the clustering quality index values for the $\bar{G}_+$ index with increasing dimensionality, when evaluated on ground truth cluster labels. (**a**) Overlapping clusters. (**b**) Well-separated clusters

7.3 Quantifying the Influence of Hubs

It turns out that the behavior of many of the evaluated clustering quality indexes can be partially captured and quantified by hubness in the data. Previous studies have shown hubs to cluster poorly in some cases [63] and that proper handling of hub points is required in order to improve clustering quality. Further studies have found that the contribution of hub points to the Silhouette index depends on their hubness and that hubs contribute more to the final Silhouette index estimate [83].

In our experiments we have been able to confirm that it is not only Silhouette index that is susceptible to being influenced by the hubness in the data and the positioning of hub points in the cluster configuration. Many different indexes exhibit similar types of dependencies on hub points, though some also exhibit the reverse trends.

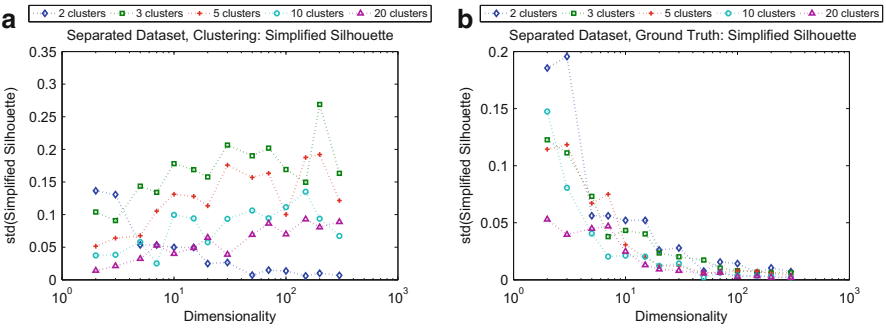

Fig. 29 Different trends in stability of the simplified Silhouette index with increasing dimensionality, depending on whether it was applied to the ground truth cluster labels or the K-means clustering results. These comparisons were performed on synthetic data with well-separated clusters. (**a**) Ground truth evaluation. (**b**) Evaluating K-means results

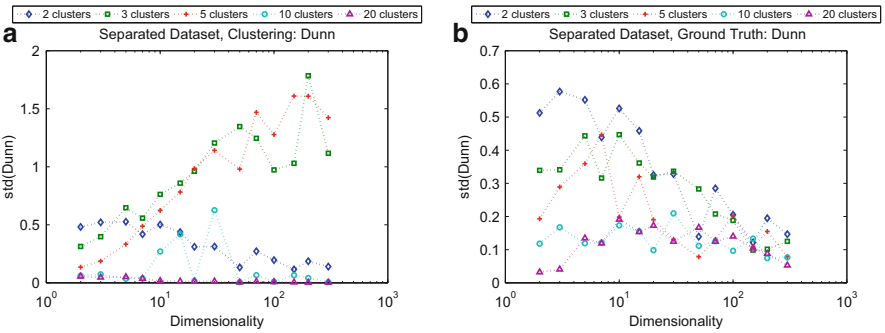

Fig. 30 Different trends in stability of the Dunn index with increasing dimensionality, depending on whether it was applied to the ground truth cluster labels or the K-means clustering results. These comparisons were performed on synthetic data with well-separated clusters. (**a**) Ground truth evaluation. (**b**) Evaluating K-means results

Formally, we denote **hubs** are points $x_h \in D$ such that $N_k(x_h)$ belongs to the top one third of the N_k distribution. We will denote the set of all hubs in T by H_k^T. In contrast, anti-hubs are the rarely occurring points in the data and their number is by convention taken to match the number of hubs on a dataset. The remaining points are referred to as being regular.

Figure 32 shows the contributions of hubs, regular points, and anti-hubs to the final point-biserial index score with increasing dimensionality, for cluster configurations produced by K-means. The contribution of hubs is higher for this index for well-separated configurations while significantly lower for overlapping configurations where anti-hubs dominate the scores. This is unlike the Silhouette index and simplified Silhouette index, where hubs contribute substantially more in both cases. The contributions to the Silhouette index and simplified Silhouette index for overlapping configurations are shown in Fig. 33.

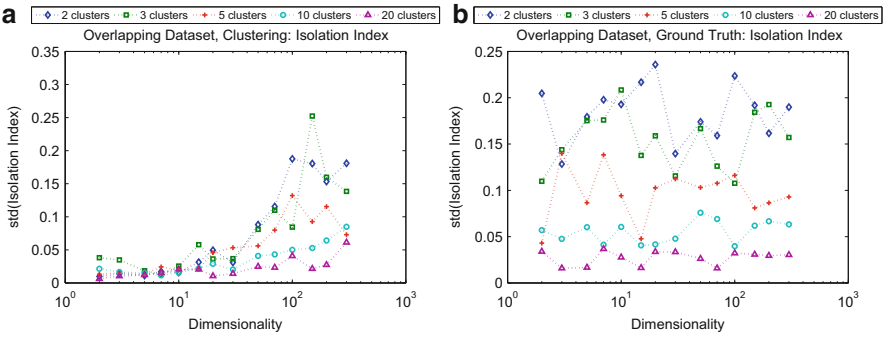

Fig. 31 Different trends in stability of the isolation index with increasing dimensionality, depending on whether it was applied to the ground truth cluster labels or the K-means clustering results. These comparisons were performed on synthetic data with overlapping clusters. (**a**) Ground truth evaluation. (**b**) Evaluating K-means results

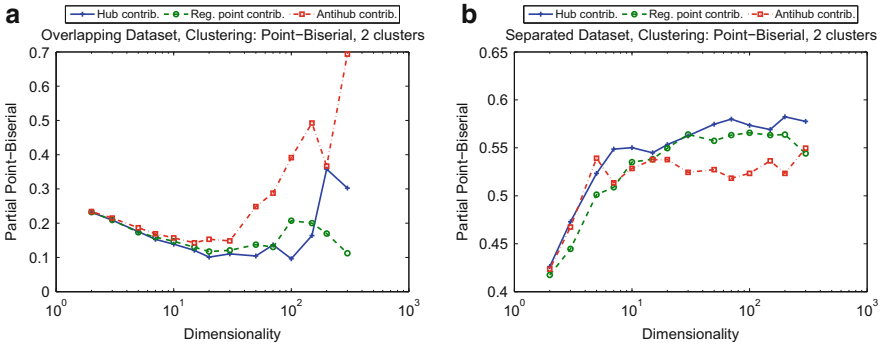

Fig. 32 The contributions of different point types to final index values for the point-biserial index in case of 2 clusters, for cluster configurations produced by K-means. (**a**) Overlapping clusters. (**b**) Well-separated clusters

Hubs seem to be the most influential points in determining the $\bar{G}_+$ and Davies–Bouldin scores as well, as seen in Fig. 34.

In contrast, anti-hubs seem to be contributing more to the partial sums in McClain–Rao index and the difference becomes more apparent with increasing the number of clusters in the data, as shown in Fig. 35.

Whether hub points contribute substantially more or less than regular points for any given index might affect the robustness of the index and its sensitivity to increasing data dimensionality.

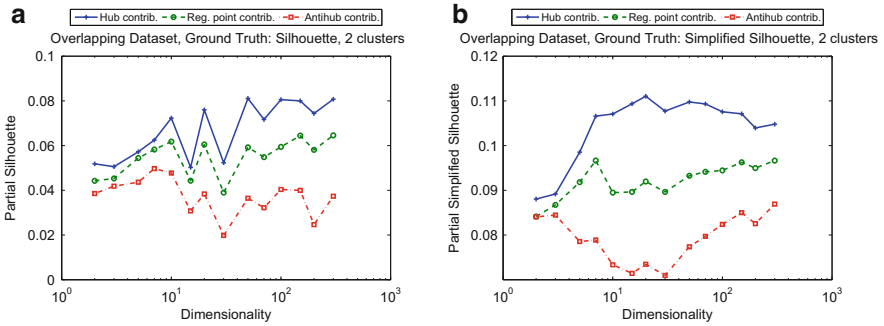

Fig. 33 The contributions of different point types to final index values for the Silhouette and simplified Silhouette index in case of 2 overlapping clusters, for ground truth cluster configurations. (**a**) Silhouette index. (**b**) Simplified Silhouette index

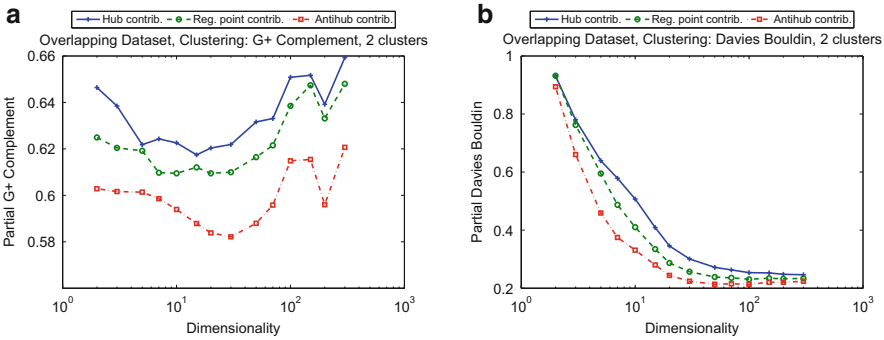

Fig. 34 The contributions of different point types to final index values for the $\bar{G}_+$ and Davies–Bouldin index in case of 2 overlapping clusters, for cluster configurations produced by K-means. (**a**) $\bar{G}_+$ index, Overlapping clusters. (**b**) Davies–Bouldin, Well-separated clusters

8 Perspectives and Future Directions

Robust and effective clustering evaluation plays an essential role in selecting the optimal clustering methods and cluster configurations for a given task. It is therefore important to understand the behavior of clustering quality measures in challenging contexts, like high data dimensionality.

We have examined a series of frequently used clustering quality measures under increasing data dimensionality in synthetic Gaussian data. The experiments have demonstrated that different quality metrics are influenced in different ways by rising dimensionality in the data. Dimensionality of the data can affect both the mean quality value assigned by an index as well as the stability of the quality estimation. Furthermore, the same index can behave differently depending on the presence or absence of overlap between different clusters as well as whether the comparisons are being made to the ground truth or an output of a clustering algorithm.

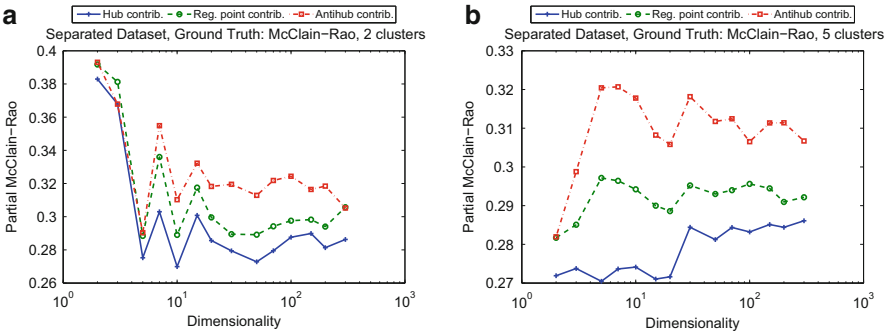

Fig. 35 The contributions of different point types to final index values for the McClain–Rao index in case of clearly separated clusters, for ground truth cluster configurations. (**a**) McClain–Rao index, 2 clusters. (**b**) McClain–Rao index, 5 clusters

These initial results show that selecting the appropriate clustering quality index for high-dimensional data clustering is non-trivial and should be approached carefully.

Robustness to increasing data dimensionality would be highly beneficial in practical clustering quality estimation. Our results indicate that the dimensionality of the test data needs to be factored in any meaningful cross-index comparison, as it is otherwise possible for the reported results to be an artifact of the dimensionality of the test data. This is another reason why many surveys of clustering evaluation that were conducted in the past on low-dimensional data might not be relevant for practical clustering tasks, where data is usually high-dimensional.

This study was performed on synthetic data, as such an evaluation gives more control over the experimental protocol and the parameters of the context. In order for the findings to be truly relevant, a similar detailed study should be performed on real-world data as well, by using repeated sub-sampling of larger high-dimensional datasets. Of course, the availability of high-dimensional clustering benchmarks with ground truth and low violation of the cluster assumption is not ideal.

Additionally, since it was demonstrated on several occasions that a better handling of hub points may result in better overall clustering quality in many-dimensional problems [40, 83, 84], we intend to consider either extending the existing clustering quality indexes or proposing new ones that would incorporate this finding into account.

References

1. Achtert, E.: Hierarchical Subspace Clustering. Ludwig Maximilians Universitat, Munich (2007)
2. Aggarwal, C.: On high dimensional projected clustering of uncertain data streams. In: Proceedings of the 25th IEEE International Conference on Data Engineering (ICDE), pp. 1152–1154 (2009)

3. Aggarwal, C.C.: Re-designing distance functions and distance-based applications for high dimensional data. ACM Sigmod Rec. **30**(1), 13–18 (2001)
4. Aggarwal, C.C.: On k-anonymity and the curse of dimensionality. In: Proceedings of the 31st International Conference on Very Large Data Bases (VLDB), pp. 901–909 (2005)
5. Aggarwal, C.C.: On randomization, public information and the curse of dimensionality. In: Proceedings of the 23rd IEEE International Conference on Data Engineering (ICDE), pp. 136–145 (2007)
6. Bai, L., Liang, J., Dang, C., Cao, F.: A novel attribute weighting algorithm for clustering high-dimensional categorical data. Pattern Recogn. **44**(12), 2843–2861 (2011)
7. Baker, F.B., Hubert, L.J.: Measuring the power of hierarchical cluster analysis. J. Am. Stat. Assoc. **70**(349), 31–38 (1975)
8. Bellman, R.E.: Adaptive Control Processes – A Guided Tour. Princeton University Press, Princeton (1961)
9. Bengio, Y., Delalleau, O., Le Roux, N.: The curse of dimensionality for local kernel machines. Technical Report 1258, Departement d'Informatique et Recherche Operationnelle, Université de Montréal, Montreal, Canada (2005)
10. Bennett, K.P., Fayyad, U.M., Geiger, D.: Density-based indexing for approximate nearest-neighbor queries. In: Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 233–243 (1999)
11. Beyer, K., Goldstein, J., Ramakrishnan, R., Shaft, U.: When is "nearest neighbor" meaningful? In: Proceedings of the International Conference on Database Theory (ICDT), ACM, New York, NY, pp. 217–235 (1999)
12. Bohm, C., Kailing, K., Kriegel, H.P., Kroger, P.: Density connected clustering with local subspace preferences. In: Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM), pp. 27–34 (2004)
13. Bouguila, N., Almakadmeh, K., Boutemedjet, S.: A finite mixture model for simultaneous high-dimensional clustering, localized feature selection and outlier rejection. Expert Syst. Appl. **39**(7), 6641–6656 (2012)
14. Bouveyron, C., Brunet-Saumard, C.: Model-based clustering of high-dimensional data: a review. Comput. Stat. Data Anal. **71**, 52–78 (2014)
15. Buza, K., Nanopoulos, A., Schmidt-Thieme, L.: INSIGHT: efficient and effective instance selection for time-series classification. In: Proceedings of the 15th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Part II, pp. 149–160 (2011)
16. Caliński, T., Harabasz, J.: A dendrite method for cluster analysis. Commun. Stat. Simul. Comput. **3**(1), 1–27 (1974)
17. Carter, K., Raich, R., Hero, A.: On local intrinsic dimension estimation and its applications. IEEE Trans. Signal Process. **58**(2), 650–663 (2010)
18. Celebi, M.E. (ed.): Partitional Clustering Algorithms. Springer, Berlin (2014)
19. Chen, X., Ye, Y., Xu, X., Huang, J.Z.: A feature group weighting method for subspace clustering of high-dimensional data. Pattern Recogn. **45**(1), 434–446 (2012)
20. Chávez, E., Navarro, G.: A probabilistic spell for the curse of dimensionality. In: Algorithm Engineering and Experimentation, pp. 147–160. Springer, Berlin (2001)
21. Chávez, E., Navarro, G.: Probabilistic proximity search: fighting the curse of dimensionality in metric spaces. Inf. Process. Lett. **85**(1), 39–46 (2003)
22. Davies, D.L., Bouldin, D.W.: A cluster separation measure. IEEE Trans. Pattern Anal. Mach. Intell. **1**(2), 224–227 (1979)
23. Draper, B., Elliott, D., Hayes, J., Baek, K.: EM in high-dimensional spaces. IEEE Trans. Syst. Man Cybern. **35**(3), 571–577 (2005)
24. Draszawka, K., Szymanski, J.: External validation measures for nested clustering of text documents. In: Ryżko, D., Rybinski, H, Gawrysiak, P., Kryszkiewicz, M. (eds.) Emerging Intelligent Technologies in Industry. Studies in Computational Intelligence, vol. 369, pp. 207–225. Springer, Berlin (2011)
25. Dunn, J.C.: Well-separated clusters and optimal fuzzy partitions. J. Cybern. **4**(1), 95–104 (1974)

26. Durrant, R.J., Kabán, A.: When is 'nearest neighbour' meaningful: a converse theorem and implications. J. Complex. **25**(4), 385–397 (2009)
27. Evangelista, P.F., Embrechts, M.J., Szymanski, B.K.: Taming the curse of dimensionality in kernels and novelty detection. In: Applied Soft Computing Technologies: The Challenge of Complexity, pp. 425–438. Springer, Berlin (2006)
28. Fan, W., Bouguila, N., Ziou, D.: Unsupervised hybrid feature extraction selection for high-dimensional non-Gaussian data clustering with variational inference. IEEE Trans. Knowl. Data Eng. **25**(7), 1670–1685 (2013)
29. Farahmand, A.M., Szepesvári, C.: Manifold-adaptive dimension estimation. In: Proceedings of the 24th International Conference on Machine Learning (ICML), ACM, New York, NY, pp. 265–272 (2007)
30. Färber, I., Günnemann, S., Kriegel, H.P., Kröger, P., Müller, E., Schubert, E., Seidl, T., Zimek, A.: On using class-labels in evaluation of clusterings. In: MultiClust: 1st International Workshop on Discovering, Summarizing and Using Multiple Clusterings Held in Conjunction with KDD (2010)
31. Fern, X.Z., Brodley, C.E.: Random projection for high dimensional data clustering: a cluster ensemble approach. In: Proceedings of 20th International Conference on Machine learning (ICML), pp. 186–193 (2003)
32. Fern, X.Z., Brodley, C.E.: Cluster ensembles for high dimensional clustering: an empirical study. Technical Report CS06-30-02, Oregon State University (2004)
33. Fowlkes, E.B., Mallows, C.L.: A method for comparing two hierarchical clusterings. J. Am. Stat. Assoc. **78**(383), 553–569 (1983)
34. François, D., Wertz, V., Verleysen, M.: The concentration of fractional distances. IEEE Trans. Knowl. Data Eng. **19**(7), 873–886 (2007)
35. France, S., Carroll, D.: Is the distance compression effect overstated? Some theory and experimentation. In: Proceedings of the 6th International Conference on Machine Learning and Data Mining in Pattern Recognition (MLDM), pp. 280–294 (2009)
36. Frederix, G., Pauwels, E.J.: Shape-invariant cluster validity indices. In: Proceedings of the 4th Industrial Conference on Data Mining (ICDM), pp. 96–105 (2004)
37. Gupta, M.D., Huang, T.S.: Regularized maximum likelihood for intrinsic dimension estimation. Comput. Res. Rep. (2012). CoRR abs/1203.3483
38. Halkidi, M., Batistakis, Y., Vazirgiannis, M.: On clustering validation techniques. J. Intell. Inf. Syst. **17**, 107–145 (2001)
39. Hinneburg, A., Aggarwal, C., Keim, D.A.: What is the nearest neighbor in high dimensional spaces? In: Proceedings of the 26th International Conference on Very Large Data Bases (VLDB), pp. 506–515. Morgan Kaufmann, New York, NY (2000)
40. Hou, J., Nayak, R.: The heterogeneous cluster ensemble method using hubness for clustering text documents. In: Lin, X., Manolopoulos, Y., Srivastava, D., Huang, G. (eds.) Proceedings of the 14th International Conference on Web Information Systems Engineering (WISE). Lecture Notes in Computer Science, vol. 8180, pp. 102–110. Springer, Berlin (2013)
41. Houle, M.E.: The relevant-set correlation model for data clustering. J. Stat. Anal. Data Min. **1**(3), 157–176 (2008)
42. Houle, M.E., Kriegel, H.P., Kröger, P., Schubert, E., Zimek, A.: Can shared-neighbor distances defeat the curse of dimensionality? In: Proceedings of the 22nd International Conference on Scientific and Statistical Database Management (SSDBM), pp. 482–500 (2010)
43. Hsu, C.M., Chen, M.S.: On the design and applicability of distance functions in high-dimensional data space. IEEE Trans. Knowl. Data Eng. **21**(4), 523–536 (2009)
44. Hubert, L., Arabie, P.: Comparing partitions. J. Classif. **2**(1), 193–218 (1985)
45. Hubert, L.J., Levin, J.R.: A general statistical framework for assessing categorical clustering in free recall. Psychol. Bull. **83**(6), 1072 (1976)
46. Jarvis, R.A., Patrick, E.A.: Clustering using a similarity measure based on shared near neighbors. IEEE Trans. Comput. **22**, 1025–1034 (1973)
47. Jing, L., Ng, M., Huang, J.: An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data. IEEE Trans. Knowl. Data Eng. **19**(8), 1026–1041 (2007)

48. Kabán, A.: On the distance concentration awareness of certain data reduction techniques. Pattern Recogn. **44**(2), 265–277 (2011)
49. Kaban, A.: Non-parametric detection of meaningless distances in high dimensional data. Stat. Comput. **22**(2), 375–385 (2012)
50. Li, T., Ma, S., Ogihara, M.: Document clustering via adaptive subspace iteration. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 218–225 (2004)
51. Liu, B., Xia, Y., Yu, P.S.: Clustering through decision tree construction. In: Proceedings of the 26th ACM SIGMOD International Conference on Management of Data, pp. 20–29 (2000)
52. Liu, Y., Li, Z., Xiong, H., Gao, X., Wu, J.: Understanding of internal clustering validation measures. In: Proceedings of the 10th IEEE International Conference on Data Mining (ICDM), pp. 911–916 (2010)
53. Lu, Y., Wang, S., Li, S., Zhou, C.: Particle swarm optimizer for variable weighting in clustering high-dimensional data. Mach. Learn. **82**(1), 43–70 (2011)
54. Maulik, U., Bandyopadhyay, S.: Performance evaluation of some clustering algorithms and validity indices. IEEE Trans. Pattern Anal. Mach. Intell. **24**(12), 1650–1654 (2002)
55. Milligan, G.W.: A monte carlo study of thirty internal criterion measures for cluster analysis. Psychometrika **46**(2), 187–199 (1981)
56. Moise, G., Zimek, A., Kröger, P., Kriegel, H.P., Sander, J.: Subspace and projected clustering: experimental evaluation and analysis. Knowl. Inf. Syst. **21**(3), 299–326 (2009)
57. Müller, E., Günnemann, S., Assent, I., Seidl, T.: Evaluating clustering in subspace projections of high dimensional data. Proc. VLDB Endowment **2**(1), 1270–1281 (2009)
58. Naldi, M.C., Carvalho, A., Campello, R.J.: Cluster ensemble selection based on relative validity indexes. Data Min. Knowl. Disc. **27**(2), 259–289 (2013)
59. Ntoutsi, I., Zimek, A., Palpanas, T., Kröger, P., Kriegel, H.P.: Density-based projected clustering over high dimensional data streams. In: Proceedings of the 12th SIAM International Conference on Data Mining (SDM), pp. 987–998 (2012)
60. Pauwels, E.J., Frederix, G.: Cluster-based segmentation of natural scenes. In: Proceedings of the 7th IEEE International Conference on Computer Vision (ICCV), vol. 2, pp. 997–1002 (1999)
61. Pestov, V.: On the geometry of similarity search: Dimensionality curse and concentration of measure. Inf. Process. Lett. **73**(1–2), 47–51 (2000)
62. Pettis, K.W., Bailey, T.A., Jain, A.K., Dubes, R.C.: An intrinsic dimensionality estimator from near-neighbor information. IEEE Trans. Pattern Anal. Mach. Intell. **1**(1), 25–37 (1979)
63. Radovanović, M.: Representations and Metrics in High-Dimensional Data Mining. Izdavačka knjižarnica Zorana Stojanovića, Novi Sad, Serbia (2011)
64. Radovanović, M., Nanopoulos, A., Ivanović, M.: Nearest neighbors in high-dimensional data: The emergence and influence of hubs. In: Proceedings of the 26th International Conference on Machine Learning (ICML), pp. 865–872 (2009)
65. Radovanović, M., Nanopoulos, A., Ivanović, M.: Hubs in space: popular nearest neighbors in high-dimensional data. J. Mach. Learn. Res. **11**, 2487–2531 (2010)
66. Radovanović, M., Nanopoulos, A., Ivanović, M.: Time-series classification in many intrinsic dimensions. In: Proceedings of the 10th SIAM International Conference on Data Mining (SDM), pp. 677–688 (2010)
67. Radovanović, M., Nanopoulos, A., Ivanović, M.: Reverse nearest neighbors in unsupervised distance-based outlier detection. IEEE Trans. Knowl. Data Eng. **27**(5), 1369–1382 (2015)
68. Ratkowsky, D., Lance, G.: A criterion for determining the number of groups in a classification. Aust. Comput. J. **10**(3), 115–117 (1978)
69. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. **20**, 53–65 (1987)
70. Santos, J.M., Embrechts, M.: On the use of the adjusted rand index as a metric for evaluating supervised classification. In: Proceedings of the 19th International Conference on Artificial Neural Networks (ICANN), Part II. Lecture Notes in Computer Science, vol. 5769, pp. 175–184. Springer, Berlin (2009)

71. Schnitzer, D., Flexer, A., Schedl, M., Widmer, G.: Using mutual proximity to improve content-based audio similarity. In: Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR), pp. 79–84 (2011)
72. Schnitzer, D., Flexer, A., Schedl, M., Widmer, G.: Local and global scaling reduce hubs in space. J. Mach. Learn. Res. **13**(1), 2871–2902 (2012)
73. Serpen, G., Pathical, S.: Classification in high-dimensional feature spaces: Random subsample ensemble. In: Proceedings of the International Conference on Machine Learning and Applications (ICMLA), pp. 740–745 (2009)
74. Talwalkar, A., Kumar, S., Rowley, H.A.: Large-scale manifold learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–8 (2008)
75. Tomašev, N.: Hub miner: A hubness-aware machine learning library. (2014). http://dx.doi.org/10.5281/zenodo.12599
76. Tomašev, N.: Taming the empirical hubness risk in many dimensions. In: Proceedings of the 15th SIAM International Conference on Data Mining (SDM), pp. 1–9 (2015)
77. Tomašev, N., Mladenić, D.: Nearest neighbor voting in high dimensional data: learning from past occurrences. Comput. Sci. Inf. Syst. **9**(2), 691–712 (2012)
78. Tomašev, N., Mladenić, D.: Hub co-occurrence modeling for robust high-dimensional kNN classification. In: Proceedings of the European Conference on Machine Learning (ECML), pp. 643–659. Springer, Berlin (2013)
79. Tomašev, N., Mladenić, D.: Hubness-aware shared neighbor distances for high-dimensional k-nearest neighbor classification. Knowl. Inf. Syst. **39**(1), 89–122 (2013)
80. Tomašev, N., Leban, G., Mladenić, D.: Exploiting hubs for self-adaptive secondary re-ranking in bug report duplicate detection. In: Proceedings of the Conference on Information Technology Interfaces (ITI) (2013)
81. Tomašev, N., Radovanović, M., Mladenić, D., Ivanović, M.: Hubness-based fuzzy measures for high-dimensional k-nearest neighbor classification. Int. J. Mach. Learn. Cybern. **5**(3), 445–458 (2014)
82. Tomašev, N., Rupnik, J., Mladenić, D.: The role of hubs in cross-lingual supervised document retrieval. In: Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), pp. 185–196. Springer, Berlin (2013)
83. Tomašev, N., Radovanović, M., Mladenić, D., Ivanović, M.: The role of hubness in clustering high-dimensional data. IEEE Trans. Knowl. Data Eng. **26**(3), 739–751 (2014)
84. Tomašev, N., Radovanović, M., Mladenić, D., Ivanović, M.: Hubness-based clustering of high-dimensional data. In: Celebi, M.E. (ed.) Partitional Clustering Algorithms, pp. 353–386. Springer, Berlin (2015)
85. Vendramin, L., Campello, R.J.G.B., Hruschka, E.R.: Relative clustering validity criteria: a comparative overview. Stat. Anal. Data Min. **3**(4), 209–235 (2010)
86. Vendramin, L., Jaskowiak, P.A., Campello, R.J.: On the combination of relative clustering validity criteria. In: Proceedings of the 25th International Conference on Scientific and Statistical Database Management, p. 4 (2013)
87. Verleysen, M., Francois, D., Simon, G., Wertz, V.: On the effects of dimensionality on data analysis with neural networks. In: Proceedings of the 7th International Work-Conference on Artificial and Natural Neural Networks, Part II: Artificial Neural Nets Problem Solving Methods, pp. 105–112. Springer, Berlin (2003)
88. Vinh, N.X., Houle, M.E.: A set correlation model for partitional clustering. In: Zaki, M.J., Yu, J.X., Ravindran, B., Pudi, V. (eds.) Advances in Knowledge Discovery and Data Mining. Lecture Notes in Computer Science, vol. 6118, pp. 4–15. Springer, Berlin (2010)
89. Wu, S., Feng, X., Zhou, W.: Spectral clustering of high-dimensional data exploiting sparse representation vectors. Neurocomputing **135**, 229–239 (2014)
90. Yianilos, P.N.: Locally lifting the curse of dimensionality for nearest neighbor search. In: Proceedings of the Eleventh Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), pp. 361–370 (2000)

91. Yin, J., Fan, X., Chen, Y., Ren, J.: High-dimensional shared nearest neighbor clustering algorithm. In: Fuzzy Systems and Knowledge Discovery. Lecture Notes in Computer Science, vol. 3614, pp. 484–484. Springer, Berlin (2005)
92. Zhang, Z., Wang, J., Zha, H.: Adaptive manifold learning. IEEE Trans. Pattern Anal. Mach. Intell. **34**(2), 253–265 (2012)
93. Zheng, L., Huang, D.: Outlier detection and semi-supervised clustering algorithm based on shared nearest neighbors. Comput. Syst. Appl. **29**, 117–121 (2012)
94. Zimek, A., Schubert, E., Kriegel, H.P.: A survey on unsupervised outlier detection in high-dimensional numerical data. Stat. Anal. Data Min. **5**(5), 363–387 (2012)