Bauhaus-Universität Weimar
Faculty of Media
Degree Programme Computer Science and Media

Clusterability in Model Selection

Master's Thesis

Johannes Kiesel

1. Referee: Prof. Dr. Benno Stein
2. Referee: Prof. Dr. Volker Rodehorst

Submission date: May 19, 2014

Declaration

Unless otherwise indicated in the text or references, this thesis is entirely the product of my own scholarly work.

Weimar, May 19, 2014

. . . . . .
Johannes Kiesel

Ripples from a pebble thrown

—*Ian Anderson*

Abstract

Cluster analysis, the discovery of a sound categorization of entities based on their pairwise differences or distances, is applied in a great variety of different problem domains. The key to its versatility is its two-step procedure. In the first step, an abstract model of the differences among the entities is created. In the second step, generic clustering algorithms find categories in the abstract model. Despite its importance for the success of the cluster analysis, model formation is barely discussed in the literature as it depends heavily on the particular task and can only partially be generalized.

Clusterability, the assessment of the extent to which a model is structured in clusters, can guide model formation by model evaluation. Literature on clusterability introduces it as a sanity check of models: if the model is not clusterable, any categorization found by a clustering algorithm is questionable. However, we propose the use of clusterability in the related but distinct task of model selection: when different models of the same data are considered, clusterability analysis identifies which of them contains the most evident clusters and is thus preferable.

We motivate the use of clusterability as a generic tool for model selection in cluster analysis, detail different methods for assessing the clusterability of a model and analyze these methods on both synthetic and real world data. We lay emphasis on a theoretical justification of clusterability in the context of cluster analysis. We categorize, explain and compare 5 clusterability indices. We do so in a discussion of their rationales, with the use of 5 formalized index properties and in a demonstration on small example datasets. We provide empirical evidence that clusterability is indeed suitable for model selection but also show its limits in this regard: fallacious models, which fail to represent the original data, can be the most clusterable nevertheless.

In order to keep this thesis on an intuitively understandable level, we only consider models of moderate sizes and for which all pairwise dissimilarities are known. The application of some of the detailed methods on large-scale datasets might require additional considerations or may not be possible at all. Most of the clusterability indices that we detail were actually proposed in a different context and it might be possible to optimize them for the task of clusterability analysis. We therefore refrain from a premature efficiency analysis of the indices.

Contents

Chapter 1

Introduction

> The validation of clustering structures is the most difficult and frustrating part of cluster analysis. Without a strong effort in this direction, cluster analysis will remain a black art accessible only to those true believers who have experience and great courage.
>
> —Jain and Dubes, *Algorithms for Clustering Data*, 1988

More than 25 years ago, Jain and Dubes stated the need for research directed towards an understandable and well-founded cluster validation or, as we refer to it, *cluster evaluation*: given a *clustering* of a dataset into subsets, do these *clusters* correspond to evident groups within the dataset, or are they only based on meaningless variations in the data? *Clusterability* takes this question one step further: are there evident groups in the data at all?

The identification of groups of similar entities can be useful in many fields. When facing data too big to analyze every entity (e.g., in biotechnology), cluster analysis provides insights on a higher level. Connections between groups can be of interest, for example in social network analysis. Detecting entities that do *not* fit into groups is used for detecting outliers or malicious inputs. Another variant is the tracking of groups over time, like in trend detection.

Cluster analysis casts categorization problems into a mathematical setting, where the pairwise similarities of the entities (called *object*) represent the particular problem. In this setting, objects with high pairwise similarity among each other and high dissimilarity to the other objects form evident groups.

However, as hinted in the first paragraph, a clustering algorithm might also detect groups where there are none. Moreover, some popular clustering algorithms like K-Means (MacQueen, 1967, chap. 3.6) always find the user-specified number of groups (*clusters*), no matter the data. Therefore, *cluster evaluation* is necessary in order to make sure not to base follow-up data analysis on wrong assumptions. But what if evaluation suggests that the clustering

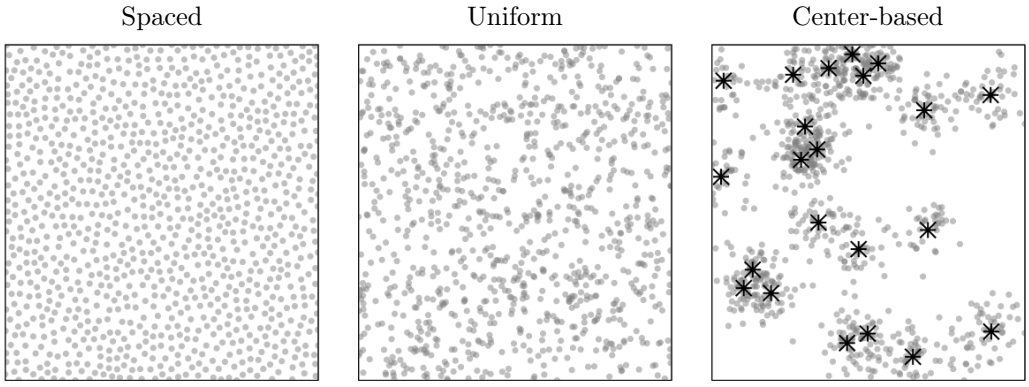

Figure 1.1: Data generated by different random processes ($\sim 1\,000$ points each). For the center-based data, the stars show the 22 uniformly sampled centers that were used in the creation of the data.

algorithm found a set of clusters (called *clustering*) that does not correspond to an evident categorization? It might be that, with respect to the data (1) the method for cluster evaluation is unsuited to the task at hand, (2) the chosen algorithm is, maybe due to unfortunate parameter settings, not able to find the groups or (3) there are no evident groups at all.

Clusterability tries to identify or rule out case (3): are there evident groups in the data? For example, consider clusters within the three datasets shown in figure 1.1: for the spaced data, the most meaningful "clustering" separates every object from all others; multiple clusterings of the uniform data seem plausible, but all are somewhat arbitrary; the center-based data contain evident clusters, although a strict separation is still not trivial. The latter data are clusterable, the others are not.

Clusterability analysis can guide a cluster analysis in different steps:

- choosing the clustering algorithm (e.g., the algorithm by Ostrovsky et al. (2006) takes advantage of clusterable data)
- choosing the mathematical model of the data (e.g., Vinay et al., 2006)
- whether to cluster at all (e.g., for document retrieval, El-Hamdouchi and Willett (1987) decide based on the clusterability of the data)

The focus of this thesis is the clusterability of data: can the existence of evident groups be detected independently of clusterings? We put clusterability into the context of cluster analysis in chapter 2. After that, chapter 3 discusses ideas for the quantification of clusterability. These ideas are employed and evaluated in a practical study of model selection on publicly available machine-learning datasets in chapter 4. We then give concluding remarks in chapter 5.

Chapter 2

Cluster Analysis

Cluster analysis is the task of discovering a sound categorization of a set of objects in terms of the similarities between these objects. According to Jain (2010), such a categorization can be used to (1) gain insight into the underlying structure of the data, (2) identify a natural classification, and (3) summarize or compress the data by representative and distinctive elements. A good introduction to the topic is provided by Jain and Dubes (1988).

However, the object similarities (or dissimilarities) are only in few cases intrinsic to the objects. Because of this, different sets of dissimilarities might be suitable and the appropriate selection is likely to have a great influence on the discovered categories. Despite the importance of this selection, it is less studied in the general cluster analysis literature as it heavily depends on the particular problem at hand (Hastie et al., 2009, chap. 14.3). Clusterability analysis can provide guidance in this situation.

In order to allow for a mathematical treatment of cluster analysis, formal definitions of the terms introduced so far are necessary.

Definition 2.1 (Object x, dataset X, ground-truth) Each *object* stands for 1 entity, like a person, a document or a node in a computer network. A *dataset* is denoted by X and contains $|X|$ objects. If the referenced dataset is clear from the context, we usually use n for the number of objects instead of $|X|$. We use $\mathbf{x}_i$ to denote the i-th object within X.

If there exists a natural categorization of the entities in the dataset, we refer to such a categorization as the ground-truth of the dataset. For simplicity, we assume that every dataset has exactly 1 ground-truth associated with it.

In the literature, the objects are also called inputs, features, patterns or independent variables. Cluster analysis requires the choice of a model which represents the dataset through pairwise dissimilarities.

Definition 2.2 (Dissimilarity function ψ) A function $\psi : X \times X \mapsto \mathbb{R}^+$ that is defined over the objects of a dataset X. The mandatory properties are $\forall_{\mathbf{x} \in X} (\psi(\mathbf{x}, \mathbf{x}) = 0)$ and $\forall_{\mathbf{x}, \mathbf{x}' \in X} (\psi(\mathbf{x}, \mathbf{x}') = \psi(\mathbf{x}', \mathbf{x}))$ (symmetry). An object $\mathbf{x}$ is said to be more similar (less dissimilar) to $\mathbf{x}'$ than to $\mathbf{x}''$ if $\psi(\mathbf{x}, \mathbf{x}') < \psi(\mathbf{x}, \mathbf{x}'')$.

Note that, although it is not always the case in practice, we assume for simplicity that the dissimilarities of all pairs of objects $(\mathbf{x}, \mathbf{x}') \in X \times X$ are known.

Definition 2.3 (Model X) A representation $\mathbf{X}$ of objects of a dataset that defines the pairwise dissimilarities of the objects in terms of some dissimilarity function $\psi_{\mathbf{X}}$. The model is said to be *meaningful* with respect to the dataset if the ground-truth is reflected in the dissimilarities of $\psi_{\mathbf{X}}$.

The *vector model* is a special case where objects are represented as vectors in some space $\mathbb{X}$ and the dissimilarities are defined by a general dissimilarity function $\psi : \mathbb{X} \times \mathbb{X} \mapsto \mathbb{R}^+$. We employ vector models with $\mathbb{X} = \mathbb{R}^2$ and ψ defined as the Euclidean distance for illustrations like in figure 1.1.

Definition 2.4 (Cluster analysis, clustering $\mathcal{C}$, cluster C) A *cluster* C is a non-empty subset of a dataset X. A *clustering* $\mathcal{C}$ of X is a partition of the objects of X into a set of clusters. In this thesis, we require clusterings to be *complete* $(\forall_{\mathbf{x} \in X} (\exists_{C \in \mathcal{C}} (\mathbf{x} \in C)))$ and *strict* $(\forall_{C \neq C' \in \mathcal{C}} (C \cap C' = \{\}))$. The number of clusters in $\mathcal{C}$ is given by $|\mathcal{C}|$ and is referred to as k when $\mathcal{C}$ is clear from the context. In the same situation, C_i denotes the i-th cluster in $\mathcal{C}$. The task of finding a suitable clustering of a dataset is called *cluster analysis*.

The base component in cluster analysis is the clustering algorithm. A clustering algorithm searches the space of possible clusterings of the dataset for an optimal one. This search is based on the dissimilarities defined by a model of the dataset.

Definition 2.5 (Clustering algorithm γ) A method for choosing a clustering of a dataset X based on a model of X.

In order to investigate the behaviour of cluster analysis methods in a fully controllable environment, *synthetic* models drawn from *model distributions* are often used in theoretical analysis. The introduction of randomness through distributions has the advantage over directly fixed models that the gained insights are less likely to depend on specific and perhaps unnoticed circumstances in the models.

Definition 2.6 (Model distribution $\mathscr{X}$) A model distribution $\mathscr{X}$ allows to randomly sample models $\mathbf{X}$ of some model representation (e.g, vector models), denoted by $\mathbf{X} \leftarrow \mathscr{X}$. It either specifies fixed values or distributions for all variables that define models of this representation. These are the number of objects, the dissimilarities between them and maybe representation specific variables. The distributions are not necessarily independent of each other. In all matters, models from model distributions are treated as if they stem from a "phantom" dataset with the corresponding number of objects.

Models and clusterings are evaluated in order to make sure that they actually contain or correspond to meaningful structure. Not every model contains clustered structure or is suited for clustering. Furthermore, a clustering algorithm only searches for a "best" clustering, which might still not be "good." Methods which are able to quantify the meaningfulness of a clustering seem thus necessary. This task is known as cluster evaluation, while the model evaluation problem of distinguishing models that contain clustered structure from those that do not is called clusterability analysis.

Definition 2.7 (Cluster evaluation) The problem of assessing the accordance of a clustering with the ground-truth or models of a dataset. If the ground-truth is not known, a model replaces it as the reference. If the model is meaningful, one can assume that both approaches lead to similar results.

Definition 2.8 (Clusterability analysis) The problem of assessing the (relative) extent to which groups are evident in a model.

Next, we discuss cluster evaluation (section 2.1), clustering algorithms (section 2.2) and model evaluation and selection (section 2.3). We show how clusterability fits into the context of model evaluation and is related to the other parts of cluster analysis. We choose this order as we believe that it facilitates the understanding of model selection if the reader is already familiar with the consecutive steps. Since the task of model creation depends on the concrete problem at hand, a discussion of it lies outside the scope of this thesis.

The implementation of cluster analysis in the form of the consecutive steps of model creation, model evaluation, model clustering and cluster evaluation is common in clustering theory. However, it should be mentioned that other approaches exist, too. For example, Law et al. (2004) propose an algorithm that iteratively clusters the data (clustering) while it evaluates for each dimension in the object space if it contains structure (model evaluation) and removes the dimensions if appropriate (model selection). Nevertheless, we discuss the parts of cluster analysis as separated steps for theoretical simplicity.

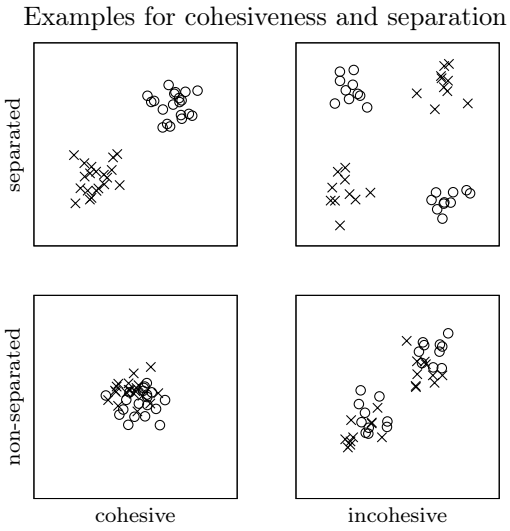

Figure 2.1: The two clusterings on the left hand side are cohesive and the two at the top are separated. The assignment of objects to the different clusters is depicted by the different symbols (circles and crosses).

2.1 Cluster Evaluation

There are two common methodologies for assessing the quality of a clustering: internal and external cluster evaluation. *Internal* evaluation is based on the accordance of the clusters in the clustering with the dissimilarities in a model of the dataset. *External* evaluation compares the clustering with the ground-truth of the dataset, which is usually also represented in the form of a clustering. Note that internal evaluation assumes that the model represents the dataset while external evaluation relies on the knowledge of the ground-truth. We focus in this thesis on the practical case in which the ground-truth is unknown and thus on internal evaluation.

The clusters of a clustering that is in accordance with the dissimilarities in a model are cohesive and separated from objects of other clusters. Since mathematical definitions of cohesiveness and separation differ, a variety of evaluation indices ν for assessing the agreement of $\mathcal{C}$ and $\mathbf{X}$ exist. We only give a vague definition of both concepts that encompasses all the different notions we are aware of. Intuitive examples that illustrate cohesiveness and separation are shown in figure 2.1.

Definition 2.9 (Cohesiveness and separation) A cluster C is *cohesive* if all of the $\mathbf{x} \in C$ form a consistent whole. C is *separated* if the addition of other objects $\mathbf{x} \in X \backslash C$ to it has a significant negative impact on its cohesiveness.

A common idea of "consistent whole" is that there exists a model distribution that is likely to have generated the objects. For example, objects in a cluster of a vector model can be approximately normal distributed for some center and standard deviation. However, in order to measure the likelihood under specific model distributions a set of possible "cluster-like" distributions has to be specified. This choice both allows and forces considerations on problem-specific properties of a cohesive cluster. Since likeliness comparisons between distribution families are not trivial, usually only one family is employed (e.g., only normal distributions but with different centers and standard deviations).

In general, the more of the objects $\mathbf{x} \in X \backslash C$ have a negative impact on the cohesiveness of the cluster C and the higher the impact, the more separated is C. When the addition of any single $\mathbf{x}$ has already a significant impact, C is very separated. Common notions of separation consider either the closest object $\mathbf{x} \in X \backslash C$ or all the $\mathbf{x} \in C'$ for the closest cluster $C' \neq C$.

The notions of cohesiveness and separation of clusters are also helpful for the assessment of the clusterability of a model. A clusterable model contains cohesive and separated clusters of objects. A more detailed look into this topic follows in chapter 3, when we discuss how clusterable structure can be identified.

Internal cluster evaluation employs an *evaluation index* ν, which quantifies the agreement between a clustering $\mathcal{C}$ and a model $\mathbf{X}$. We define internal evaluation indices and discuss their properties in section 2.1.1. All evaluation indices can be used to rank clusterings by their agreement with a model. Section 2.1.2 discusses the special case of evaluation indices which allow for a direct interpretation of the agreement. We discuss the use of the ground-truth for the evaluation of clusterings and models in section 2.1.3. Although not relevant in practice, such methods can provide a baseline for model selection purposes and can thus be used in the evaluation of clusterability indices as in chapter 4.

2.1.1 Internal Evaluation Indices

While studies on the theoretical properties of clusterability indices are rare, the properties of evaluation indices are well-studied in the literature. Since clusterability and cluster evaluation are related, properties of evaluation indices can also be relevant for clusterability indices. Specific sets of properties are discussed, for example, by Wright (1973), Puzicha et al. (2000) or Ackerman and Ben-David (2008). Our discussion focuses on properties that are indeed relevant with respect to clusterability. Based on this section, we analyze specific clusterability properties in section 3.1.

Definition 2.10 (Internal evaluation index ν) A permutation invariant and consistent mapping from clusterings and models of some common dataset to real valued scores. A clustering $\mathcal{C}$ is said to be in better accordance with a model $\mathbf{X}$ than $\mathcal{C}'$ if it achieves a higher score, that is $\nu(\mathcal{C}, \mathbf{X}) > \nu(\mathcal{C}', \mathbf{X})$.

We use definitions of permutation invariance and consistency adapted from Puzicha et al. (2000).[1] Permutation invariance is necessary as the ordering of objects and clusters is usually seen as arbitrary in cluster analysis. If the dataset is ordered in some respect, for example timeline data, the order is instead used as one of the factors which determine the dissimilarities of objects. Consistency assures that an unambiguously more cohesive and more separated clustering does not achieve a worse score.

Definition 2.11 (Permutation invariant evaluation index) The score of the evaluation index ν does not depend on the order of objects or clusters.

More formally, let π_r be a permutation of the set $\{1, \ldots, r\}$ and $\pi_r(i)$ its i-th element. Furthermore, let $\mathbf{x}'$ be the objects of a model $\mathbf{X}'$, $\mathbf{x}$ be those of $\mathbf{X}$, C' be the clusters of a clustering $\mathcal{C}'$ and C be those of $\mathcal{C}$. Then ν is permutation invariant *with respect to objects* if, for all $\mathbf{X}, \mathbf{X}'$ and $\mathcal{C}$

$$\exists_{\pi_n}(\forall_{i,j\in\{1,\ldots,n\}}(\psi_{\mathbf{X}'}(\mathbf{x}'_{\pi_n(i)}, \mathbf{x}'_{\pi_n(j)}) = \psi_{\mathbf{X}}(\mathbf{x}_i, \mathbf{x}_j))) \quad \Rightarrow \quad \nu(\mathcal{C}, \mathbf{X}') = \nu(\mathcal{C}, \mathbf{X})$$

It is permutation invariant *with respect to clusters* if, for all $\mathbf{X}, \mathcal{C}$ and $\mathcal{C}'$

$$\exists_{\pi_k}(\forall_{\mathbf{x}\in X, i\in\{1,\ldots,k\}}(\mathbf{x} \in C'_{\pi_k(i)} \Leftrightarrow \mathbf{x} \in C_i)) \quad \Rightarrow \quad \nu(\mathcal{C}', \mathbf{X}) = \nu(\mathcal{C}, \mathbf{X})$$

And it is permutation invariant if it is permutation invariant with respect to both clusters and objects.

Definition 2.12 (Consistent evaluation index) Consistency guarantees that an unambiguously more cohesive and separated clustering achieves at least an equal evaluation score. A clustering $\mathcal{C}'$ is at least as cohesive/separated as another clustering $\mathcal{C}$ of a model $\mathbf{X}$ if there exists a combined permutation of objects and clusters, $\forall_{i\in\{1,\ldots,n\}, j\in\{1,\ldots,k\}}(\mathbf{x}_i \in C'_j \Leftrightarrow \mathbf{x}_{\pi_n(i)} \in C_{\pi_k(j)})$, such that all dissimilarities within clusters in $\mathcal{C}'$ are smaller or equal to their corresponding dissimilarities in $\mathcal{C}$ and all corresponding dissimilarities between clusters are larger or equal.

The formal relationship is actually easier to understand if differences in two models are used and the clustering $\mathcal{C}$ kept constant. The result is equivalent.

[1] Consistency is actually called monotonicity by Puzicha et al. (2000). However, we deem consistency—used by Ackerman and Ben-David (2008)—to be more intuitive.

Let $\mathbf{X}$ be any model and let the model $\mathbf{X}_{\Delta\mathcal{C}}$ be defined by

$$\psi_{\mathbf{X}_{\Delta\mathcal{C}}}(\mathbf{x}_i, \mathbf{x}_j) = \begin{cases} \psi_{\mathbf{X}}(\mathbf{x}_i, \mathbf{x}_j) + d_{i,j} & \text{if} \quad \exists_{C \in \mathcal{C}}(\mathbf{x}_i \in C \wedge \mathbf{x}_j \in C) \\ \psi_{\mathbf{X}}(\mathbf{x}_i, \mathbf{x}_j) - d_{i,j} & \text{else} \end{cases}$$

Where $\forall_{i,j \in \{1,\ldots,n\}}(d_{i,j} = d_{j,i} \wedge d_{j,i} \geqslant 0 \wedge d_{i,i} = 0)$. Then ν is consistent if, for all $\mathbf{X}$, $\mathcal{C}$ and values $d_{i,j}$ restricted as above,

$$\nu(\mathcal{C}, \mathbf{X}) \geqslant \nu(\mathcal{C}, \mathbf{X}_{\Delta\mathcal{C}})$$

We continue with further properties that allow to reason about the behaviour of evaluation indices and can be a criterion for choosing one. Especially when the task at hand does not suggest a specific magnitude of within- and between-cluster dissimilarities, scale invariant indices (Ackerman and Ben-David, 2008) are preferable. In some cases, it can be advantageous to choose an evaluation index that is robust with respect to outliers (Puzicha et al., 2000). For example, measurement errors are a common cause of outliers.

Definition 2.13 (Scale invariant evaluation index) A uniform scaling of all dissimilarities in the model has no effect on the score of an evaluation index ν.

Formally, for all models $\mathbf{X}, \mathbf{X}'$ and all clusterings $\mathcal{C}$,

$$\exists_{a \in \mathbb{R}} \left(\forall_{\mathbf{x}_i, \mathbf{x}_j \in X} \left(\psi_{\mathbf{X}}(\mathbf{x}_i, \mathbf{x}_j) = \frac{\psi_{\mathbf{X}'}(\mathbf{x}_i, \mathbf{x}_j)}{a} \right) \right) \Rightarrow \nu(\mathcal{C}, \mathbf{X}) = \nu(\mathcal{C}, \mathbf{X}')$$

Note that, because of permutation invariance, the objects are actually not required to be in the same order in $\mathbf{X}$ and $\mathbf{X}'$.

Definition 2.14 (Strongly/weakly robust evaluation index) The effect of single dissimilarity changes in $\mathbf{X}$ on the score of the evaluation index ν is limited and tends towards 0 for sufficiently large datasets. *Weak* robustness assures a limited effect when one single dissimilarity is changed, while *strong* robustness assures the same for changes to dissimilarities of one object.

For any model $\mathbf{X}$ we define models $\mathbf{X}_{\Delta r,d}$ by

$$\psi_{\mathbf{X}_{\Delta r,d}}(\mathbf{x}_i, \mathbf{x}_j) = \begin{cases} \psi_{\mathbf{X}}(\mathbf{x}_i, \mathbf{x}_j) + d_j & \text{if} \quad \mathbf{x}_i = \mathbf{x}_r \\ \psi_{\mathbf{X}}(\mathbf{x}_i, \mathbf{x}_j) + d_i & \text{if} \quad \mathbf{x}_j = \mathbf{x}_r \\ \psi_{\mathbf{X}}(\mathbf{x}_i, \mathbf{x}_j) & \text{else} \end{cases}$$

Where $\forall_{i\in\{1,\ldots,n\}}(|d_i| \leqslant d)$, $d_r = 0$ and additionally, in the case of only weakly robust evaluation indices, $\exists_j(\forall_{i\in\{1,\ldots,n\}}(i \neq j \Rightarrow d_i = 0))$. Then, for all $\mathbf{X}$, clusterings $\mathcal{C}$, d_i as restricted above and positive real values d and e,

$$\exists_{n_0\in\mathbb{Z}^+}(n \geqslant n_0 \Rightarrow (|\nu(\mathcal{C},\mathbf{X}) - \nu(\mathcal{C},\mathbf{X}_{\Delta r,d})| < e))$$

While all internal evaluation indices can be used to decide on a "best" of a set of clusterings, not all of them evaluate if a clustering fits the model in an absolute manner.

Definition 2.15 (Absolute evaluation index) The score of a clustering and a model can be interpreted independently of the number and sizes of the clusters, the number of objects, the average dissimilarity of objects, and additional model parameters (e.g., the dimensionality for vector models).

For some evaluation indices ν and models $\mathbf{X}^0$ drawn from a specific model distribution $\mathscr{X}^0$, the optimum score $\max_{\mathcal{C}}(\nu(\mathcal{C},\mathbf{X}^0))$ of a random model $\mathbf{X}^0$ has an analytically known expected value and variance. This allows for an interpretation of any score with respect to the typical score of models drawn from $\mathscr{X}^0$. Usually, the model distribution chosen for $\mathscr{X}^0$ is unlikely to generate models that contain any clusters. Therefore, if a score $\nu(\mathcal{C},\mathbf{X})$ is relatively high—as measured by the known expected value and variance—then it is likely that the clustering $\mathcal{C}$ corresponds to evident groups in $\mathbf{X}$. This is because such a score is unlikely to be reachable at all for a model without any clusters. If the complete distribution of the optimum score is known, a probability of the score under $\mathscr{X}^0$ can be calculated, which can then be used as an absolute evaluation index (Jain and Dubes, 1988).

Definition 2.16 (Distribution normalized evaluation index) The score can be interpreted with respect to the known expected value and variance of the optimal score for models from some specific model distribution.

More formally, let $\mathrm{opt}_\nu(\mathbf{X}) = \max_{\mathcal{C}}(\nu(\mathcal{C},\mathbf{X}))$ be the optimum achievable score for clusterings of model $\mathbf{X}$ with respect to an evaluation index ν. Then, ν is called distribution normalized or $\mathscr{X}^0$-normalized if—for models drawn from some fixed model distribution $\mathscr{X}^0$—the expected value (e_ν) and standard deviation (s_ν) of the optimal score are known.

$$e_\nu = \mathrm{E}_{\mathbf{X}^0\leftarrow\mathscr{X}^0}\left[\mathrm{opt}_\nu(\mathbf{X}^0)\right] \qquad s_\nu = \sqrt{\mathrm{E}_{\mathbf{X}^0\leftarrow\mathscr{X}^0}\left[(\mathrm{opt}_\nu(\mathbf{X}^0) - e_\nu)^2\right]}$$

However, depending on the task at hand, different model distributions $\mathscr{X}^0$ may be suitable. For example, if cluster evaluation is performed for clusterings of 2-dimensional vector models with coordinates in the range $[-1, 1]$, the $\mathscr{X}^0$ that samples objects uniformly from the same 2-dimensional subspace is reasonable. Moreover, in most cases the optimum score depends on the number of objects in the models. Therefore, when interpreting a score of a model $\mathbf{X}$, it is necessary to make sure that e_ν and s_ν correspond to $\mathbf{X}^0 \leftarrow \mathscr{X}^0$ with the same number of objects. Nevertheless, even if no analytically known values exist, it may still be feasible to approximate e_ν and s_ν. This is further discussed in section 2.1.2. We want to point out that this kind of normalization differs from being normalized with respect to a random clustering, which is a property of some evaluation indices for external cluster evaluation (Jain and Dubes, 1988).

The same idea of distribution normalization with respect to cluster-less models exists for measuring the clusterability of a model $\mathbf{X}$. In this line of thought, for any distribution normalized evaluation index ν, all clusterable $\mathbf{X}$ and reasonable $\mathcal{C}$, we assume that $\nu(\mathcal{C}, \mathbf{X}) > e_\nu + a \cdot s_\nu$ for some relatively large a. On the other hand, a value close to e_ν does not imply that $\mathbf{X}$ is not clusterable, as a different clustering may achieve a high score.

Finally, we want to point out that all evaluation indices correspond to some specific view of how cohesiveness and separation should be measured and weighted. Therefore, the evaluation index should be chosen with care and the concrete task in mind. As noted by Lange et al. (2004), some evaluation indices are biased towards clusters of certain shapes or sizes. For example, many evaluation indices, like the Gap-statistic (Tibshirani et al., 2001) and the scatter separability criterion (cf. Dy and Brodley, 2000), use the notion that *all* pairwise dissimilarities in a cohesive cluster should be small. In terms of a vector model, this corresponds to spherical clusters. On the other hand, some evaluation indices only measure either cohesiveness or separation and are therefore unsuited for the comparison of $\mathcal{C}$ with different number of clusters. These indices are said to have a refinement or coarsening preference respectively (Ackerman and Ben-David, 2008). For example the compactness index, $\nu(\mathcal{C}, \mathbf{X}) = \sum_{\mathbf{x},\mathbf{x}' \in X} \psi_{\mathbf{X}}(\mathbf{x}, \mathbf{x}') - \sum_{i=1}^{k} \sum_{\mathbf{x},\mathbf{x}' \in C_i} \psi_{\mathbf{X}}(\mathbf{x}, \mathbf{x}')$, increases even when a cohesive cluster is split. Note that evaluation indices with a refinement or coarsening preference are never absolute as the interpretation of the score varies with the number of clusters.

2.1.2 Absolute Internal Evaluation

As the task of absolute internal cluster evaluation is to judge whether the structure implied by a clustering $\mathcal{C}$ can be justified by the dissimilarities in a model $\mathbf{X}$, the employed evaluation index ν has to be absolute (def. 2.15). If ν

is not, the score might not only depend on the agreement between $\mathcal{C}$ and $\mathbf{X}$, but also, for example, on the number of clusters, the sizes of the clusters or the number of dimensions for a vector model. Therefore, an interpretation of the score requires normalization—which essentially makes ν absolute. Some methods for normalization with respect to a model distribution are introduced below.

Absolute evaluation is especially interesting for our discussion as the normalization of indices can be directly adopted for clusterability analysis. Similar to cluster evaluation, most of the methods for the assessment of model clusterability are suited for a comparison of different models, but not for an absolute assessment. For some indices, the comparable models are even further restricted, for example to contain the same number of objects. Fortunately, the methods we detail below for interpreting evaluation index scores apply to clusterability scores, as well. Furthermore, because clusterability does not rely on clusterings, the application of these methods is more straightforward.

Moreover, absolute cluster evaluation is related to clusterability, as specific results of one task imply results in the other one. If the model $\mathbf{X}$ is not clusterable and therefore contains no clustered structure, $\nu(\mathcal{C}, \mathbf{X})$ will be low for all reasonable clusterings $\mathcal{C}$. On the other hand, if there exists a $\mathcal{C}$ for which $\nu(\mathcal{C}, \mathbf{X})$ is high, $\mathbf{X}$ has to be clusterable. However, the implications do not hold in the other direction. Note that the requirement of an absolute index allows for the specification of meaningful values for "low" and "high."

One method for interpreting the score of a clustering $\mathcal{C}$ is to use the probability that, for models $\mathbf{X}^0$ from a cluster-less model distribution $\mathscr{X}^0$, there exists no clustering of $\mathbf{X}^0$ that achieves a score as high as $\nu(\mathcal{C}, \mathbf{X})$. By using this idea, it is possible to create an absolute variant ν_a of ν:

$$\nu_a(\mathcal{C}, \mathbf{X}) = \Pr_{\mathbf{X}^0 \leftarrow \mathscr{X}^0}\left[\max_{\mathcal{C}'}\left(\nu\left(\mathcal{C}', \mathbf{X}^0\right)\right) < \nu(\mathcal{C}, \mathbf{X})\right] \tag{2.1}$$

However, distributions of evaluation indices are rarely known, especially since they often depend on the number of objects, the number of clusters, the sizes of the clusters, and in some cases also model specific parameters. For example, while there has been some work on employing the distribution of the log-likelihood ratio, ν_{lr}, McLachlan (1987) notes that the corresponding theorem by Wilks (1938) can not be applied in cluster analysis due to a violated regularity condition. Therefore, it is in general necessary to use an empirical approximation by a sampling distribution instead (Jain and Dubes, 1988, chap. 4.4). One idea for such an empirical method is to use a sufficiently large number (r) of random models $\mathbf{X}_i^0 \leftarrow \mathscr{X}^0$:

$$\nu_{\hat{a}}(\mathcal{C}, \mathbf{X}) = \frac{1}{r} \cdot \sum_{i=1}^{r} \mathbf{1}\left(\nu(\gamma(\mathbf{X}_i^0), \mathbf{X}_i^0) < \nu(\mathcal{C}, \mathbf{X})\right) \tag{2.2}$$

Where $\mathbf{1}(e)$ is 1 if e is true and 0 otherwise. Since the calculation of the optimal clustering (cf. equation 2.1) is in general not feasible, $\nu_{\hat{a}}$ uses a clustering algorithm γ for further approximation. Unfortunately, due to this approximation, $\nu_{\hat{a}}$ tends to be above the correct probability. In order to limit this effect, it is necessary to evaluate multiple clusterings and thus to use different γ. This increases the chances of finding an optimal or close-to-optimal clustering.

A similar approach is used in the Gap-statistic by Tibshirani et al. (2001) in order to determine the number of clusters in a model. However, since this task requires only the relative evaluation of clusterings with different number of clusters but on the same model, they do not need to calculate absolute probabilities. Instead, they subtract the expected value from the score: $\nu_{\text{gap}}(\mathcal{C}, \mathbf{X}) = \nu(\mathcal{C}, \mathbf{X}) - \frac{1}{r} \cdot \sum_{i=1}^{r} \nu(\gamma^{|\mathcal{C}|}(\mathbf{X}_i^0), \mathbf{X}_i^0)$ where γ^l only considers clusterings with l clusters. Then, Tibshirani et al. choose the number of clusters by considering $\nu_{\text{gap}}(\gamma^l(\mathbf{X}), \mathbf{X})$ for different l and the variance of the empirical distribution over $\mathbf{X}_i^0$. Note that they only consider the expected value and the variance of the empirical distribution instead of relying on distribution percentiles like equation 2.2. This is advantageous, as reliable estimations of the distribution tails usually require a much larger r.

Although the Gap-statistic is not absolute, it is similar in spirit to a second method for creating an absolute evaluation index. This method calculates the standard score (or z-score) of the observed value:

$$\nu_z(\mathcal{C}, \mathbf{X}) = \frac{\nu(\mathcal{C}, \mathbf{X}) - \hat{\mu}(\nu(\gamma(\mathbf{X}_1^0), \mathbf{X}_1^0), \cdots, \nu(\gamma(\mathbf{X}_r^0), \mathbf{X}_r^0))}{\hat{\sigma}(\nu(\gamma(\mathbf{X}_1^0), \mathbf{X}_1^0), \cdots, \nu(\gamma(\mathbf{X}_r^0), \mathbf{X}_r^0))} \tag{2.3}$$

Where $\hat{\mu}$ and $\hat{\sigma}$ calculate the empirical average and standard deviation respectively. Under the assumption that the $\nu(\gamma(\mathbf{X}_i^0), \mathbf{X}_i^0)$ are normally distributed, ν_z can be directly translated into a probability under the hypothesis that $\mathbf{X}$ is sampled from $\mathscr{X}^0$. Again, $\mathscr{X}^0$ should be a cluster-less model distribution and thus the probability should be low for a good fit of $\mathcal{C}$ and $\mathbf{X}$. Like equation 2.2, this evaluation index depends on the employed clustering algorithm and the use of different γ might be necessary in order to remove some bias from it.

Experiment 2.1 ($\bar{L}_1^{\hat{\sigma}}$-distance normalization) In order to facilitate the understanding of absolute index interpretation, we give an example related to clusterability. Given a vector model, we want to decide whether it was sampled from a normal distribution $\mathscr{X}_{\mathcal{N}}^1$ (not clusterable) or a uniform mixture of 2 normal distributions, $\mathscr{X}_{\mathcal{N}}^2$ (clusterable, standard deviation $\sigma = 1$).

For this experiment, the centers of the distributions in $\mathscr{X}_{\mathcal{N}}^2$ have a distance of 5 and lie on the x_1 axis. 1 000 models $\mathbf{X}_i^2$ with n objects in m dimensions are sampled from $\mathscr{X}_{\mathcal{N}}^2$. The empirical averages and standard deviations for each dimension $(\hat{\mu}(\mathbf{X}^2), \hat{\sigma}(\mathbf{X}^2))$ of the $1\,000 \cdot n$ objects $\mathbf{x} \in \mathbf{X}^2$ are used as parameters

for $\mathscr{X}^1_{\mathcal{N}}$. In the same manner, 1 000 models $\mathbf{X}^1_i$ with n objects are sampled from $\mathscr{X}^1_{\mathcal{N}}$. One example with 64 objects in 2 dimensions is shown in figure 2.2 for each case (a,b). For the inner evaluation index (cf. ν in equation 2.2), the average variance-weighted L_1-distance[2] to the center is used:

$$\bar{L}_1^{\hat{\sigma}}(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{inv}(\hat{\sigma}(\mathbf{X}))^T \cdot (\mathbf{x}_i - \hat{\mu}(\mathbf{X}))$$

Where $\mathrm{inv}(\mathbf{v})^T$ is the row-vector $(1/v_1, \ldots, 1/v_m)$ and v_i is i-th coordinate of $\mathbf{v}$. Since we do not want to measure differences in variances but in structure, we use the weighted average in order to remove the effect of the different variances in the different dimensions. Note that no clustering is employed in this experiment as we want to measure clusterability instead.

The absolute evaluation index is then given by the empirical complementary cumulative distribution function $\mathrm{ccdf}_1(d) = \Pr_i\left[\bar{L}_1^{\hat{\sigma}}(\mathbf{X}^1_i) \geqslant d\right]$. We show this function together with $\mathrm{cdf}_2(d) = \Pr_i\left[\bar{L}_1^{\hat{\sigma}}(\mathbf{X}^2_i) \leqslant d\right]$ for 64 objects in 2 dimensions in figure 2.2 (c). The figure also shows the decision threshold: if $\bar{L}_1^{\hat{\sigma}}(\mathbf{X})$ is smaller than this threshold we decide for $\mathscr{X}^1_{\mathcal{N}}$ and else for $\mathscr{X}^2_{\mathcal{N}}$.

The misclassification risk of this decision, that is, the fraction of cases in which a wrong decision is made, is shown in figure 2.2 (d). It can be seen that the risk grows with dimensionality. This is not surprising, as only one of the dimensions contains the information that allows to distinguish the cases. Since $\bar{L}_1^{\hat{\sigma}}$ calculates a sum over all dimensions, the one important dimension has less and less impact on the result. For example, consider increasing the distance of the Gaussian centers in the 256 objects in 64 dimensions case from 5 to 100. Although this makes the clusters extremely salient with regard to one dimension, the misclassification risk drops still merely from 0.26 to 0.17. We also want to point out that more objects can lessen this effect and that the risk does not change much if n/m is kept constant. We revisit $\bar{L}_1^{\hat{\sigma}}$ in section 3.2.

2.1.3 External Evaluation

In *external* cluster evaluation, a clustering $\mathcal{C}$ is compared to the ground-truth $\mathcal{C}_{\mathrm{gt}}$ of the dataset (def. 2.1) by means of an evaluation index ν^e. In the evaluation, $\mathcal{C}$ achieves the highest score $\nu^e(\mathcal{C}, \mathcal{C}_{\mathrm{gt}})$ if it is identical to $\mathcal{C}_{\mathrm{gt}}$ and a lower score the more it differs from $\mathcal{C}_{\mathrm{gt}}$. Note that, since knowledge of the ground-truth eliminates the need of cluster analysis, external evaluation is usually limited to showcases of clustering algorithms or models. Similarly, we use external knowledge to showcase clusterability indices in chapter 4. A good discussion of external indices is provided by Hubert and Arabie (1985).

[2] Also called "Manhattan distance"

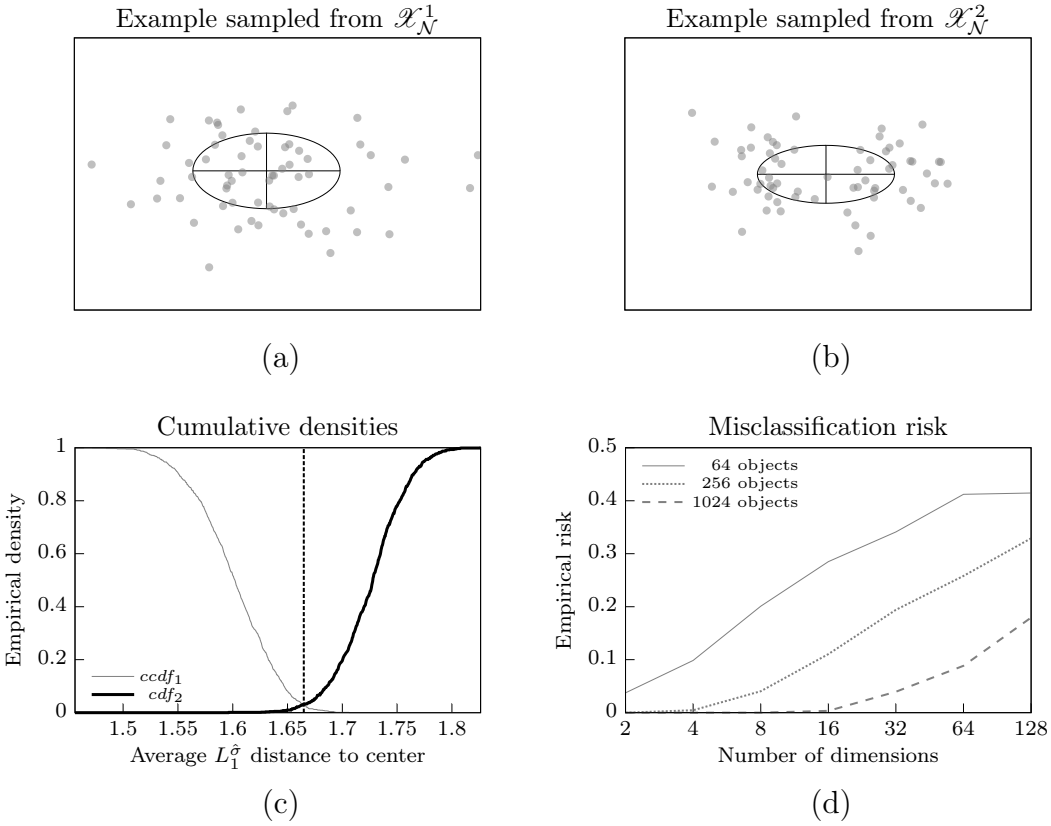

Figure 2.2: 64 objects sampled from (*a*) 1 Gaussian or (*b*) 2 Gaussians with approximately the same overall average and standard deviation. The figures also show the centers ($\hat{\mu}(\mathbf{X})$) as cross junctions and the doubled standard deviations ($2 \cdot \hat{\sigma}(\mathbf{X})$) as cross-beam lengths. (*c*) Empirical (complementary) cumulative density functions of $\bar{L}_1^{\hat{\sigma}}$ for 64 objects in 2 dimensions (ccdf_1, cdf_2) and the decision threshold for classification as $\mathscr{X}_{\mathcal{N}}^1$ or $\mathscr{X}_{\mathcal{N}}^2$ as vertical line. (*d*) Misclassification risks by the number of objects and dimensions.

In the same manner, the meaningfulness of a model $\mathbf{X}$ with respect to its dataset can be evaluated through the use of external knowledge about a natural clustering of the dataset, $\mathcal{C}_{\mathrm{gt}}$. For the moment, let us define an accordant pair of model and clustering as one for which the value of an (absolute) evaluation index ν is above a threshold t. Then the model is meaningful if $\nu(\mathcal{C}_{\mathrm{gt}}, \mathbf{X}) > t$. This is related to clusterability, which asks if there is *any* evident partition of objects into clusters for the model, $\approx \exists_{\mathcal{C}}(\nu(\mathcal{C}, \mathbf{X}) > t)$.[3] These ideas can be connected by the assumption that $\nu(\mathcal{C}_{\mathrm{gt}}, \mathbf{X}) \approx \max_{\mathcal{C}}(\nu(\mathcal{C}, \mathbf{X}))$, which heavily depends on the notion of a "good" cluster as defined by ν.

[3]We use $\approx$ as we actually exclude the clusterings that contain only 1 cluster or only clusters of size 1 from this consideration.

In a second methodology for external model evaluation, the internal index is replaced by an external index ν^e and a clustering algorithm γ. Therefore, the model is referred to as meaningful if $\nu^e(\gamma(\mathbf{X}), \mathcal{C}_{\text{gt}}) > t$. Since clustering algorithms are optimization algorithms that do not necessarily converge to the global optimum, it makes sense to lower the dependency on γ. In order to achieve this, one can use a suitable set of clustering algorithms and $\exists_\gamma(\nu^e(\gamma(\mathbf{X}), \mathcal{C}_{\text{gt}}) > t)$ instead. This uses the assumption that, for meaningful models, $\mathcal{C}_{\text{gt}}$ is similar to the optimal clustering with respect to γ. We define optimum clusterings in definition 2.17 below.

2.2 Clustering Algorithms

A clustering algorithm γ maps a model $\mathbf{X}$ of a dataset X to a clustering $\mathcal{C}$ of X (cf. def. 2.5). Most cluster algorithms can be adjusted to specific tasks through algorithm parameters. Some algorithms also depend on randomness. In our notation, γ includes a specific parameter setting and, if required, a source of random but fixed inputs. Thus, the mapping of γ is deterministic even if the "base" algorithm contains randomness. In accordance with Jain and Dubes (1988, chap. 3.3), the number of possible clusterings for a dataset with n objects is:

$$\text{number of clusterings}(n) = \sum_{k=1}^{n} \frac{1}{k!} \cdot \sum_{i=1}^{k} (-1)^{k-i} \cdot \binom{k}{i} \cdot i^n$$

For as few as 15 objects, this already gives more than 10^9 possible clusterings.

Since the number of clusterings increases exponentially with n, an exhaustive scan of all clusterings for the optimal $\mathcal{C}_{\text{opt}}$ is infeasible even for small datasets. Therefore, all clustering algorithms are search algorithms that use certain assumptions to make the search feasible. Most clustering algorithms search by iteratively improving a clustering (its "state") until a local optimum is found. If the algorithm considers changing its state from one clustering to another one, these 2 clusterings are called neighbors. The preference of the algorithm when selecting a clustering from the neighbors of the current state can be described as a partial order relation.

Definition 2.17 (Clustering relation $\leqslant_\gamma$, (local) optimum clustering) The clustering relation is a partial order on clusterings $\mathcal{C}$. Therefore, for all $\mathcal{C}, \mathcal{C}'$ and $\mathcal{C}''$: $\mathcal{C} \leqslant_\gamma \mathcal{C}$ (reflexivity), $(\mathcal{C} \leqslant_\gamma \mathcal{C}' \wedge \mathcal{C}' \leqslant_\gamma \mathcal{C}) \Rightarrow (\mathcal{C} = \mathcal{C}')$ (antisymmetry) and $(\mathcal{C} \leqslant_\gamma \mathcal{C}' \wedge \mathcal{C}' \leqslant_\gamma \mathcal{C}'') \Rightarrow (\mathcal{C} \leqslant_\gamma \mathcal{C}'')$ (transitivity). If $\mathcal{C} \leqslant_\gamma \mathcal{C}'$, then the clustering algorithm prefers $\mathcal{C}'$ over $\mathcal{C}$.

An *optimum clustering* $\mathcal{C}_{\text{opt}}$ is defined by $\forall_{\mathcal{C}}(\mathcal{C} \leqslant_\gamma \mathcal{C}_{\text{opt}} \vee \neg(\mathcal{C}_{\text{opt}} \leqslant_\gamma \mathcal{C}))$. That is, it is either preferred over any other clustering or incomparable to it.

A *local optimum clustering* $\hat{\mathcal{C}}_{\text{opt}}$ is defined in terms of neighboring clusterings: $\forall_{\mathcal{C}}(\mathcal{C} \leqslant_\gamma \hat{\mathcal{C}}_{\text{opt}} \vee \neg(\text{neighbor}_\gamma(\hat{\mathcal{C}}_{\text{opt}}, \mathcal{C})))$,

While some clustering relations can be interpreted as maximizing the likelihood of the model under assumptions of model distributions (Kamvar et al., 2002), others rely on a "variety of ad hoc rules and tricks" (Jain et al., 2004) and lack a statistical justification. A discussion of clustering algorithm taxonomies is provided by Jain et al. (2004).

Some of the clustering relations $\leqslant_\gamma$ that clustering algorithms γ use to choose a clustering can be defined by means of evaluation indices ν. With the use of ν, $\leqslant_\nu$ can be defined as $(\mathcal{C} \leqslant_\nu \mathcal{C}') \approx (\nu(\mathcal{C}, \mathbf{X}) \leqslant \nu(\mathcal{C}', \mathbf{X}))$.[4] However, because clustering algorithms restrict the search for an optimum clustering, the clustering relation does not have to be total. For example, since K-Means (MacQueen, 1967, chap. 3.6) only considers $\mathcal{C}$ with some fixed number of clusters, $\leqslant_{\text{K-Means}}$ does not need to be able to compare $\mathcal{C}$ with different number of clusters. In fact, while the employed clustering relation would theoretically be able to do so, it has a strong preference for $\mathcal{C}$ that contain a higher number of clusters and would therefore be unsuited for this task.

Since a variety of clustering algorithms without statistical justification exist, it is difficult to relate clustering algorithms in general to the concept of clusterability. Although there are already few results in this regard, it is still an open field for future work. Special clustering algorithms γ can be designed to rely on the result that a model $\mathbf{X}$ contains a highly clustered structure. For example, knowledge on the clusterability of a model can allow to find a clustering close to the global optimum clustering $\mathcal{C}_{\text{opt}}$ with high probability in a feasible amount of time. Furthermore, it can be possible to provide error bounds in the form of $\Pr[\nu(\gamma(\mathbf{X}), \mathbf{X}) < e \cdot \nu(\mathcal{C}_{\text{opt}}, \mathbf{X})] < b$ with $e, b \in [0, 1]$ depending on the clusterability of $\mathbf{X}$ (Ostrovsky et al., 2006). Unfortunately, the time required to check if $\mathbf{X}$ is clusterable with respect to the definition by Ostrovsky et al. grows exponentially with the number of objects. It is therefore infeasible to employ it in practice.

A possible implication of a clusterable model to a clustering algorithm might be an increased stability. Consider the case that the objects or their dissimilarities are drawn identically and independently from some distribution with multiple clusters to form the model $\mathbf{X}$. We now consider randomly sampled subsets of $\mathbf{X}$, yielding the sub-models $\mathbf{X}_1, \ldots, \mathbf{X}_r$. A clustering algorithm γ is said to be stable with respect to $\mathbf{X}$ if the clusterings $\gamma(\mathbf{X}_1), \ldots, \gamma(\mathbf{X}_r)$

[4] This is not exactly equal since some additional method is required to fulfill antisymmetry (cf. def. 2.17) when $\nu(\mathcal{C}, \mathbf{X}) = \nu(\mathcal{C}', \mathbf{X})$ and $\mathcal{C} \neq \mathcal{C}'$

are similar (Lange et al., 2004). Note that the clusterings can not be compared directly, since the models contain different objects. Lange et al. propose the use of classification methods for an indirect comparison. They show that stability can be useful in the selection of clustering algorithm parameters,[5] as parameters that are suitable for a model are assumed to lead to more stable results. A statistical justification for this assumption is provided by Shamir and Tishby (2007). By using this line of argument, we assume that a clusterable model has a high difference in stability between suitable and unsuitable parameter settings. However, in order to limit the scope of our discussion, we leave an analysis of this conjecture to future publications.

2.3 Model Evaluation

Cluster analysis is performed in two steps: modeling and clustering. In the last sections we have discussed clustering, which is a general method of identifying cohesive and separated clusters in a model that defines pairwise dissimilarities between single objects. Modeling, also called the measurement problem (Wright, 1973), is the task of assigning a dissimilarity to each pair of objects such that clustering can take place. Although the modeling of a dataset is more important for the success of a cluster analysis than the clustering, less literature exists on this topic since it relies to a large extent on the problem at hand and can thus only scarcely be examined in a general setting (Hastie et al., 2009, chap. 14.3).

The model can be defined using various different mathematical structures. The structure should be chosen depending on the problem at hand. Examples are (1) graphs where the vertices represent objects $\mathbf{x}$ and the edges are weighted by their dissimilarities;[6] (2) matrices where an entry in the i-th row and j-th column is the dissimilarity between $\mathbf{x}_i$ and $\mathbf{x}_j$; or (3) a vector space $\mathbb{X}$ in which the $\mathbf{x}$ are represented by vectors and their dissimilarity is calculated by means of a function $\psi : \mathbb{X} \times \mathbb{X} \mapsto \mathbb{R}^+$. In practice, such *vector models* are often defined by measurements of certain attributes on the objects. Each attribute corresponds to one dimensions of $\mathbb{X}$ and the coordinate of each $\mathbf{x}$ is determined by the measured value therein. Most publications that consider the vector model require that all attributes have a ratio domain, that is $\mathbb{X} = \mathbb{R}^m$ for some $m \in \mathbb{N}$ and a semantical scaling is performed by multiplying attribute values with a scalar.[7] We also assume ratio attributes in this thesis. This simplifies

[5] Their discussion is actually limited to the target number of clusters but can be expanded to other parameters, as well

[6] The graph model is usually employed when not all pairwise dissimilarities are known, which we excluded from our discussion for simplicity.

[7] For example, there is a "twice as hot" on the Kelvin scale, but not on the Celsius scale.

our considerations as identical object vectors and therefore dissimilarities of 0 are unlikely for measured ratio attributes. As can be seen in section 3.2, some measures of clusterability are undefined when dissimilarities of 0 are possible.

An alternative to the use of a dissimilarity function is the similarity function, which—although quite common—will not be considered in this thesis. Opposite to dissimilarity functions, similarity functions use 0 for lowest and $+\infty$ for maximum similarity. Although conversions are easily possible, they should be chosen carefully and with the concrete situation in mind.

This thesis continues with an analysis on desired model properties and how a model can be changed in order to improve it. Section 2.3.1 addresses diverse aspects of model quality, which includes clusterability. After that, section 2.3.2 discusses methods that modify models in different ways. The generality of these methods allows for a domain-independent discussion of them. Nevertheless, it is the problem at hand that has to justify their application.

2.3.1 Aspects of Model Quality

When given the choice between multiple models of a dataset, which one should be chosen? The model should be meaningful with respect to the dataset and contain as little noise and outlier objects as possible. Models with a low complexity are preferable. In order to receive a cohesive and separated clustering, the model should be clusterable.

A model of a dataset has to be motivated by the dataset and the problem at hand. What is a meaningful measure of dissimilarity? In case of a vector model, what are reasonable attributes? How should they be weighted? Which measurements on the objects are important and which might mislead clustering algorithms? Formalizing answers to these questions and choosing the model appropriately is crucial for a successful cluster analysis. After all, a clustering has only a meaning when the model is also meaningful to the dataset.

In practice, models $\mathbf{X}$ often contain noise and outlier objects that can have a negative impact on the search for clusterings. Although the removal of noise and outliers is a classical pre-processing task in statistics and thus part of model selection, there are also some clustering algorithms that cover it.

Definition 2.18 (Outlier) An object that has, likely due to some error, a high dissimilarity to practically every other object in the dataset. A typical origin of outliers are measurements errors.

If $\mathbf{X}$ contains outliers, the clustering algorithm might be misled in its local search when attempting to group the outliers with other objects. An example of an algorithm that identifies outliers is the possibilistic clustering algorithm

by Krishnapuram and Keller (1993). The function assigns each object a probability that it belongs to a certain cluster. If this probability is near zero for all clusters,[8] the object is seen as an outlier.

Definition 2.19 (Noise object) Although it might be relatively close to even multiple clusters, adding the object to any of them would reduce their cohesiveness significantly.

The term "noise" stems from statistics where it refers to variation in the data that is not explained by a mathematical formalization. When clusters are seen as objects that are sampled from a common distribution, noise objects are thus unlikely with respect to each of the distributions in the clustering.

The declaration of some objects as noise, which then demands a separate treatment, is often accepted in order to keep the clustering cohesive. In detail, clustering algorithms might be misled by noise objects to merge actually cohesive clusters in order to incorporate the noise objects. Different strategies for a treatment of noise objects are listed in section 2.3.2. An example of a clustering algorithm that identifies noise objects and ignores them in the determination of the clusters is DBScan (Ester et al., 1996).

Experiment 2.2 (Minimum spanning forests and noise) We want to illustrate one of the effects of noise objects: clusters that are separated can lose their separation under noise.

A model $\mathbf{X}$ that contains 200 objects from $\mathbb{R}^2$ is sampled from a mixture of two multivariate normal distributions. The centers of the distributions have a distance of 5 standard deviations (fig. 2.3, left). In order to demonstrate the separation of the two clusters, we add the shortest edges of the minimum spanning tree of $\mathbf{X}$ such that the clusters are still not linked (center). Since these are nearly all edges, the clusters have a good though not perfect separation.

We then add 200 "noise objects" uniformly over the window spanned by the objects in $\mathbf{X}$. The use of uniformly distributed noise is common in the literature (e.g., Aggarwal, 2001b or Houle et al., 2010). After this step, the clusters have lost with respect to their separation (fig. 2.3, right).

This experiment demonstrates the "chaining" effect, which occurs especially for clustering algorithms that define clusters by trees over the edge graph (as in fig. 2.3). Because of this effect, noise objects can have a large influence on such clustering algorithms. Therefore, these algorithms often include a method for noise reduction (like DBScan, Ester et al., 1996).

[8] The authors use an extended definition of a clustering that allows violation of strictness (objects can be part of multiple clusters) and completeness (objects can be part of no cluster).

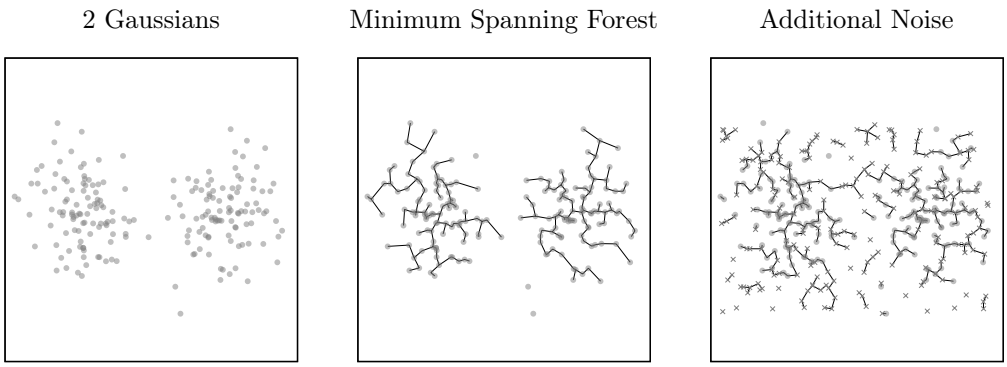

Figure 2.3: (*Left*) 200 objects generated from two Gaussians. (*Center*) 195 shortest edges of the minimum spanning tree. (*Right*) Additional 200 uniformly distributed "noise objects" marked as "x" and 365 shortest edges. The formerly separated clusters are now linked by noise.

Simple models are preferable in order to minimize uninformative statistical variation and reduce runtime and storage requirements of clustering algorithms and models. In the case of vector models, we say that a model becomes more complex with a higher number of attributes, as this weakens the constraints on the dissimilarities between objects. This corresponds to an increase in the degrees of freedom of the objects. However, sometimes it is possible to project the objects into spaces of lower dimensionality without much change in the dissimilarities due to dependencies among the attributes. For example, if both attributes of a 2-dimensional model are directly linearly related, all objects lie on a (1-dimensional) straight line. This refers to the concept of *intrinsic dimensionality* (e.g., cf. Levina and Bickel, 2004).

Some definitions of intrinsic dimensionality require only the dissimilarities of object pairs, which allows to apply the discussion of vector model complexity to models in general. Although it is unclear how to interpret this complexity, there is some evidence that it can be seen as a measure of the clusterability of a model (cf. section 3.2.3). An explanation of this evidence is that clusterable models also contain dependencies among the attributes. We show that, for some measures of intrinsic dimensionality, a low dimensionality is estimated because more objects lie at a "border" in clusterable models.

High-dimensional models bear the risk that the clustered structure that exists in some attributes is overshadowed by the statistical variation in uninformative ones (cf. experiment 2.1 or Houle et al., 2010). Sometimes, this effect can be diminished by an appropriate choice of the dissimilarity measure (Aggarwal, 2001a; Aggarwal et al., 2001). However, Francois et al. (2007) show that such a choice is not trivial and the optimal selection depends on the

model. We want to note that, although there are proofs of the inexpressiveness of dissimilarities for certain high-dimensional object distributions (Beyer et al., 1999), these do not apply to clustered data (Bennett et al., 1999).

The selection of a more clusterable model facilitates the detection of evident groups by the clustering algorithm. The corresponding assessment of models is called clusterability (cf. def. 2.8). According to Jain and Dubes (1988), *clusterability*[9] is "the problem of deciding whether data exhibit a predisposition to cluster into natural groups without identifying the groups themselves." Ackerman and Ben-David (2009) define it simply as "a measure of clustered structure." Opposite to Jain and Dubes, we follow Ackerman and Ben-David to allow the identification of groups during clusterability analysis. Other publications relate the question *if* there are clusters to *how many* there are (e.g., Havens et al., 2009). In this sense, if one has determined that a model contains multiple clusters (only 1 or n clusters), one might also reason that it is (not) clusterable. As we detail in chapter 3, not every proposed measure of clusterability follows this reasoning. Moreover, the question if a model that contains $n-1$ clusters should still be seen as clusterable remains open. In our terms, a model is clusterable if it has a structure of mutually separated and cohesive parts (cf. def. 3.1, page 26). Clusterability is the focus of chapter 3, where it is discussed in more detail. With regard to model quality, we want to point out that noise—either in the form of noise objects or irrelevant attributes—is intuitively also likely to have a negative impact on the clusterability of models.

2.3.2 Model Transformations

In order to improve a model, several general methods exist that change the dissimilarities of object pairs or remove objects from the model. One well-studied subtopic is the embedding of a model into a low-dimensional vector space. The following list of model transformations is incomplete and is also missing transformations that are relevant to specific problems only.

The embedding of a model in a vector space can be useful to gain further insight on the data, reduce model complexity or visualize it. Several methods for embedding have been proposed in the literature. Some minimize the differences in all pairwise dissimilarities (e.g, Sammon, 1969) while others focus on the preservation of the dissimilarities between close objects (e.g, Roweis and Saul, 2000). Some of these methods, however, require also a vector model as input and can thus only be used for dimensionality and complexity reduction. The transformation of a general model into a vector model, on the other hand, can provide insight with regard to model complexity. Such a transformation corresponds to modeling the data via a set of "latent" attributes, which

[9] Clusterability is called "clustering tendency" by Jain and Dubes.

have to be motivated with regard to the problem at hand. For example, topic models try to identify latent topics in text document collections and model each document as a mixture of these (e.g., Blei et al., 2003). As can be seen, this approach requires a careful distinction between the meaning of latent attributes (here topics) and clusters (e.g., categories of news articles). Finally, an embedding into low-dimensional spaces can be helpful for the visualization of the data if the embedding error is still acceptable.

A special form of dimensionality reduction is the orthogonal projection, which simply removes some attributes from the model. Such methods usually score either (1) single dimensions or (2) sets of dimensions by their usefulness for the problem at hand. We also want to note that clustering algorithms exist that score attributes (e.g., Law et al., 2004) or find a separate subspace of relevant attributes for each cluster (e.g., Agrawal et al., 1998 or Aggarwal and Yu, 2000).

If a score has been calculated for each dimension, dimensions are removed that achieve only a small score. Sometimes the eigenvectors of the model are used instead of the original attributes (Aggarwal, 2001b). Aggarwal notes that the eigenvalues corresponding to the eigenvectors are not a good measure for usefulness in the general case. The eigenvalues correlate with the amount of variation along the eigenvectors. However, also data distributed uniformly along the eigenvector can have a high variance. He therefore suggests to use instead the so-called coherence probability, a measure of correlation between the dimension and the object coordinates. This is similar to a clusterability analysis with respect to each dimension and choosing the most clusterable attributes.

If the measure of attribute usefulness, on the other hand, evaluates sets of attributes, then different sets are evaluated and the one with the highest score is chosen. As the cost of an exhaustive search grows exponentially with the number of dimensions, search algorithms are employed in order to find locally optimal sets. For example, Dash et al. (1997) propose to select dimensions such that the spread of the object dissimilarities is high. They argue that in the clustered case dissimilarities are either low (same cluster) or high (different clusters) but rarely in between.

Another type of model transformations adapts the dissimilarities to additional knowledge. We especially want to point out the method of pairwise instance-based constraints proposed by Wagstaff and Cardie (2000). This method allows the incorporation of constraints in the form of "these two objects must be in the same cluster" and "these two can not be in the same cluster." It is believed in the cluster analysis community that these constraints provide a promising approach to represent domain knowledge and user input in a human-understandable and general way (Jain, 2010). In the last years, several

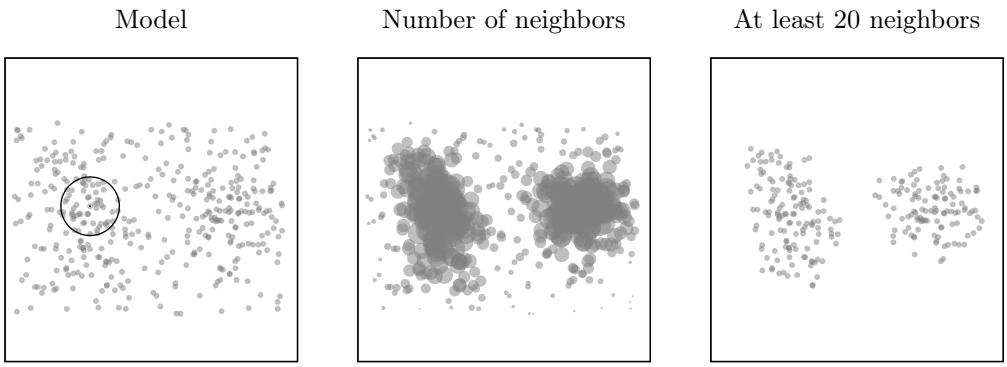

Figure 2.4: (*Left*) The noisy model of experiment 2.2. Also shows the neighborhood for one object as a circle. (*Center*) Objects with radii proportional to the number of neighbors. (*Right*) The 247 objects with at least 20 neighbors. Among these are 166 of the 200 original objects. Thus, a good part of the unwanted noise has been removed without removing too many non-noise objects.

methods have been developed to adapt clustering algorithms and models to such constraints. Some of the model adaption techniques work on general models (e.g., Klein et al., 2002), while others require vector models (e.g., Bar-Hillel et al., 2005). Some methods do not enforce the constraints, but instead adjust the dissimilarities to them. However, it should be noted that even correct constraints with respect to the ground-truth can lead to a decreased clustering performance when they are incoherent within the metric space (Davidson et al., 2006). The effect of such misleading constraints on the clusterability of the model remains an interesting open question for future work.

A kind of model transformation which is not directly related to object dissimilarities is the removal of outliers and noise objects before the use of the clustering algorithm. Depending on the concrete problem, the removed objects can then after the clustering either be added to the nearest cluster, be assigned to a cluster on their own or simply be ignored in subsequent analysis. A well-known approach for the removal of noise objects is the one incorporated into the DBScan algorithm (Ester et al., 1996): ignore objects that have less than a certain number of "neighbors," which are other objects with a dissimilarity below a specific threshold. The successful application of this approach to the noisy data of experiment 2.2 is shown in figure 2.4. A similar approach by Steinbach et al. (2003) uses the number of common objects in the two lists of the r nearest neighbors of two objects as a meta-similarity function. With regard to this similarity, noise objects become outliers as they tend to share only few neighbors with other objects. While this method automatically adjusts itself to the dissimilarities in the model, the choice of r remains.

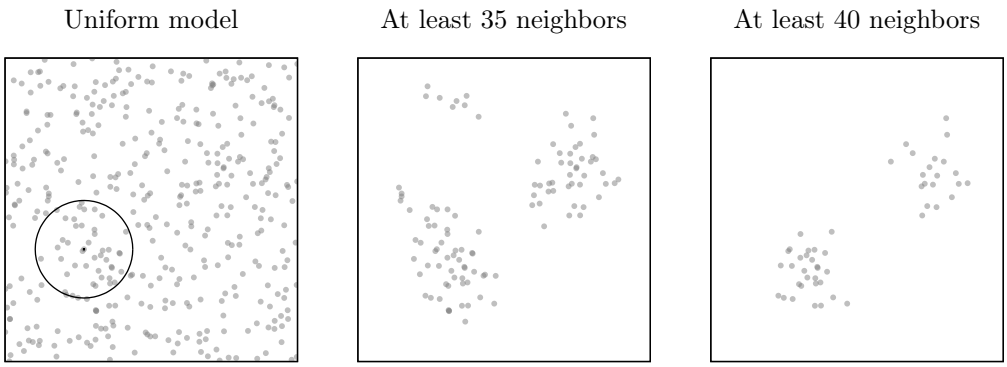

Figure 2.5: (*Left*) 400 objects from a uniform distribution and size of neighborhoods (circle). (*Center*) The 105 objects with at least 35 neighbors. (*Right*) The 50 objects with at least 40 neighbors. The central and right plots show evident clusters that are not present in the original data.

However, we want to point out that, with respect to clusterability, a not carefully motivated model transformation can lead to seemingly good but incorrect results. Figure 2.5 shows the removal of "noise objects" in the case of uniform data (compare to fig. 2.4). As shown, "clustered structure" that randomly exists in the model is highlighted by the removal and forms separated and cohesive clusters. Although the "cleaned" model is indeed clusterable, it is meaningless with respect to the original (uniform) data. This shows the importance of a mindful restriction of model transformations in model selection.

Summary Cluster analysis can be divided into 2 steps: modeling and clustering. For both steps, a unique best solution is not known in general. Instead, it is often recommended to consider multiple suitable models/clusterings, then evaluate which is "best," and proceed with this. Clusterability can be used to decide on a "best" model, as a more clusterable model facilitates the successive clustering. Clusterability is also related to the well-established cluster evaluation and measures from the latter can be adapted for the former. We continue with a discussion of such clusterability indices. Some are based on evaluation indices, some on other ideas. The relation of the clusterability and the meaningfulness of models is empirically explored in chapter 4 with the use of external knowledge (cf. section 2.1.3).

Chapter 3

Measuring Clusterability

How can the clusterability of a model be measured? The question of clusterability is a question of model structure: is the model structured as separated object clusters and how evident is this structure?

Definition 3.1 (Clusterable model) A model which has a dominant structure of mutually separated parts that are cohesive groups of objects.

For Stein and Niggemann (1999), "structure defines the organization of parts as dominated by the general character of the whole." The general character of a clusterable model is that of separated object clusters. For illustration, consider again the introductory examples in figure 1.1. The uniform model contains no structure: the objects are independent of each other. The spaced model contains no *clustered* structure: the parts, although separated, are the single objects and not clusters. Indeed, Jain and Dubes (1988) pose the task of clusterability analysis as the classification of a model as either random, regularly spaced or clustered. We want to point out that this is not the way structure is intuitively perceived by humans. Instead, we see the "random structure" in the uniform model and the lack of distinguishing features in the spaced model (Köppen, 2000).

The notion of an overall separated structure of cohesive parts is also found in many internal evaluation indices. The well-known index by Dunn (1974) is the quotient of the minimum dissimilarity between clusters divided by the maximum dissimilarity within any cluster (cf. section 3.2.1). Thus, models in which the separation within clusters is dominated by the separation between clusters achieve a high score. Similarly, the expected density index (Stein et al., 2003) compares the dissimilarities within the clusters with an expectation based on the overall dissimilarities of the model. In a good clustering, the cohesiveness of the parts dominates the cohesiveness of the whole.

We continue with an experiment that further illustrates how cohesiveness and separation are related to clusterability. After that, section 3.1 discusses clusterability indices on a general level and introduces properties that characterize them. Then, various different approaches to the identification and quantification of clustered structure are the topic of section 3.2. The analysis of some representative clusterability indices uses the properties defined in section 3.1. For a comparison of these indices, we then also apply them to the models of the following experiment.

Experiment 3.1 (Cohesiveness and separation in clusterability) For this experiment, we use models of 180 objects sampled from a uniform distribution over an $\mathbb{R}^2$ square with an area of 1. This model distribution is denoted by $\mathscr{X}_{\boxplus}^{0}$. Apart from such models $\mathbf{X}_i^0$, we also generate models $\mathbf{X}_i^{1,s}$, $\mathbf{X}_i^{4,s}$, $\mathbf{X}_i^{9,s}$ and $\mathbf{X}_i^{n,s}$ from the distributions $\mathscr{X}_{\boxplus}^{1,s}$, $\mathscr{X}_{\boxplus}^{4,s}$, $\mathscr{X}_{\boxplus}^{9,s}$ and $\mathscr{X}_{\boxplus}^{n,s}$ with distribution parameter s. Examples from each distribution and the corresponding constraints are shown in figure 3.1. The parameter s controls the size of the squares from which objects are sampled for $\mathscr{X}_{\boxplus}^{1,s}$, $\mathscr{X}_{\boxplus}^{4,s}$ and $\mathscr{X}_{\boxplus}^{9,s}$ and a sequential inhibition algorithm prevents the sampling of "close" objects for models from $\mathscr{X}_{\boxplus}^{n,s}$. For $\mathscr{X}_{\boxplus}^{n,s}$, we sample 2 000 objects uniformly. Then we iteratively select one of the objects with the fewest other objects within a distance $d_s = 2 \cdot r_s$ and remove these other objects. This assures that the disks of radius r_s centered at the objects do not overlap (cf. fig. 3.1 bottom right). Additionally, we assure that all such disks lie within the sampling region of $\mathscr{X}_{\boxplus}^{0}$. When only selected objects remain, we randomly select 180 of these. If fewer than 180 remain, we restart the process. The choice of $r_s = \sqrt{(2 \cdot s - s^2)/(180 \cdot \pi)}$ is such that the "constrained area"—the combined area of the disks in the case of $\mathscr{X}_{\boxplus}^{n,s}$—has an equal size for the 4 parametric distributions.

The parameter s controls the deviation from the purley uniform model distribution. If $s = 0$, all distributions are identical to $\mathscr{X}_{\boxplus}^{0}$. For higher values of s, the sampled models tend to get more cohesive ($\mathscr{X}_{\boxplus}^{1,s}$), more separated ($\mathscr{X}_{\boxplus}^{n,s}$) or both ($\mathscr{X}_{\boxplus}^{4,s}$, $\mathscr{X}_{\boxplus}^{9,s}$).

For an assessment of model structure we use $|\mathrm{mst}_t|$, which denotes the number of edges in the minimum spanning tree of the model that are smaller or equal to a threshold t. Intuitively, cohesive clusters result in a large $|\mathrm{mst}_t|$ for small t ("small scale structure") while separated clusters lead to a lower value for relatively large t ("overall structure").

The cohesiveness or separation of structures at different t can be compared in terms of the cumulative distribution functions of $|\mathrm{mst}_t|$. We sampled 100 000 models $\mathbf{X}_i^0$ and 1 000 models $\mathbf{X}^{k,s}$ for each combination of the two parameters $k \in \{1, 4, 9, n\}$ and $s \in \{0.1, 0.2, 0.3\}$. The minimum distance of two objects in the spaced models $\mathscr{X}_{\boxplus}^{n,s}$ is therefore about 0.037, 0.050 and 0.060 respec-

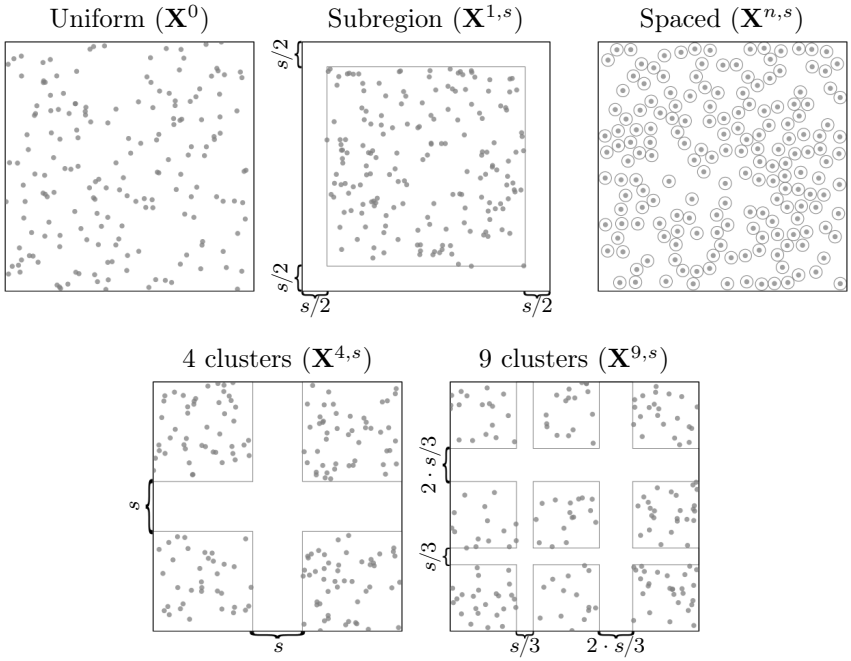

Figure 3.1: Examples from the 5 distributions of experiment 3.1 ($s = 0.2$). The constrained regions are shown as enclosed by thin lines.

tively. The two empirical (complementary) cumulative distribution functions $\mathrm{cdf}_t(v) = \Pr_i\left[|\mathrm{mst}_t(\mathbf{X}_i^0)| \leqslant v\right]$ and $\mathrm{ccdf}_t(v) = \Pr_i\left[|\mathrm{mst}_t(\mathbf{X}_i^0)| \geqslant v\right]$ give the cumulative empirical probability of observing a $|\mathrm{mst}_t|$ as small/large as v in models drawn from $\mathscr{X}_{\boxplus}^0$. Then we define the average separation and cohesiveness indices, $\overline{\mathrm{si}}_{k,s}(t)$ and $\overline{\mathrm{ci}}_{k,s}(t)$ as the empirical probability of observing a less separated/cohesive model $\mathbf{X}_i^0$, averaged over the models $\mathbf{X}_i^{k,s}$:

$$\overline{\mathrm{si}}_{k,s}(t) = 1 - \sum_{i=1}^{1\,000} \frac{\mathrm{cdf}_t\left(\left|\mathrm{mst}_t\left(\mathbf{X}_i^{k,s}\right)\right|\right)}{1\,000} \qquad \overline{\mathrm{ci}}_{k,s}(t) = 1 - \sum_{i=1}^{1\,000} \frac{\mathrm{ccdf}_t\left(\left|\mathrm{mst}_t\left(\mathbf{X}_i^{k,s}\right)\right|\right)}{1\,000}$$

In order to provide a line for comparison, we use $\overline{\mathrm{si}}_0$ and $\overline{\mathrm{ci}}_0$, which are defined in the same manner as above but with respect to $\mathbf{X}_i^0$. The uniform $\mathscr{X}_{\boxplus}^0$ corresponds to a balanced case, where cohesiveness and separation are in an equilibrium for all relevant values of t. Then, if $\overline{\mathrm{si}}_{k,s}(t)$ is greater than $\overline{\mathrm{si}}_0(t)$, we say that the separation in the $\mathbf{X}_i^{k,s}$ dominates its cohesiveness at t. It should be noted that $|\mathrm{mst}_t(\mathbf{X})| \in \{0, \ldots, 179\}$. Due to this finite and rather small range, it is usually *not* the case that $\overline{\mathrm{si}}_{k,s}(t) + \overline{\mathrm{ci}}_{k,s}(t) = 1$. For the same reason, $\overline{\mathrm{si}}_0$ and $\overline{\mathrm{ci}}_0$ are generally below 0.5 and tend towards 0 the more the probability mass of $|\mathrm{mst}_t|$ concentrates on its mode. This is the case for either very small ($|\mathrm{mst}_t(\mathbf{X})| \to 0$) and large values of t ($|\mathrm{mst}_t(\mathbf{X})| \to 179$).

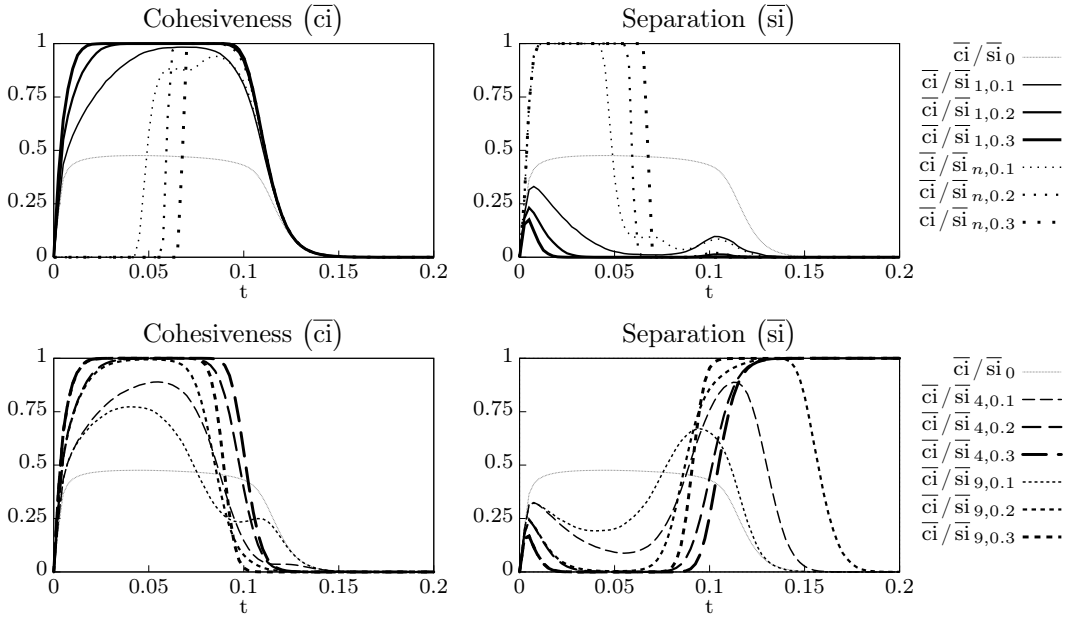

Figure 3.2: Average cohesiveness and separation index of the subregion and spaced ($\overline{\text{ci}}/\overline{\text{si}}_{1,s}$, $\overline{\text{ci}}/\overline{\text{si}}_{n,s}$; *top*) as well as the clustered models ($\overline{\text{ci}}/\overline{\text{si}}_{4,s}$, $\overline{\text{ci}}/\overline{\text{si}}_{9,s}$; *bottom*) compared for different values of s and t. Values above the uniform case ($\overline{\text{si}}_0/\overline{\text{ci}}_0$) correspond to a more cohesive ($\overline{\text{ci}}$)/separated ($\overline{\text{si}}$) structure at t.

Figure 3.2 shows the score for the models from the different model distributions with respect to the average separation and cohesiveness indices.

The subregion and spaced models represent the not clusterable case (fig. 3.2, top). Unsurprisingly, the subregion models are on average at least as and typically more cohesive than the uniform baseline. The spaced models, on the other hand, have an unusually separated small scale structure and turn cohesive for t somewhat above $2r_s$. This is because the objects fill the space more evenly and it is thus less likely that separated clusters occur.

In contrast, we expect a clusterable model to be relatively separated for large t ("dominant structure of mutually separated parts") and relatively cohesive for small t ("parts of cohesive subsets"). This complies with the results shown for models from $\mathscr{X}_{\boxplus}^{4,s}$ and $\mathscr{X}_{\boxplus}^{9,s}$ (fig. 3.2, bottom). However, there is also a visible difference in the graphs for the 4-cluster and 9-cluster models. Especially for $s = 0.1$, the 9-cluster models are relatively separated already on a smaller scale but overall to a lesser extent. We attribute this to the facts that more objects lie at the boundary of a cluster in the case of 9-cluster models and the constrained areas that separate the clusters are thinner.

3.1 Properties of Clusterability Indices

Clusterability indices quantify the clusterability of models. Most of the definitions related to internal evaluation indices that are introduced in section 2.1.1 can be directly adapted to clusterability indices. Furthermore, clusterability indices have to be permutation invariant for the same reason that applies to evaluation indices: the ordering of the dataset is assumed to be arbitrary in general cluster analysis. All clusterability indices we introduce in section 3.2 are indeed permutation invariant.

Definition 3.2 (Clusterability index η) A permutation invariant mapping from the space of models to real valued scores. A higher score corresponds to a more clusterable model.

Definition 3.3 (Permutation invariant clusterability index) The order of the objects within a model has no effect on its score.

Formally, let π_r be a permutation of the set $\{1, \ldots, r\}$ and $\pi_r(i)$ its i-th element. Furthermore, let $\mathbf{x}'$ be the objects of a model $\mathbf{X}'$, $\mathbf{x}$ be those of $\mathbf{X}$. Then η is permutation invariant if, for all models $\mathbf{X}$ and $\mathbf{X}'$

$$\exists_{\pi_n}(\forall_{i,j\in\{1,\ldots,n\}}(\psi_{\mathbf{X}'}(\mathbf{x}'_{\pi_n(i)}, \mathbf{x}'_{\pi_n(j)}) = \psi_{\mathbf{X}}(\mathbf{x}_i, \mathbf{x}_j))) \quad \Rightarrow \quad \eta(\mathbf{X}') = \eta(\mathbf{X})$$

Consistency is defined for evaluation indices in terms of the effects of straightforward changes in the model, which can not be directly adapted to clusterability. In detail, for evaluation indices, the sole decrease of a dissimilarity within a cluster or the sole increase of a dissimilarity between clusters must not decrease the evaluation score (cf. def. 2.12). Since there are no fixed clusters in clusterability evaluation, the formalization of a straightforward change is difficult. For instance, consider the models shown in figure 3.3. Any ranking of the 4 models from least to most clusterable is debatable. In fact, each of the 5 different clusterability indices we analyze in section 3.2 produces a different ranking. The detailed scores are shown in figure 3.4. On the other hand, a ranking with respect to a 4-cluster clustering that corresponds to the 4 multivariate normal distributions from which the objects were sampled[1] is much more straightforward.

Although, due to the scope of it, we have to leave the formalization of consistency to future work, we want to stress its importance for further research on and with clusterability. Consistency can be a guideline in the design of

[1] In the model with $d_v = 0$, there are actually 2 pairs of 2 identical distributions each.

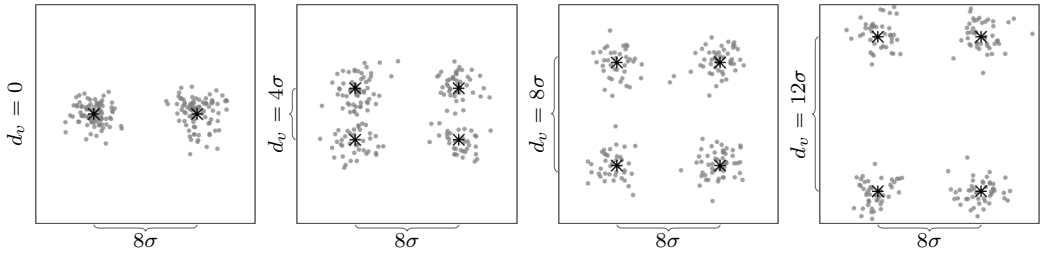

Figure 3.3: Which model is the most clusterable? Models with 180 objects sampled from multiple multivariate normal distributions. The distances between the distribution centers (stars) are specified in terms of σ, the shared standard deviation of the distributions.

new indices. Moreover, all of the clusterability indices we detail in this thesis are based on different assumptions of how clusterability can be measured. A formalized consistency property could help justify or falsify these assumptions. It might be especially interesting with respect to a ranking of models with 1 or n evident clusters. The examples of clusterability indices we detail in section 3.2 score cohesive 1-cluster models as about as clusterable as a uniform model—at least, if no additional knowledge about the generation process is provided. On the other hand, spaced models are most times ranked even below a uniform model. The details are provided in section 3.2. Nevertheless, this evidence for a ranking can only support a sound formalization based on theory, but it can not replace it.

Identically to definition 2.13, scale invariance assures that the overall dissimilarities between the objects are of no importance for the clusterability score of the model. Scale invariance allows the application and interpretation of a clusterability index without knowledge about the scale of the dissimilarities in the model. For example, for scale invariant indices the score is the same whether dissimilarities are measured in meters or centimeters. Note that this only applies to the scale of the dissimilarities and not to the scale of the single attributes of a vector model. Nevertheless, when the actual scale is known and thus appropriate dissimilarity thresholds for "cohesive" and "separated" are known, scale invariance might not be desired.

Definition 3.4 (Scale invariant clusterability index) A uniform scaling of all dissimilarities in the model has no effect on the score of the clusterability index η.

Formally, η is said to be scale invariant if, for all models $\mathbf{X}$ and $\mathbf{X}'$,

$$\exists_{a\in\mathbb{R}} \left(\forall_{\mathbf{x}_i,\mathbf{x}_j} \left(\psi_{\mathbf{X}}(\mathbf{x}_i,\mathbf{x}_j) = \frac{\psi_{\mathbf{X}'}(\mathbf{x}_i,\mathbf{x}_j)}{a} \right) \right) \Rightarrow \eta(\mathbf{X}) = \eta(\mathbf{X}')$$

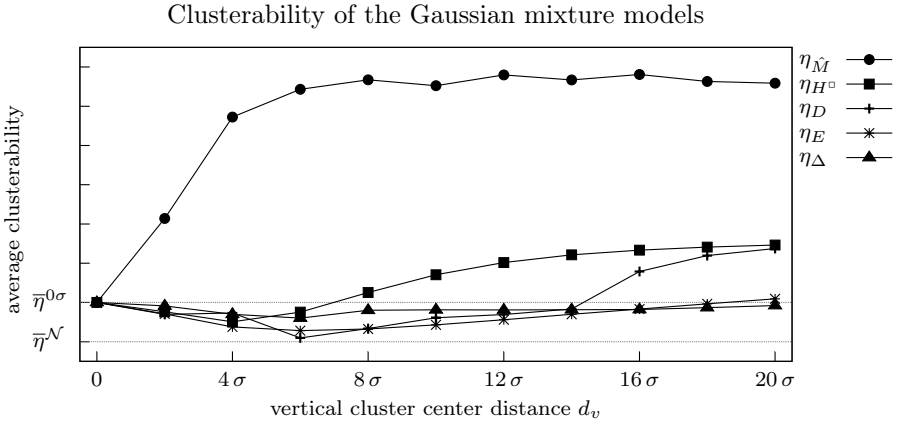

Figure 3.4: Average scores of the 5 different clusterability indices which are analyzed in section 3.2 on different normal mixture distributions (setup as in fig. 3.3; average over 1 000 models for each value $d_v \in \{0, 2\,\sigma, 4\,\sigma, \ldots, 20\,\sigma\}$). Scores are linearly scaled such that the average score on both models from 1 normal distribution ($\overline{\eta}^{\mathcal{N}}$, not shown in fig. 3.3) and on models from 2 distributions ($\overline{\eta}^{0\sigma}$) are on the same level.

Robustness essentially guards against outliers (cf. def. 2.18), which can have a large unwanted effect on the score for not robust η. However, if a robust η is chosen since outliers are possible, the employed clustering algorithm and evaluation index should be robust, as well.

Definition 3.5 (Strongly/weakly robust clusterability index) As the number of objects in a model $\mathbf{X}$ grows, the maximal effect of single dissimilarity changes in $\mathbf{X}$ on its score as measured by the clusterability index η tends towards 0. *Weak* robustness assures a limited effect when one single dissimilarity is changed, while *strong* robustness assures the same for changes to dissimilarities of one object.

A clusterability index η is robust if, for all models $\mathbf{X}$, d_i and models $\mathbf{X}_{\Delta r,d}$ as defined in definition 2.14 and positive real values d and e,

$$\exists_{n_0 \in \mathbb{Z}^+}(n \geqslant n_0 \Rightarrow (|\eta(\mathbf{X}) - \eta(\mathbf{X}_{\Delta r,d})| < e))$$

Where for only weak robust clusterability indices the difference of $\mathbf{X}$ and $\mathbf{X}_{\Delta r,d}$ is restricted to one single dissimilarity.

Baseline models facilitate the interpretation of clusterability scores and can be employed, as detailed for evaluation indices in section 2.1.2, to create abso-

lute variants of clusterability indices. As it is the case for cluster evaluation, cluster-less model distributions can provide a baseline for unstructured models (cf. section 2.1.1). This includes, for example, models with independent dissimilarities from a uniform distribution or object vectors drawn uniformly from a hypersphere. Distribution normalization is much more straightforward for clusterability indices than for evaluation indices as there is no dependence on the optimum clustering (cf. def. 2.16).

Definition 3.6 (Distribution normalized clusterability index) The expected clusterability score and the standard deviation of scores of models from some specific model distribution are known.

Formally, η is called distribution normalized or $\mathscr{X}^0$-normalized if, for models drawn from some fixed model distribution $\mathscr{X}^0$, the expected value (e_η) and standard deviation (s_η) of the score are known.

$$e_\eta = \mathrm{E}_{\mathbf{X}^0 \leftarrow \mathscr{X}^0}\left[\eta(\mathbf{X}^0)\right] \qquad s_\eta = \sqrt{\mathrm{E}_{\mathbf{X}^0 \leftarrow \mathscr{X}^0}\left[\eta(\mathbf{X}^0) - e_\nu)^2\right]}$$

Definition 3.7 (Absolute clusterability index) The score of a model can be interpreted without knowledge on the number of objects in the model or additional model parameters (e.g., model dimensionality in the case of vector models).

Similar to evaluation indices, clusterability indices represent certain assumptions on clusterable models. This can take the form of preferences as section 2.1.1 details for evaluation indices. However, with regard to our definition of clusterable models (def. 3.1), the most cohesive and most separated models are not clusterable at all. These models contain 1 very compact cluster and n spaced "clusters" respectively. Nevertheless, also measures of structure that do not cover these special cases can be of interest in a clusterability analysis. For example, a combination of a pure measure of cohesiveness and a pure measure of separation, like in experiment 3.1, might be able to provide deeper insight into the clusterability of a model.

3.2 Indications of Structure

How can clusterable models be distinguished from not-clusterable ones? As clusterability is an analysis of the structure of a model, measures of structure allow for a comparison of models with respect to their clusterability. This

allows for a ranking of different models of the same dataset. If the measure of structure, the clusterability index, is absolute, models with a score above some index-dependent threshold can be seen as clusterable.

This section introduces some clusterability indices based on related publications. We categorize the indices into three groups based on their basic approach to the measurement of clusterability. Section 3.2.1 details clusterability indices that evaluate clusterings. In section 3.2.2, we present tests for unstructured/random models. Indices based on the spread or concentration of dissimilarities within the model are analyzed in section 3.2.3. Not all of the indices are actually published as clusterability indices. Instead, some of them stem from related topics and are adjusted in order to form an index for clusterability analysis.

The simple indices we introduced for experiment 2.1 and 3.1 have some weaknesses that limit their use in practice. For experiment 2.1, one can use

$$\eta_{\bar{L}_1^{\hat{\sigma}}}(\mathbf{X}) = 1 - \mathrm{ccdf}_1(\bar{L}_1^{\hat{\sigma}}(\mathbf{X})) = \Pr_{\mathbf{X}^0 \leftarrow \mathscr{X}_{\mathcal{N}}^1}\left[\bar{L}_1^{\hat{\sigma}}(\mathbf{X}^0) < \bar{L}_1^{\hat{\sigma}}(\mathbf{X})\right]$$

First and foremost, $\bar{L}_1^{\hat{\sigma}}$ is a measure of spread and not of structure. Spread and structure do not necessarily correlate in general like they do in the experiment. Moreover, a common disadvantage in the use of the empirical cumulative distribution function is the reliance on the distribution tails, which need a large amount of samples (here: models) for their estimation. Furthermore, already moderately clusterable models achieve a perfect score. For example, of the 1 000 models drawn from $\mathscr{X}_{\mathcal{N}}^2$ with 256 objects in 64 dimensions, already 17% achieve a score at or above 0.99 although the empirical risk is still at 0.26. While this can suffice for a classification as in experiment 2.1, it is a drawback in an optimization setting like model selection (cf. section 2.3).

The reliance of $\eta_{\bar{L}_1^{\hat{\sigma}}}$ and similar clusterability indices on distribution tails, as well as an upper limit on the score, can be alleviated when the measurement on the baseline models has a known distribution. For example, we noted from visual inspection that $|\mathrm{mst}_t(\mathbf{X}^0)|$ ($\mathbf{X}^0 \leftarrow \mathscr{X}_{\boxplus}^0$, cf. experiment 3.1) approximately follows a discretized normal distribution for $0.03 \leqslant t \leqslant 0.09$. For these cases, $\overline{\mathrm{ci}}$ and $\overline{\mathrm{si}}$ could be redefined to employ the z-score (as in equation 2.3) instead of the cumulative distribution (equation 2.2). Tests on the equality of an empirical and a theoretical distribution, like Kolmogorov-Smirnov (Smirnov, 1948), can justify the approximation. A drawback of η that make use of $\overline{\mathrm{ci}}$ and $\overline{\mathrm{si}}$ can be their reliance on the minimum spanning tree, which makes them susceptible to the chaining effect (cf. experiment 2.2).

For a comparison of the clusterability indices, we show for each of the ones we detail below the average result on the models of experiment 3.1. Each of the clusterability indices is applied to 1 000 models from the model distributions $\mathscr{X}_{\boxplus}^{1,s}$, $\mathscr{X}_{\boxplus}^{4,s}$, $\mathscr{X}_{\boxplus}^{9,s}$ and $\mathscr{X}_{\boxplus}^{n,s}$ for $s \in \{0.1, 0.2, 0.3\}$ each. Moreover, the indices

are also applied to 1 000 models from $\mathscr{X}_{\boxplus}^{0}$, which corresponds to $s = 0$ for all of the above model types. The respective graphs show the mean score over the 1 000 models. The variance of these scores is shown by shaded regions. These regions have a width of 2 standard deviations and are centered at the averages.

3.2.1 Salient Clusters

As detailed in section 2.1, clusterability is related to cluster evaluation analysis and evaluation indices can be applied in clusterability, as well. Specifically, the score $\eta_\nu(\mathbf{X}) = \max_{\mathcal{C}}(\nu(\mathcal{C}, \mathbf{X}))$ can be a measurement of the clusterability of the model $\mathbf{X}$. If ν is absolute, then η_ν is absolute, too. Other schemes of clusterability indices that make use of evaluation indices exist, as well. For example, one can search for different optimum clusterings under different strong and mutually exclusive assumptions on the structure of $\mathbf{X}$. If one clustering fits $\mathbf{X}$ far better than the other clusterings, this gives an indication of the type and quality of the structure of $\mathbf{X}$. As a disadvantage of these approaches, it is often not feasible to find the optimum clustering and clustering algorithms have to be used for approximation. Next, we introduce one example for each of these schemes.

Dunn index family As an example for the direct use of an evaluation index for clusterability assessment, we choose the well-known index by Dunn (1974). The Dunn index family (Stein et al., 2003) is characterized as

$$\nu_D(\mathcal{C}, \mathbf{X}) = \frac{\min_{C \neq C' \in \mathcal{C}}(\Psi_b(C, C'))}{\max_{C \in \mathcal{C}}(\Psi_w(C))} \tag{3.1}$$

where Ψ_b is a measure of dissimilarity *between* two clusters, while Ψ_w is a measure of the dissimilarity *within* one cluster. The corresponding clusterability index is then

$$\eta_D(\mathbf{X}) = \max_{\mathcal{C}} \left(\nu_D(\mathcal{C}, \mathbf{X})\right)$$

As measures of between- and within-cluster dissimilarity we use the smallest and largest edge length:

$$\Psi_b(C, C') = \min_{\mathbf{x} \in C, \mathbf{x}' \in C'} \psi(\mathbf{x}, \mathbf{x}') \qquad \Psi_w(C) = \max_{e \in \mathrm{mst}(C)} \|e\| \tag{3.2}$$

where $\mathrm{mst}(C)$ are the edges of the minimum spanning tree of the objects in C and $\|e\|$ is the length of the edge e. In this configuration ν_D is also called worst pair ratio (Ackerman and Ben-David, 2009). It should be noted that ν_D is undefined when there is only one cluster or all objects are in a cluster on their

own. This is a weakness of other common evaluation indices, as well (Tibshirani et al., 2001). Because of this, ν_D is not able to distinguish unstructured and spaced models. Additionally, as we show below, η_D relies completely on the minimum spanning tree of $\mathbf{X}$, which makes it sensitive to noise objects (cf. experiment 2.2). Furthermore, outliers also have a large effect on the score as they cause unusually large edges in the minimum spanning tree. While this is not a problem in the noise and outlier-free setting of experiment 3.1, it can be a disadvantage in practice (cf. chap. 4).

Our choice of Ψ_b and Ψ_w allows for a feasible exact computation of η_D. First we note that the optimum clustering of a model $\mathbf{X}$ with respect to ν_D, $\mathcal{C}_{\text{opt}} = \arg\max_{\mathcal{C}}(\nu_D(\mathcal{C}, \mathbf{X}))$,[2] is defined by a threshold t on the edge lengths: all edges less or equal to t are within a cluster and all edges above t are between clusters. This can be proven by contradiction. Let $t_w = \max_{C \in \mathcal{C}}(\Psi_w(C))$ and $t_b = \min_{C \neq C' \in \mathcal{C}}(\Psi_b(C, C'))$. We assume that $\mathcal{C}_{\text{opt}}$ is not defined by a single threshold t. Thus, there exists at least one edge that is smaller than t_w and links two objects of different clusters. In this case $t_b < t_w$. We can therefore merge all clusters that are connected by edges smaller than t_w. This does not change the denominator in equation 3.1 but increases the numerator and thus the overall score. Therefore, the original clustering has not been optimal. It is then easy to see that t_w is the length of the largest edge of the minimum spanning tree of $\mathbf{X}$ that is smaller than t. Correspondingly, t_b is the length of the smallest edge of the tree that is larger than t. Because of this, we get

$$\eta_D(\mathbf{X}) = \max_{k \in \{2,\dots,n-1\}} \frac{\{\text{mst}(\mathbf{X})\}_{n-k+1}}{\{\text{mst}(\mathbf{X})\}_{n-k}} \tag{3.3}$$

where $\{\text{mst}(\mathbf{X})\}_i$ is the i-th smallest edge in the minimum spanning tree of $\mathbf{X}$. We have formatted the equation such that k is the number of clusters in the corresponding clustering. The computational cost is dominated by the calculation of the minimum spanning tree of a fully connected undirected graph with positive weights, which can be done in $\mathcal{O}(n^2)$ by the well-known algorithm by Prim (1957).

The clusterability index η_D is absolute and scale invariant. The score can be directly interpreted as the best possible ratio of between- and within-cluster dissimilarities (def. 3.7). It is scale invariant (def. 3.4) as a uniform scaling also directly scales the edge lengths of the minimum spanning tree, which leaves the ratio in equation 3.3 constant. As the effect of a change of the dissimilarity that corresponds to $\{\text{mst}(\mathbf{X})\}_{n-k}$ is not limited by the number of objects, it is not robust (def. 3.5). Furthermore, it is not $\mathscr{X}_{\boxplus}^{0}$-normalized (def. 3.6).

[2] This corresponds to the definition of optimum clustering with respect to a total clustering relation based on ν_D (def. 2.17).

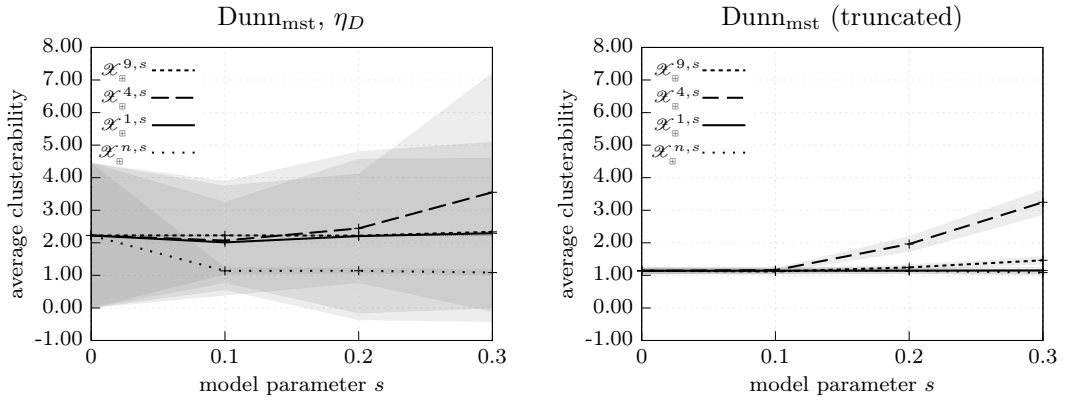

Figure 3.5: (*Left*) Average η_D-score of models from the model distributions of experiment 3.1. The shaded regions cover the area from 1 standard deviation below to 1 standard deviation above the averages. (*Right*) The same but when considering only clusterings up to 13 clusters.

The average η_D-score of the models of experiment 3.1 are shown in the left plot of figure 3.5. The large shaded region corresponds to a very large standard deviation for models of the clustered and uniform model distributions. This stems from models in which two objects are very close to each other in comparison to all other dissimilarities. In these cases, $\{\text{mst}(\mathbf{X})\}_2/\{\text{mst}(\mathbf{X})\}_1$, which corresponds to a clustering with $n-1$ clusters, is very high. Note that the score for the spaced models has a low standard deviation since this effect can not occur in this case. In order to compensate for this effect, we also show the average score when only considering values $k \leqslant 13$ in the right plot.[3] In this case, the subregion and spaced models achieve similar results. Because η_D is scale invariant, the subregion and complete uniform models can not be distinguished. The structure in the models with 9 clusters leads only to a small improvement in the score. In fact, the small gaps of size $s/3$ within these models actually decrease the score, as they lead to a wider range of edge lengths between the natural clusters and thus to a smaller ratio ν_D. This is counter-intuitive for a measure of clusterability.

Evident number of clusters The index by Ostrovsky et al. (2006) compares the quality of optimal clusterings for different number of clusters. The idea is that if one of the optimum clusterings fits the model $\mathbf{X}$ far better than the other clusterings, then $\mathbf{X}$ has a salient structure with the corresponding number of clusters. For the assessment of clustering quality they use RSS, the

[3] We use 13 as it is about $\sqrt{n}$, which we deem to be a plausible number of clusters.

residual sum of squares:

$$\mathrm{RSS}(\mathcal{C}, \mathbf{X}) = \sum_{i=1}^{k} \sum_{\mathbf{x} \in C_i} \left(\|\mathbf{x} - \mathbf{c}_i\|_2^2 \right) \qquad \mathbf{c}_i = \frac{1}{|C_i|} \sum_{\mathbf{x} \in C_i} \mathbf{x} \tag{3.4}$$

Where $\mathbf{c}_i$ is the centroid of cluster C_i and $\|\mathbf{x}\|_2$ is the Frobenius- or L2-norm of $\mathbf{x}$, which is often referred to as the length of the vector $\mathbf{x}$. Thus, RSS requires a vector model and prefers spherical clusters. Note that RSS is not a evaluation index since a larger value indicates a worse clustering quality. The clusterability for a specific number of clusters, k, is given by Λ_k:

$$\Lambda_k(\mathbf{X}) = 1 - \frac{\Delta_k(\mathbf{X})}{\Delta_{k-1}(\mathbf{X})} \qquad \Delta_k(\mathbf{X}) = \min_{\mathcal{C}^k} \mathrm{RSS}\left(\mathcal{C}^k, \mathbf{X}\right) \tag{3.5}$$

Where $\mathcal{C}^k$ is a clustering with exactly k clusters. The clusterability index η_Δ is then defined as:

$$\eta_\Delta(\mathbf{X}) = \max_{k \in \{2,\ldots,k_{\max}\}} \Lambda_k(\mathbf{X})$$

We want to point out that, for all models $\mathbf{X}$, $\Delta_k(\mathbf{X}) \leqslant \Delta_{k-1}(\mathbf{X})$ where an equality requires that all dissimilarities within clusters in the optimal clustering of Δ_{k-1} are 0. Also, since for all models $\Delta_n(\mathbf{X}) = 0$, which is the best possible value, the range of values which are considered for k has to be limited. Moreover, when $k \approx n$, the index suffers heavily from missing robustness similar to η_D. For our experiments we use $k_{\max} = 10$.

Unfortunately, an exact computation of $\Delta_k(\mathbf{X})$ is only feasible in one dimension or if k is either 1 or about n. We thus employ the clustering algorithm K-Means (MacQueen, 1967, chap. 3.6), γ_{km}, in our experiments. K-Means is known to optimize RSS for a selectable number of clusters. Moreover, it is an algorithm that converges to a local minimum of RSS (cf. local optimum clustering, def. 2.17). It does so by iteratively assigning the $\mathbf{x}$ to the nearest of the k centroids $\mathbf{c}_i$ and then recomputing $\mathbf{c}_i$ as defined in equation 3.4. The local minimum the algorithm converges to depends on the initial position of the centroids. We select the centroids by random from the $\mathbf{x}$. In order to improve the approximation of the optimum clustering, we repeat K-Means with different initial centroid sets: $\gamma_{\mathrm{km}}^{k,i}$ for $i \in \{1, \ldots, 1\,000\}$ where i is a seed for the pseudo-random number generation used in the selection of the initial centroids and k is the number of clusters to find. $\Delta_k(\mathbf{X})$ in equation 3.5 is thus replaced by:

$$\hat{\Delta}_k(\mathbf{X}) = \min_i \mathrm{RSS}\left(\gamma_{\mathrm{km}}^{k,i}(\mathbf{X}), \mathbf{X}\right)$$

Out of the properties we introduced in section 3.1, η_Δ only fulfills scale invariance and the mandatory permutation invariance. A uniform scaling of the dissimilarities by a increases the RSS by a^2. This factor, however, is canceled out by the ratio in equation 3.5 (cf. def. 3.4). Since the interpretation of Λ_k depends also on k and the dimensionality of the vector space, η_Δ is not absolute (def. 3.7). When $\Delta_{k-1}(\mathbf{X})$ is relatively small, even a small absolute change in $\Delta_{k-1}(\mathbf{X})$ due to a change in the dissimilarities can have a large effect on the score. Therefore, it is not (strong) robust (def. 3.5). We omit the formal proof and only briefly sketch the counterexample that shows the lack of robustness. Consider two very tightly packed clusters with sufficiently large distance from each other and one outlier. The optimal clustering contains three clusters: the two actual clusters and the outlier. For the clustering with two clusters, the outlier is added to one of the clusters. We demand the clusters to be packed so tight that the dissimilarity of the outlier to the centroid is the major factor in the RSS. Note that in order to achieve this, the actual dissimilarity of the outlier does not have to be large in an absolute sense. Instead, it suffices that the clusters are really tight. As long as the clusters are tight enough, the actual number of objects is unimportant and does therefore not limit the change in the score. This, however, is required for robustness. As a clusterability index that requires vector models, weak robustness as such can not be applied since a single dissimilarity can not be changed without a change in others.

The average scores on the models of experiment 3.1 are shown in figure 3.6. The left plot shows the average score for the different model distributions. Models from the spaced distribution achieve, on average, a lower score than the uniform models. Since no score can be calculated for $k = 1$ and $k = n$, the corresponding clusterings that would actually fit the model well are not considered. Thus, the score is relatively low in both cases. While the score for models from $\mathscr{X}_{\boxplus}^{4,s}$ increases appropriately with a higher separation s, this occurs only to a much smaller extend for $\mathbf{X}^{9,s} \leftarrow \mathscr{X}_{\boxplus}^{9,s}$. We show the average Λ_k for models from $\mathscr{X}_{\boxplus}^{9,s}$ in the right plot. First, a sudden decrease in the score from 4 to 5 clusters can be noticed. We assume that this stems from the rectangular setup with 4 corners. Second, although there is a noticeable increase in the score for $k = 9$ with an increased separation s, the score for 2 or 3 clusters is still higher. In order to rule out the possibility that the 9 clusters are not detected by K-Means, we confirmed that they are indeed detected. This incidence of a high relative but small absolute score is a symptom of the missing absoluteness of the Λ_k and therefore of η_Δ. In detail, the low overall score despite the large increase at $k = 9$ is a result of the direct comparison of Λ_k for different k, although the interpretation of them depends, among others, on k.

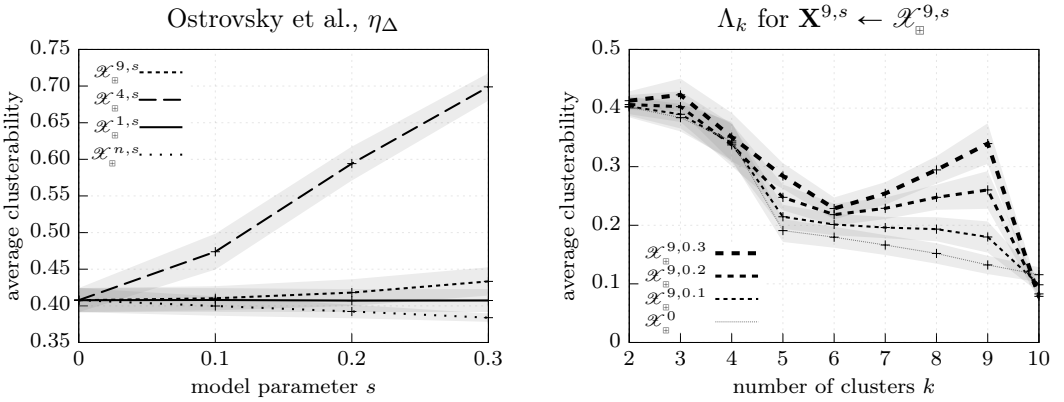

Figure 3.6: (*Left*) Average η_Δ-score of models from the model distributions of experiment 3.1. (*Right*) Λ_k for models from $\mathscr{X}_\boxplus^{9,s}$. The shaded regions cover the area from 1 standard deviation below to 1 standard deviation above the averages. Please note the different scaling of the two plots.

Ostrovsky et al. introduce a polynomial time approximation scheme to the problem of finding a near-optimal clustering for η_Δ-clusterable models. In their proofs, they use different lower bounds on η_Δ, which can be seen as a threshold above which a model is deemed clusterable. One such bound in the analysis of a part of their scheme is $(1-\eta_\Delta(\mathbf{X}))/(p\cdot\eta_\Delta(\mathbf{X})) \leqslant 1/14$ for $p \leqslant 0.1$ (Ostrovsky et al., 2006, section 4.1.2), which can be weakened to $\eta_\Delta(\mathbf{X}) \geqslant 0.999$. In order to put this threshold into context, we sampled 1 000 models from $\mathscr{X}^{4,0.9}$, for which the objects are concentrated on only 1% of the total area. The highest achieved score on these models is still below 0.997. We thus argue that this result of Ostrovsky et al. is unlikely to be relevant in practice.

3.2.2 Tests for the Lack of Structure

As clusterable models are structured, not clusterable models are lacking structure, which might be easier to identify. Unstructured models are sometimes referred to as random (Jain and Dubes, 1988). However, even structured model distributions are random in some sense and we thus refrain from the usage of "random" in this context in order to avoid ambiguity. The clusterability indices we detail in this section perform statistical tests under the assumption of a certain unstructured model. The actual score of the indices is a statistic that is measured on the model. If the score is unusually high, the test is said to have failed and the hypothesis that the model is unstructured is rejected. Equivalently, the model is said to be clusterable.

Statistical tests for unstructured models can be categorized based on the model type they apply to. Next, we list some methods that only require the

graph representation of a model, although they assume that the graph is fully connected. The usual unstructured model that is tested for in this case is that of independent and uniformly distributed pairwise dissimilarities. On the other hand, a significant portion of the tests stem from fields in which the objects lie in a vector space. Jain and Dubes list tests for applications in astronomy, ecology, forestry and geography. In these cases, the unstructured hypothesis is that of a uniform distribution over a certain area, which is called the sampling window of the model distribution. We will discuss this kind of tests after the tests on graphs and detail one such test that is based on the statistic by Hopkins and Skellam (1954).

Unstructured Graphs

These methods are similar in spirit to $\overline{\text{si}}$ and $\overline{\text{ci}}$ in experiment 3.1, although it should be noted that experiment 3.1 tests for a lack of *spatial* structure. Indeed, one of the methods mentioned by Jain and Dubes (1988) is the number of connected components of the model $\mathbf{X}$ at a certain threshold t, which is just $n - |\text{mst}_t(\mathbf{X})|$. The expected number of components for a model with uniform pairwise dissimilarities can be found for some values of n and t in the tables by Ling and Killough (1976). The appropriate values can be subtracted from the score such that unstructured models achieve, on average, a score of 0. This makes the clusterability index distribution normalized.

Another possible statistic is the number of edges that are necessary to connect every object in the model when they are added from shortest to longest edge. This is equal to the number of edges that are shorter or of equal length than the longest edge in the minimum spanning tree. For clusterable models, it is expected that it takes a relatively large amount of the edges to connect all objects. Since the objects reside in separated clusters, it is expected that the edges within the clusters are added before the edges between clusters are considered. As some of the between-cluster edges are also part of the minimum spanning tree, this increases this statistic. In comparison to $\overline{\text{si}}$, this removes the dependency on a parameter (t) and makes the clusterability index scale invariant. The score can again be distribution normalized by using tables by Ling and Killough. They also note that the exact distribution of the statistic can be obtained, which could then be used to interpret the score. By incorporating this distribution in the clusterability index, it can thus become absolute.

Lack of Spatial Structure

Spatial statistics assume a uniform distribution of objects over a sampling window.

Definition 3.8 (Sampling window (of a model distribution of vector models)) The subspace of the vector space that contains all objects that could possibly be sampled from the distribution.

For example, for distributions $\mathscr{X}_{\boxplus}^{1,0.2}$, $\mathscr{X}_{\boxplus}^{4,0.2}$ and $\mathscr{X}_{\boxplus}^{9,0.2}$ (cf. experiment 3.1), the sampling window consists of the 1, 4 and 9 squares respectively that are shown in the corresponding plots in figure 3.1. The sampling window that is used to determine the clusterability of the models, on the other hand, is the whole square that is used by $\mathscr{X}_{\boxplus}^{0}$.

The reliance on a sampling window can be a disadvantage for these clusterability indices. If the sampling window is not known, it has to be estimated in order to allow the usage of the clusterability index. Estimations are usually either the smallest axis-aligned hyperrectangle that contains all objects or the convex hull (Jain and Dubes, 1988). However, as demonstrated by several authors, including Jain and Dubes, a wrong sampling window can easily mislead the clusterability index that relies on it. In detail, models can be deemed clusterable under one sampling window but not under another one. For example, models $\mathbf{X} \leftarrow \mathscr{X}_{\boxplus}^{1,0.3}$ are clusterable with respect to sampling window of $\mathscr{X}_{\boxplus}^{0}$, but not with respect to the sampling window of $\mathscr{X}_{\boxplus}^{1,0.3}$. Therefore, a different type of clusterability indices may be more adequate when the sampling window is not known.

Tests on spatial structure usually focus on the relative density of subregions of the sampling window S. A dense region contains, with respect to its volume, relatively many objects as compared to other subregions of S. A clusterable model has dense and sparse regions, while a non-clusterable model is usually evenly distributed. Note that from this intuition, one can see that these methods do also categorize models with only one cluster as clusterable as long as the cluster is dense. Tests on the density rely, for instance, on the number of objects in the most dense subregion or on the distribution of object counts in the different cells of an m-dimensional grid. More detailed examples are provided by Jain and Dubes (1988).

One statistic that relies on nearest neighbor dissimilarities is proposed by Cox and Lewis (1976). In the computation of the statistic, r objects $\mathbf{x}_i^S$ are sampled from the sampling window S. Let $\psi_1(\mathbf{x}_i^S)$ be the dissimilarity to the nearest (i.e., least dissimilar) object in the actual model $\mathbf{X}$, $\mathbf{x}_i^{\mathbf{X}}$, and let $\psi_1(\mathbf{x}_i^{\mathbf{X}})$ be the dissimilarity of this object to its nearest neighbor. Examples for these distances are shown in the left plot of figure 3.7. The statistic then considers the ratio $\psi_1(\mathbf{x}_i^S)/\psi_1(\mathbf{x}_i^{\mathbf{X}})$. The details can be found in the publication by Cox and Lewis, who normalize the statistic to be uniform in the range $[0, 1]$ when the objects are uniformly sampled from a plane. An adaption to vector spaces with more than 2 dimensions is provided by Panayirci and Dubes (1983). In-

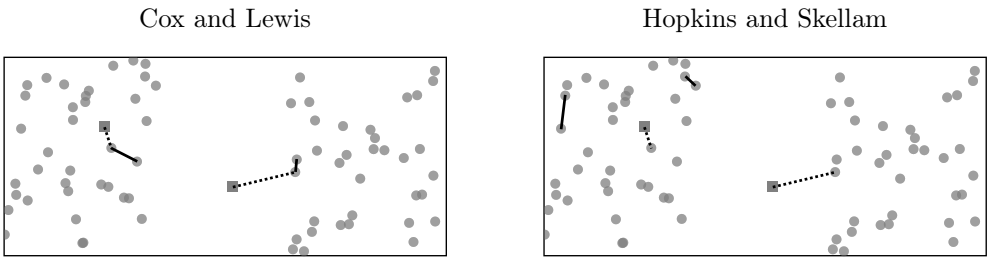

Figure 3.7: Examples of sampled pairwise dissimilarities that are used for the statistics by Cox and Lewis (1976, *left*) and Hopkins and Skellam (1954, *right*). Original objects are depicted as disks and objects sampled uniformly from the entire shown space as squares. The measured dissimilarities are shown as lines. The lines for dissimilarities between sampled objects and their nearest neighbor are dashed.

tuitively, clusterable models contain regions that are free of objects. If objects $\mathbf{x}_i^S$ are sampled from such a region, then $\psi_1(\mathbf{x}_i^S) > \psi_1(\mathbf{x}_i^{\mathbf{X}})$. In this case, one can see $\psi_1(\mathbf{x}_i^S)$ as the distance to the nearest cluster and $\psi_1(\mathbf{x}_i^{\mathbf{X}})$ as a distance within a cluster. If $\psi_1(\mathbf{x}_i^S) > \psi_1(\mathbf{x}_i^{\mathbf{X}})$ for most i, $\mathbf{X}$ tends to be clusterable.

Next, we detail the related statistic by Hopkins and Skellam (1954). Unlike the statistic by Cox and Lewis, it does not use the dissimilarity $\psi_1(\mathbf{x}_i^{\mathbf{X}})$. Instead it employs the dissimilarity of a randomly chosen object from $\mathbf{X}$ to its nearest neighbor (fig. 3.7, right). This difference can be important for data collection. For example in forestry, which is the application for which the Cox and Lewis statistic has been proposed, it is far easier to determine the nearest tree than to choose a tree randomly from a forest. The latter would be necessary for the statistic by Hopkins and Skellam. A comparison of these two and other similar statistics is provided by Panayirci and Dubes (1983). They note that a test based on the statistic by Cox and Lewis seems to perform worse than a test based on the one by Hopkins and Skellam if the clusters are not perfectly separated.

Hopkins and Skellam statistic The statistic H_r by Hopkins and Skellam (1954) is defined by

$$\mathrm{H}_r(\mathbf{X}, \mathbf{X}^S, \pi) = \frac{\sum_{i=1}^{r} \left(\psi_1\left(\mathbf{x}_i^S\right)\right)^m}{\sum_{i=1}^{r} \left(\psi_1\left(\mathbf{x}_i^S\right)\right)^m + \sum_{i=1}^{r} \left(\psi_1\left(\mathbf{x}_{\pi(i)}\right)\right)^m}$$

where $\mathbf{X}^S$ is an additional model that contains r objects $\mathbf{x}_i^S$ that are drawn uniformly from the sampling window S, π is a permutation of the integer numbers $\{1, \ldots, n\}$, $\psi_1(\mathbf{x})$ is the dissimilarity to the least dissimilar $\mathbf{x} \in \mathbf{X}$, n is the number of $\mathbf{x} \in \mathbf{X}$ and m is the dimensionality of the $\mathbf{x}$ and the $\mathbf{x}^S$.

A random permutation π is employed in order to select r different $\mathbf{x} \in \mathbf{X}$ at random by $\mathbf{x}_{\pi(i)}$. Examples of the dissimilarities used by the statistic are shown in figure 3.7. Note that the statistic is bounded to $[0, 1]$ and that the expected value of H_r if $\mathbf{X}$ is also drawn uniformly from S is 0.5. The latter can be deduced from the fact that $\mathrm{E}\left[\sum_{i=1}^{r}\left(\psi_1\left(\mathbf{x}_i^S\right)\right)^m\right] = \mathrm{E}\left[\sum_{i=1}^{r}\left(\psi_1\left(\mathbf{x}_{\pi(i)}\right)\right)^m\right]$ in this case.

Note that the statistic can be undefined for finite sampling windows when the model contains identical object vectors, as this can result in a division 0/0.

The corresponding clusterability index, η_{H^S}, is then defined through the randomized choice of $\mathbf{X}^S$ and π. In detail, we denote by $\mathscr{X}^{S,r}$ the uniform model distribution over the sampling window S and by Π^n a uniform distribution over all permutations of $\{1, \ldots, n\}$. A total of l models and permutations are drawn from the corresponding distributions: $\mathbf{X}_i^S \leftarrow \mathscr{X}^{S,r}$ and $\pi_i \leftarrow \Pi^n$. We define the clusterability index by

$$\eta_{H^S}(\mathbf{X}) = \frac{1}{l}\sum_{i=1}^{l}\mathrm{H}_r(\mathbf{X}, \mathbf{X}_i^S, \pi_i)$$

The average over l applications is performed in order to decrease the variance of the statistic over different samples from the same model.

If $\mathbf{X}$ is sampled uniformly from the sampling window S and the used dissimilarities are independent of each other, H_r follows a beta-distribution with shape parameters of r, $\beta_{r,r}$ (Panayirci and Dubes, 1983). According to Panayirci and Dubes, the assumption of independence is reasonable for $r < n/10$. We certify this by measurement of the statistic on 1 000 models $\mathbf{X}^0 \leftarrow \mathscr{X}_{\boxplus}^0$, where $\mathscr{X}_{\boxplus}^0$ is the uniform model distribution over the entire sampling window (cf. experiment 3.1). We symbolize the sampling window of this distribution in our notation by a filled square, $S = \blacksquare$. Since the models contain 180 objects each, we use $r = 17 < 180/10$. The theoretical distribution density as well as the empirical distribution in the experiment are shown in figure 3.8. For this shape value, 99.99% of the density mass of the distribution lies in the interval $[0.2, 0.8]$. In the case of an optimal fit, the line for $l = 1$ in the quantile-quantile plot (right) would be equal to the main diagonal. We assume that the small deviation stems from edge effects, that is objects near the border of the sampling window have a different distribution of nearest neighbor dissimilarities. As the distribution of $\eta_{H^{\blacksquare}}$ for $l = 1\,000$ concentrates on the interval $[0.4, 0.6]$, the decreased variance is clearly noticeable. However, the decreased variance also has the effect that $\eta_{H^{\blacksquare}}$ is no longer distributed as $\beta_{r,r}$. Although we noticed in experiments that, for shape parameters somewhat in the range of $[120, 160]$, the β-distribution is close to the distribution in the experiments, we have no theoretical justification for this observation.

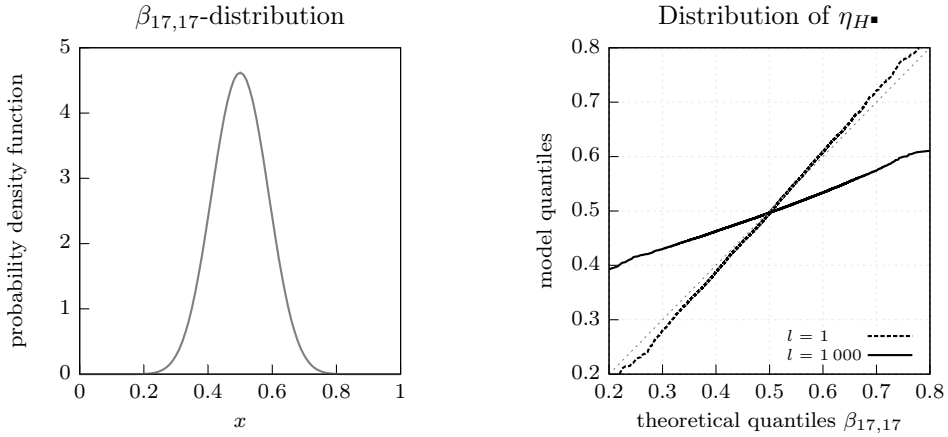

Figure 3.8: (*Left*) Density of the β distribution with both shape parameters equal to 17. (*Right*) Quantile-quantile plot of the $\beta_{17,17}$-distribution quantiles and empirical quantiles for models sampled uniformly from a 2-dimensional square. In the case of a perfect fit, the black lines would coincide with the main diagonal, which is shown as straight gray dotted line.

The clusterability index η_{H^S} is $\mathscr{X}^0_{\boxplus}$-normalized, can be made absolute when the theoretical distribution is known and included in the score, scale invariant when the sampling window is estimated from the model and not robust. The expected value for η_{H^S} for the case that the $\mathbf{x} \in \mathbf{X}$ are sampled uniformly from the sampling window S is 0.5 as both the $\mathbf{x}$ and the $\mathbf{x}^S$ follow the same distribution in this case (def. 3.6). If the theoretical distribution of η_{H^S} is known (see above), it can be used to interpret the score with respect to a uniform model distribution. However, in order to make η_{H^S} absolute, it has to be normalized with respect to the different variances for different r (def. 3.7). If S is estimated from the objects, it scales along with the dissimilarities, which makes the index scale invariant (def. 3.4). η_{H^S} with a fixed sampling window is not scale invariant as the $\psi_1\left(\mathbf{x}_i^S\right)$ scale differently from the $\psi_1\left(\mathbf{x}_{\pi(i)}\right)$ in this case. Additionally, in the case of an estimated sampling window, a single outlier, for example due to a measurement error, can lead to a too large estimation for S and thus to a very different overall score. If, on the other hand, S is known and fixed, outliers outside of S will barely have any effect on the statistic as they are unlikely to be the nearest neighbor of any $\mathbf{x}$ or $\mathbf{x}^S$. Nevertheless, $\eta_{H\blacksquare}$ is not strong robust (def. 3.5), as well, as the repositioning of a single object in the center of an originally object-free area has an impact on all the $\mathbf{x}^S$ that are sampled from this area. Especially if this area is large, for example when all $\mathbf{x}$ are close to the border of S and thus there is a large area in the center of S without any $\mathbf{x}$, the number of dissimilarities that is impacted by this repositioning does not diminish with a growing n (with $r \approx n/10$).

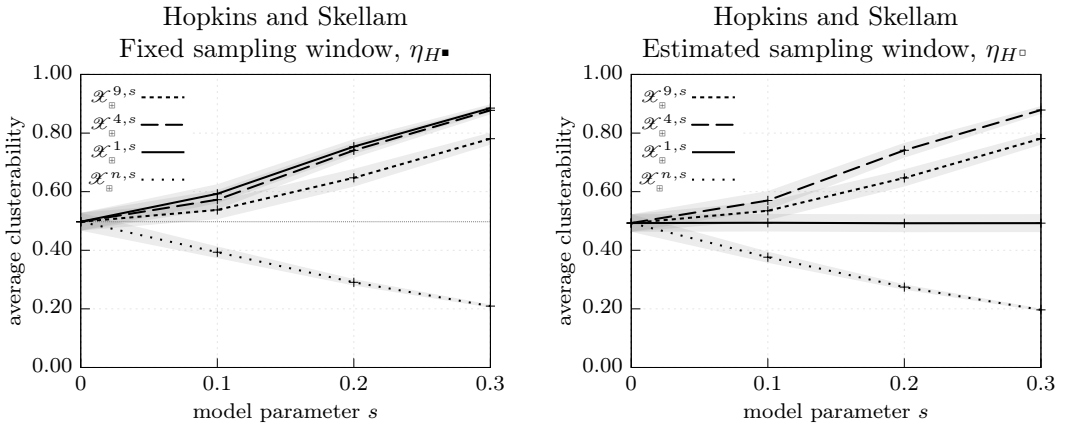

Figure 3.9: (*Left*) Average η_{H^S}-score of models from the model distributions of experiment 3.1 with a fixed sampling window of the entire square shown in figure 3.1. The thin gray line shows the average score for models sampled uniformly from the entire square. (*Right*) The same but with a sampling window estimated for each model as the smallest axis-aligned rectangle that contains all objects. The shaded regions cover the area from 1 standard deviation below to 1 standard deviation above the averages.

The average score on the models from experiment 3.1 for $l = 1\,000$ and $r = 17$ are shown in figure 3.9. The index $\eta_{H\blacksquare}$ uses the fixed sampling window over the entire square while $\eta_{H\square}$ estimates the window as the smallest axis-aligned rectangle that contains all objects. Unsurprisingly, the only noticeable difference occurs for the subregion models drawn from $\mathscr{X}_{\boxplus}^{1,s}$, as all the other models are likely to contain objects close to the borders of the entire square. Since about 99% of the models from $\mathscr{X}_{\boxplus}^{0}$ achieve a score between 0.4 and 0.6 (cf. fig. 3.8), one can say that a score below 0.4 or above 0.6 is already a significant evidence for a structured model. This applies to the models from distributions with $s \geqslant 0.2$. The clear distinction in the score of $\eta_{H\square}$ between the clustered (from $\mathscr{X}_{\boxplus}^{4,s}$, $\mathscr{X}_{\boxplus}^{9,s}$) and the spaced models (from $\mathscr{X}_{\boxplus}^{n,s}$) is apparent. It can also be seen that the 4-cluster models achieve a higher score than the 9-cluster models. This is the case as the bands that are free from objects are wider in the 4-cluster case (cf. fig. 3.1). Interestingly, the subregion models achieve a similar score as the 4-cluster models with respect to $\eta_{H\blacksquare}$. The reason for this is that, although the object-free bands have only half the width for models from $\mathscr{X}_{\boxplus}^{1,s}$ as for models from $\mathscr{X}_{\boxplus}^{4,s}$, only one side of them is adjacent to an area that contains objects. Therefore, $\mathrm{E}\left[\psi_1\left(\mathbf{x}_i^S\right)\right]$ for $\mathbf{x}^S$ from the object-free bands is identical for both types of models.

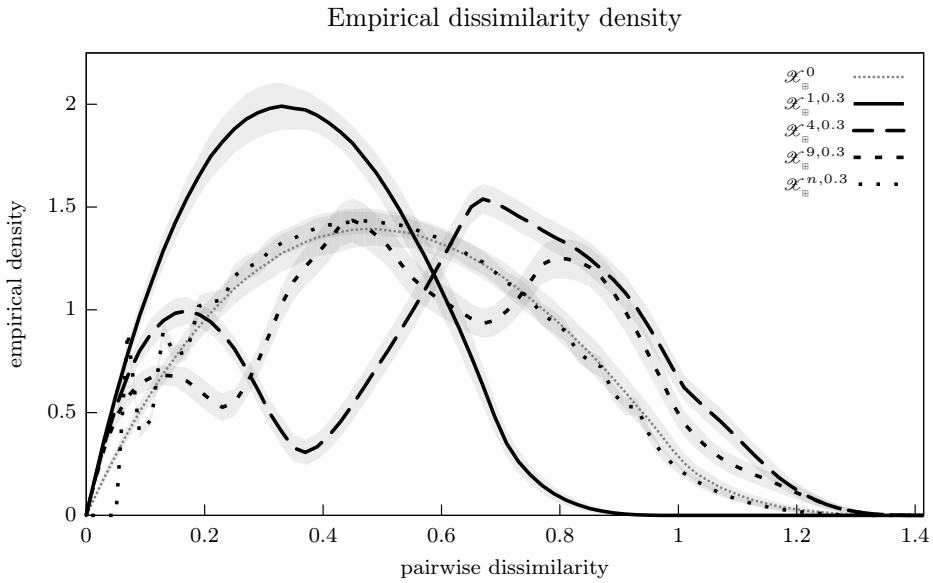

Figure 3.10: Empirical densities of pairwise dissimilarities averaged over the models from some of the model distributions of experiment 3.1. Densities are estimated over dissimilarity-bins of size 0.02. The shaded regions cover the area from 1 standard deviation below to 1 standard deviation above the averages.

3.2.3 Concentration of Dissimilarities

A third approach to clusterability is based on the empirical density of the pairwise dissimilarities in a model. In a clusterable model, it is expected that there are small dissimilarities within the clusters and large dissimilarities between the clusters. For clusterable models, the empirical density of pairwise dissimilarities is thus likely to have multiple modes, that is, clear local maxima separated by a clear local minimum. Models with only one cluster, on the other hand, are expected to have only one local maximum. This notion is illustrated in figure 3.10 for the model distributions from experiment 3.1. However, it is not guaranteed that all dissimilarity distributions of clusterable models have multiple modes or that all distributions of non-clusterable models have only one. Nevertheless, besides the number of modes also other properties of the empirical dissimilarity density, like the overall spread, can be useful to identify clusterability.

Spread of pairwise dissimilarities One index that measures the variability in the pairwise dissimilarities for an assessment of the clusterability of models is proposed by Dash et al. (1997). In detail, they use $\mathrm{E}(\mathbf{X})$, which is

defined as

$$\mathrm{E}(\mathbf{X}) = -\sum_{\mathbf{x},\mathbf{x}' \in X} \left(\varphi_{\mathbf{X}}(\mathbf{x},\mathbf{x}') \cdot \log_2(\varphi_{\mathbf{X}}(\mathbf{x},\mathbf{x}')) + (1-\varphi_{\mathbf{X}}(\mathbf{x},\mathbf{x}')) \cdot \log_2(1-\varphi_{\mathbf{X}}(\mathbf{x},\mathbf{x}')) \right)$$

where $\varphi_{\mathbf{X}}$ is a measure of object *similarity* that ranges from 0 (not similar) to 1 (practically identical). In the case two objects have a similarity of 0 or 1, the term $0 \cdot \log_2(0)$ is replaced by 0. Because the equation has some resemblance to the equation for information theoretic entropy, Dash et al. refer to $\mathrm{E}(\mathbf{X})$ as entropy of the model $\mathbf{X}$. This suggests a connection between similarities (E) and probabilities (information theoretic entropy). Under this aspect, pairs of objects that are either very similar or very dissimilar correspond to predictable events: same cluster and different clusters. On the other hand, object pairs with a similarity close to 0.5 have a high information content. As these should be relatively rare in clusterable models, $\mathrm{E}(\mathbf{X})$ should be relatively low for a clusterable model $\mathbf{X}$. In our opinion, this line of thought is somewhat devious in the connection between similarities and probabilities. Nevertheless, the measurement of similarity variance through E is also not unreasonable outside of the entropy context.

In the case of models based on object dissimilarities, Dash et al. suggest a conversion of dissimilarities to similarities such that the average dissimilarity yields a similarity of 0.5. In detail, they propose the use of $\varphi_{\mathbf{X}}(\mathbf{x},\mathbf{x}')$ as

$$\varphi_{\mathbf{X}}(\mathbf{x},\mathbf{x}') = \exp\left(\frac{\log(0.5) \cdot \psi(\mathbf{x},\mathbf{x}')}{\bar{\psi}(\mathbf{X})}\right) \qquad \bar{\psi}(\mathbf{X}) = \sum_{\mathbf{x},\mathbf{x}' \in X} \frac{\psi(\mathbf{x},\mathbf{x}')}{n \cdot (n-1)} \tag{3.6}$$

Where $\bar{\psi}(\mathbf{X})$ is the average dissimilarity between the n objects of $\mathbf{X}$.

We then define the clusterability index η_E such that a higher score corresponds to a more clusterable model:

$$\eta_E(\mathbf{X}) = 1 - \frac{\mathrm{E}(\mathbf{X})}{n \cdot (n-1)}$$

Although our experiments show a somewhat stable result for models sampled from $\mathscr{X}_{\boxplus}^{0}$ even for different number of objects, we were not able to determine the expected value for such models analytically. Without such a result, however, η_E can not be proven to be distribution normalized or absolute (def. 3.6, 3.7). On the other hand, due to the employed similarity function that normalizes with respect to the average dissimilarity, it is trivially scale invariant (def. 3.4). Moreover, it is robust in the strong sense (def. 3.5), as any change to a single object affects only $\mathcal{O}(n)$ of the $\mathcal{O}(n^2)$ dissimilarities that are used in the computation of E. Thus, for large enough n, the unaffected dissimilarities outweigh any change to the affected dissimilarities.

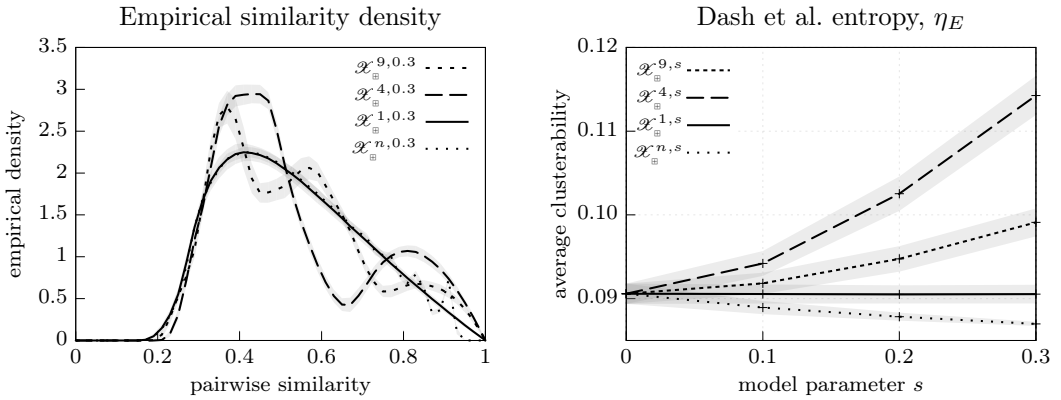

Figure 3.11: (*Left*) Empirical density of pairwise similarities ($\varphi_{\mathbf{X}}$, equation 3.6) averaged over some of the model distributions of experiment 3.1. Densities are estimated over similarity-bins of size 0.02. (*Right*) Average η_E-score. The shaded regions cover the area from 1 standard deviation below to 1 standard deviation above the averages.

The η_E-scores of the models from experiment 3.1 are shown in figure 3.11. The left plot shows the empirical density of pairwise similarities $\varphi_{\mathbf{X}}$ (equation 3.6) for models with a high separation. The most clusterable models with respect to η_E have most of their density close to 0 and close to 1. The relative high densities close to 1 of models from $\mathscr{X}_{\boxplus}^{9,0.3}$ and especially $\mathscr{X}_{\boxplus}^{4,0.3}$ lead to an increased clusterability score (right). Although the densities for models from $\mathscr{X}_{\boxplus}^{1,0.3}$ and $\mathscr{X}_{\boxplus}^{n,0.3}$ are very similar for similarities in the range $[0, 0.8]$, the difference for values above 0.8 suffices to produce a visible difference in the η_E-score (right). Since the influence of the similarities very close to 1 on the score is especially high, the low density leads to a smaller clusterability score.

Intrinsic Dimensionality

The concentration of dissimilarities that η_E exploits as a characteristic of not clusterable models is also known to be related to the dimensionality of vector models. One result by Beyer et al. (1999) is the proof that, for some model distributions and dissimilarity functions, in high-dimensional spaces the dissimilarity between any object and its most dissimilar object is relatively equal to its dissimilarity to its most similar object.[4] This result is especially a problem with regard to nearest-neighbor searches in high-dimensional data, as it

[4]This is one of a collection of effects in high-dimensional spaces that are known as the "curse of dimensionality." The reader interested in the causes of these effect is referred to the introduction by Köppen (2000).

puts the meaningfulness of this task itself into question. However, it was noted by some researches, including Chávez et al. (2001) and Korn et al. (2001), that the theoretical result does not directly apply to practical models. Based on this observation, they suggest the estimation of an "intrinsic dimensionality" from the dissimilarities of the model and show its correlation with the practicality of nearest neighbor search. Interestingly, Korn et al. show the connection between the search time of nearest neighbor searches in a special search structure and the (intrinsic) Fractal dimensionality of the model. The common box-count estimation for Fractal dimensionality they employ uses the distribution of the objects over a grid of hypercubes, which is also employed by some tests for spatial structure (cf. section 3.2.2). This can be seen as further evidence for the relationship of intrinsic dimensionality and clusterability.

Some estimators of intrinsic dimensionality can be used in the measurement of model clusterability, although—or because—assumptions they are based on are violated for clustered models. For example, Peres and Netto (2004) use dimensionality estimates at different scales in order to detect clusters. In detail, they count the number of hypercubes in a vector space grid that contain objects for different granularities of the grid. When clusters are together in one hypercube in a coarse grid and in separate hypercubes in a finer grid, the ratio of hypercubes that contain objects is likely to be lower in the second case. The assumption is based on the idea that the finer grid contains hypercubes that lie between the clusters, and which are thus empty. With respect to the box-count estimation for Fractal dimensionality, this corresponds to a decreased intrinsic dimensionality estimate. We want to point out that the actual method for estimating the intrinsic dimensionality based on box-counts assumes that the dimensionality estimate is nearly constant over a wide range of granularities (Korn et al., 2001). Similarly, the estimator by Chávez et al. (2001) is based on the spread of the pairwise dissimilarities around their mean. The dissimilarity density of models that contain well-separated clusters usually has multiple modes (cf. fig. 3.10), which is not considered in the Chávez et al. estimate. The estimator by Levina and Bickel (2004) relies on a consistent distribution of objects in a neighborhood, which is not the case near the borders of a model distribution. As models that contain multiple clusters tend to contain more objects at a border of the distribution, such estimators are somewhat misled for these models. Note that in all three cases, the intrinsic dimensionality is estimated to be lower when the objects are grouped in clusters.

We continue with a more detailed analysis of the estimator by Levina and Bickel (2004). The estimator is based on the assumption that the objects are actually distributed according to some lower-dimensional model distribution, but are smoothly mapped into the higher dimensional space. For example, the plane $x_1 + x_2 + x_3 = 0$, although being an entity in a 3-dimensional space,

is still a 2-dimensional plane. Similarly, objects sampled uniformly from a plane continue to have the usual distribution of pairwise dissimilarities of a 2-dimensional model distribution. As this dissimilarity distribution differs for model distributions with a different number of dimensions, the original dimensionality can be detected. Levina and Bickel make a dimensionality estimate at each object based on the number of "neighbored" objects in a growing hypersphere. Under the assumption of a uniform model distribution, the expected number of objects in a hypersphere with fixed radius depends on the denseness of the distribution and the volume of the sphere. In turn, this volume depends on the dimensionality of the space. In a higher-dimensional space, the volume of the hypersphere has a more steep increase by radius. Therefore, the increase in the number of objects in the growing sphere is also expected to be more steep. From this reasoning, Levina and Bickel deduce an estimator for intrinsic dimensionality based on the radius of the hyperspheres. They also adopt this reasoning in order to provide a scale-invariant estimator based on the number of neighbors to be considered, $\hat{m}_b$. For a more robust estimate, they suggest to average the estimates for different numbers of neighbors, b:

$$\hat{m}_b(\mathbf{x}) = \left(\frac{1}{b-1} \sum_{j=1}^{b-1} \log \left(\frac{\psi_b(\mathbf{x})}{\psi_j(\mathbf{x})} \right) \right)^{-1} \tag{3.7}$$

$$\hat{m}(\mathbf{X}) = \frac{1}{b_{\max} - b_{\min} + 1} \sum_{b=b_{\min}}^{b_{\max}} \sum_{i=1}^{n} \frac{\hat{m}_b(\mathbf{x}_i)}{n} \tag{3.8}$$

Where $\psi_i(\mathbf{x})$ is the dissimilaritiy of $\mathbf{x}$ to the i-th least dissimilar other object (i-th neighbor of $\mathbf{x}$).

Levina and Bickel note that their estimator can be misled by "edge effects" that they assume to become even more severe in higher dimensions. Consider again the idea of a growing hypersphere as a neighborhood. A 2-dimensional example, where the hypersphere is a circle, is shown in the left plot of figure 3.12. As long as the neighborhood remains within the sampling window of the model distribution, the expected number of objects grows in accordance with the area of a disk. However, this rate of growth decreases as the neighborhood starts to extend outside of the sampling window (depicted as circle segments in fig. 3.12). As the rate decreases, the estimator finds a relatively high number of objects within the neighborhood relative to the current rate, which would, under normal conditions, be characteristic for a lower dimensional space. Note that it is this change of the rate that misleads the estimator. Objects that lie directly on the border are not affected, as the estimator interprets the smaller but steady rate as a lower denseness of the model distributions. A measurement of the edge effect is shown on the right hand

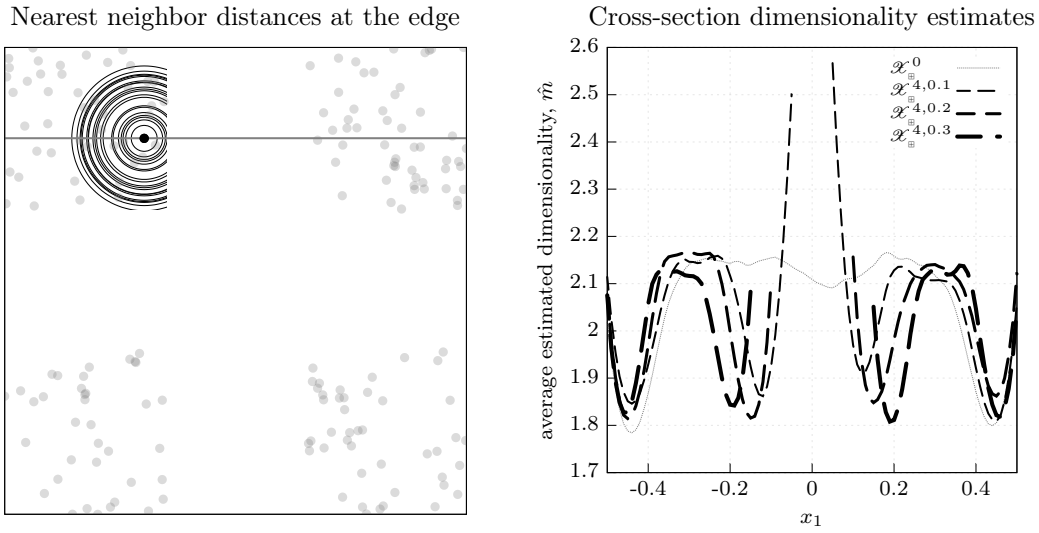

Figure 3.12: (*Left*) Model sampled from $\mathscr{X}_{\boxplus}^{4,0.3}$ (cf. experiment 3.1) with one object highlighted. The circles correspond to the neighborhoods that contain the b nearest neighbors for $b \in \{1, \ldots, 20\}$ and are clipped by the sampling window of $\mathscr{X}_{\boxplus}^{4,0.3}$. (*Right*) Estimates of dimensionality for different x_1 coordinates in the cross-section at $x_2 = 0.35$ that is shown in the left plot (horizontal line). For each estimate ($x_1 \in \{-0.50, -0.49, -0.48, \ldots, 0.50\}$), an object was created at the corresponding coordinates and its average dimensionality with respect to the models from the model distributions was measured ($b_{\min} = 10, b_{\max} = 20$).

side of figure 3.12. The steep increase in the dimensionality estimate close to the central border for models from $\mathscr{X}_{\boxplus}^{n,0.1}$ can be explained by the reverse effect as the neighborhood starts to reach into the adjacent cluster.

We want to note that dissimilarities of 0 can cause problems for this estimator. This might be unsurprising, if one considers that the estimator assumes a uniform distribution in which the probability of an exact 0 dissimilarity is 0. However, due to the necessary quantization for electronical storage, dissimilarities of 0 are possible. If single dissimilarities are 0, it is reasonable to use $\hat{m}_b(\mathbf{x}) = 0$ for the affected objects instead. This essentially performs a substitution of $x/0$ by ∞. However, if there are $b_{\min}$ objects that all have a pairwise dissimilarity of 0, $\hat{m}_b$ for any of these objects becomes ∞ which also leads to an estimated total dimensionality of ∞. The estimator is thus not usable in this case.

Since a higher clusterability is thus indirectly related to a lower dimensionality, the sign of the estimator has to be inverted for its usage as a clusterability index. In order to show the difference to the expected (true) dimensionality m,

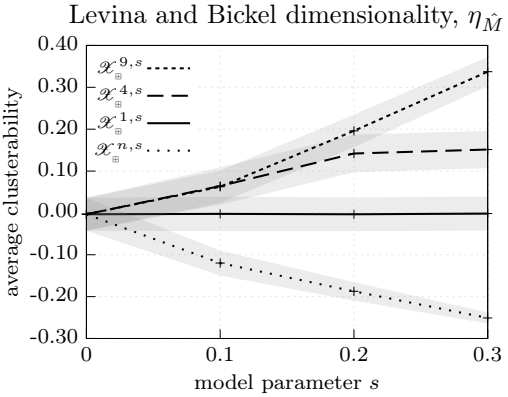

Figure 3.13: Average $\eta_{\hat{M}}$-score of models from the model distributions of experiment 3.1 ($b_{\min} = 10$, $b_{\max} = 20$ and $m = 2$). The shaded regions cover the area from 1 standard deviation below to 1 standard deviation above the averages.

we propose to subtract $\hat{m}(\mathbf{X})$ from this value if it is known:

$$\eta_{\hat{M}}(\mathbf{X}) = m - \hat{m}(\mathbf{X}) \tag{3.9}$$

The clusterability index $\eta_{\hat{M}}$ is scale invariant and robust in the weak sense. The estimator contains some bias for a higher dimensionality for too few objects (Levina and Bickel, 2004) and the extent of the edge effects depend on the dimensionality and the number of objects, as well. Therefore, the clusterability index is neither $\mathscr{X}_{\boxplus}^{0}$-normalized nor absolute (def. 3.6, 3.7). Indeed, it is mere coincidence that the bias and the edge effect negate each other for the uniform models in our experiment and result in an average score very close to 0. Due to the scale invariance of $\hat{m}_b$ (cf. equation 3.7), $\eta_{\hat{M}}$ is scale invariant, too (def. 3.4). Furthermore, the averaging of the single object-based estimates in equation 3.8 limits the effect a change in one dissimilarity can have on the $\eta_{\hat{M}}$-score.[5] Therefore, $\eta_{\hat{M}}$ is *weak* robust (def. 3.5). However, if the change can affect all dissimilarities of one object, they can all be set to a value very close to 0 such that it is nearest neighbor to every other object. As this change effects all $\hat{m}_b$-estimates, $\eta_{\hat{M}}$ is not *strong* robust.[6]

The average scores of the models from experiment 3.1 are shown in figure 3.13. We use the values $b_{\min} = 10$ and $b_{\max} = 20$ that are also employed in the experiments of Levina and Bickel (2004). The spaced models sampled from $\mathscr{X}_{\boxplus}^{n,s}$ have a lower score than the uniform models. This is because the forced free region around every object leads to more regularly packed objects,

[5]This actually assumes that it is not allowed to set the dissimilarity to 0 in order to make $\hat{m}_b$ go to ∞ (see above).

[6]We assume that $b_{\max}$ does not depend on the total number of objects.

Table 3.1: Overview of which clusterability index (section 3.2) fulfills which property (section 3.1). Permutation invariance is required for all clusterability indices, the others are optional. Distribution normalization and absoluteness are neither proven nor disproven for η_E.

	property				
index	perm. inv.	scale inv.	robust	$\mathscr{X}_{\boxplus}^{0}$-norm.	absolute
η_D	yes	yes	no	no	yes
η_Δ	yes	yes	no	no	no
$\eta_{H^\blacksquare}$	yes	no	no	yes	no
$\eta_{H^\square}$	yes	yes	no	yes	no
η_E	yes	yes	strong		
$\eta_{\hat{M}}$	yes	yes	weak	no	no

which has the effect that the dissimilarities to the nearest neighbors are very similar for most objects (cf. fig. 3.10). This, in turn, is a characteristic of high-dimensional spaces. It can also be seen in the case of models from $\mathscr{X}_{\boxplus}^{4,s}$ that, once neighborhoods do no longer reach into adjacent clusters, the $\eta_{\hat{M}}$-score stops to increase.[7] Furthermore, because models from $\mathscr{X}_{\boxplus}^{9,s}$ have more edges than those from $\mathscr{X}_{\boxplus}^{4,s}$, $\eta_{\hat{M}}$ assigns a higher score to the former.

Summary Different approaches for measuring the clusterability of models exist. We categorized them as being based on optimum clusterings, lacking structure and the concentration of dissimilarities. A solid formalization of consistency for clusterability indices is still an open topic. In this regard, we want to point out that $\eta_{\hat{M}}$ is the only clusterability index we analyzed that assigns a higher score to models from the more fine-grained $\mathscr{X}_{\boxplus}^{9,s}$ than to those from $\mathscr{X}_{\boxplus}^{4,s}$. Besides consistency, we formulated 5 other properties for clusterability indices. A summary of which clusterability indices fulfill which property is given in table 3.1.

[7] This effect is also visible in the comparison of clusterability indices in figure 3.4.

Chapter 4

Clusterability of Real-world Data

Are the most clusterable models also the most meaningful? Based on the considerations in chapter 2, especially in section 2.1.3, it appears intuitive that the most meaningful models—in which the clusters of the ground-truth are the most evident—are also clusterable. This intuition encourages the use of clusterability indices in model selection, where it is desired to select meaningful models without knowledge of the ground-truth. However, the inversion of the argument, that clusterability implies meaningfulness, does not hold in general. This chapter presents first empirical results that show that these intuitions can nevertheless be valuable in practice and discusses when they are reasonable.

In order to take a look at the correlation of the clusterability and meaningfulness of models, we consider different models of the same dataset. In detail, we generate different models of datasets with multiple attributes by removing some of the attributes and keeping the others. This corresponds to the common model transformation by orthogonal projection (cf. section 2.3.2).

We selected 2 real-world datasets with a reasonable number of attributes (7) such that we are able to consider all possible attribute sets (127). We use the *seeds* and *abalone* datasets, which are freely available from the UCI Machine Learning Repository (Bache and Lichman, 2013). The models of the seeds dataset that we analyze are more meaningful than the models of the abalone dataset. A comparison of the results provides thus insight on the relationship of clusterability and meaningfulness under different circumstances.

Clusterability and meaningfulness are measured for each model. We employ the clusterability indices detailed in section 3.2 to rank the different models by their clusterability. Furthermore, through the use of the ground-truth, we also rank the models by their meaningfulness to the dataset.

Section 4.1 details the setup of the experiments and the employed measures and methods, while section 4.2 analyzes the datasets which are used. The results are then shown in section 4.3 and discussed in section 4.4.

4.1 Experiment Setup

The experiments compare the relative scores of 6 clusterability indices and 5 indices of model meaningfulness on models of 2 datasets with up to 7 attributes each. For each dataset, we analyze different models that employ different attribute sets. The 7 attributes allow for 127 different attribute sets ranging from 7 sets with 1 attribute to 1 set with all 7 attributes. The score of each index is calculated for each attribute set.

In order to compare the indices, we consider their pairwise correlation for the different attribute sets both visually and numerically. Scatter plot matrices allow for a visual comparison of 2 indices. A scatter plot shows, in this case, each attribute set as a point in a 2-dimensional coordinate system where the coordinates correspond to the scores with respect to 2 different indices. A scatter plot matrix contains scatter plots for all combinations of attributes in a grid layout. The single plots are arranged such that the column in the grid layout determines the index that is used for the x_1 coordinate (horizontal) while the row determines the index for the x_2 coordinate (vertical). The scatter plots are scaled such that the shown range is limited by the smallest and largest score for each axis respectively. Because of this, the scatter plot is scale invariant with respect to the scores. For the numerical assessment we employ the Pearson and Spearman correlation coefficients that are detailed below.

The standard measure for linear correlation is the product-moment correlation coefficient that is usually referred to as Pearson's (Everitt, 2002). The coefficient is calculated for two variables $\mathbf{Y}$ and $\mathbf{Y}'$, where each observation y_i of $\mathbf{Y}$ is related to the observation y'_i of $\mathbf{Y}'$. In the setting of our experiment, the observations y_i and y'_i correspond to the scores of the same model with respect to different indices. On the other hand, the observations y_i and y_j for $i \neq j$ are the scores of different models with respect to the same index. Let $\bar{y}$ be the mean of the observations in $\mathbf{Y}$ and $\bar{y}'$ be defined likewise with respect to $\mathbf{Y}'$. The Pearson or product-moment correlation of the 2 variables is then defined by

$$\frac{\sum_{i=1}^{n}(y_i - \bar{y}) \cdot (y'_i - \bar{y}')}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2} \cdot \sqrt{\sum_{i=1}^{n}(y'_i - \bar{y}')^2}}$$

The correlation takes values between -1 (strong negative correlation) and 1 (strong positive correlation). With respect to scatter plots, the points (y_i, y'_i) of strong positive (negative) correlated variables lie approximately on a straight upward (downward) line. Note that, like for scatter plots, the scale of the indices is of no importance as they are normalized by the mean ($\bar{y}$) and standard deviation (denominator) of the observations.

The approach of rank correlation does not consider the observations directly, but instead uses their rank within the observations of the same variable. In detail, consider the observations of a variable $\mathbf{Y}$ in sorted order.[1] The rank of an observation is then its place in this sorted order. The rank correlation that is known as Spearman's is defined as the product-moment correlation of the observation ranks (Everitt, 2002). Trivially, the rank correlation has the same range as the product-moment correlation and a similar interpretation. However, where the product-moments require a linear relationship for maximum absolute correlation, the rank correlation only require a monotone relationship. If the points (y_i, y_i') lie on a monotone increasing (decreasing) line in a scatter plot, their ranks have a linear increasing (decreasing) relationship and the correlation is thus close to 1 (-1).

The rank correlation is better suited for our experiments as we are more interested in a relative agreement on which models are more clusterable than in a direct linear relationship. Therefore, we focus on the rank correlation coefficients in section 4.3.

We continue with a brief listing of the indices we use in our experiments. If an index is not defined in chapter 2 or chapter 3, we give a short description and refer to further sources. The first indices are the clusterability indices analyzed in section 3.2. For completeness, section 4.1.1 lists these again and states specific parameter values. After that, we detail indices which measure the meaningfulness of a model with respect to the ground-truth of the datasets as mentioned in section 2.1.3. There are 2 methodologies for assessing the meaningfulness of a model: (1, section 4.1.2) measuring the fit of the model to the ground-truth and (2, section 4.1.3) measuring the similarity of the ground-truth and an optimum clustering with respect to the model.

4.1.1 Clusterability Indices

We employ the clusterability indices based on the Dunn index and the minimum spanning tree, η_D, and the index η_Δ by Ostrovsky et al. (2006) (section 3.2.1). For η_D, we only consider clusterings with a number of clusters less than $\sqrt{n}$. For η_Δ, we use $n_{\max} = 10$ and 1 000 initial centroid sets per evaluation.

For the index based on the statistic by Hopkins and Skellam (1954), $\eta_{H^\square}$, we use $l = 1\,000$ re-samples and the highest value r such that $r < n/10$ as mentioned in section 3.2.2. For m we use the number of attributes of the particular model. The sampling window is estimated from the model as the smallest hyperrectangle that contains all objects.

[1] With regard to correlation it is unimportant if the observations are sorted in ascending or descending order as long as the same order is used every time.

We also use the "entropy" index based on the measure by Dash et al. (1997), η_E, and the index based on the intrinsic dimensionality estimator by Levina and Bickel (2004), $\eta_{\hat{M}}$ (section 3.2.3). We use $\eta_{\hat{M}}$ as defined in equation 3.9 with m as the number of attributes of the current model. Additionally, we show the scores for $m = 0$. We denote this clusterability index by $\eta_{\hat{m}}$. For both indices, we use $b_{\min} = 10$ and $b_{\max} = 20$.

4.1.2 Internal Evaluation by Ground-truth

Evaluation indices for internal cluster evaluation measure the fit of a clustering $\mathcal{C}$ to the dissimilarities in a model $\mathbf{X}$ (cf. section 2.1). The usual application assumes a fixed model and evaluates the clustering. However, when the model is to be selected and the clustering represents the ground-truth of the dataset, the same approach can be used to evaluate the model (cf. section 2.1.3). With regard to the experiments, we use the 3 internal indices which are detailed below.

The different indices base on different notions of cohesiveness and separation. Therefore, they do not necessarily agree with each other on the meaningfulness of models. This is an important observation that makes further considerations on the employed clusterability and evaluation indices necessary. In detail, clusterability indices and evaluation indices which are based on rather different notions of cohesiveness and separation are unlikely to correlate.

Dunn index The index by Dunn (1974), ν_{D}, as defined by equation 3.1 on page 35 with measures of between- and within-cluster dissimilarity given by equation 3.2.

Note that ν_{D} is susceptible to noise objects and outliers as both have a significant effect on the minimum spanning tree. It should also be noted that $\eta_D(\mathbf{X}) = \max_{\mathcal{C}}(\nu_{\mathrm{D}}(\mathcal{C}, \mathbf{X}))$. Therefore, when the ground-truth is nearly optimal and the model is thus a meaningful representation of the dataset, one can conjecture that $\eta_D(\mathbf{X}) \approx \nu_{\mathrm{D}}(\mathcal{C}_{\mathrm{gt}}, \mathbf{X})$. As shown in the experiments, this is not the case for the models we consider. This is because the clusters of the ground-truth are not separated (as defined by η_D/ν_{D}) for any of the models.

Residual sum of squares This index is the sum of the squared distances of each object to the center of its cluster.

$$\nu_{\mathrm{RSS}}(\mathcal{C}, \mathbf{X}) = \frac{\mathrm{RSS}(\{\mathbf{X}\}, \mathbf{X}) - \mathrm{RSS}(\mathcal{C}, \mathbf{X})}{(\bar{\psi}(\mathbf{X}))^2}$$

Where $\mathrm{RSS}(\mathcal{C}, \mathbf{X})$ is the residual sum of squares as defined in equation 3.4, $\{\mathbf{X}\}$ is the clustering of the dataset that contains only 1 cluster and $\bar{\psi}(\mathbf{X})$ is

the average dissimilarity of the model (cf. equation 3.6).

This index is related to η_Δ by Ostrovsky et al. (2006) which is based on $\mathrm{RSS}(\mathcal{C}, \mathbf{X})$ for different $\mathcal{C}$ (cf. section 3.2.1). RSS assigns a high score to clusters for which the objects are all close to each other. Although RSS does not directly evaluate separation, ν_{RSS} does to some extent through the comparison to the clustering that contains only 1 cluster.

Expected Density This evaluation index is proposed by Stein et al. (2003) and compares the overall cohesiveness of the model to the cohesiveness of the clusters. Stein et al. use the term "density" instead of cohesiveness due to the derivation of the index from graph models.

$$\nu_{\mathrm{ED}}(\mathcal{C}, \mathbf{X}) = \sum_{i=1}^{k} \frac{|C_i|}{n} \cdot |C_i|^{\theta_i - \theta} \qquad \theta = \frac{\log(\mathrm{weight}(\mathbf{X}))}{\log(n)} \qquad \theta_i = \frac{\log(\mathrm{weight}(C_i))}{\log(|C_i|)}$$

Where θ and the θ_i are called the densities of the model and of the clusters respectively, $\mathrm{weight}(C) = |C| + \frac{1}{2} \cdot \sum_{\mathbf{x} \neq \mathbf{x}' \in C} \varphi(\mathbf{x}, \mathbf{x}')$ and $\mathrm{weight}(\mathbf{X})$ is the same for the cluster that contains all objects of the dataset. In order to convert the dissimilarities of the model (ψ) into similarities (φ), we use a conversion scheme that normalizes with respect to the largest dissimilarity in the model: $\varphi(\mathbf{x}, \mathbf{x}') = 1 - \psi(\mathbf{x}, \mathbf{x}') \cdot (\max_{\mathbf{x}'', \mathbf{x}''' \in \mathbf{X}}(\psi(\mathbf{x}'', \mathbf{x}'''))^{-1}$.

The expected density index, ν_{ED} is similar to ν_{RSS} in that it compares the overall dissimilarities (or similarities) to the dissimilarities within the clusters.

4.1.3 External Evaluation Indices

Indices for external cluster evaluation measure the similarity of a clustering $\mathcal{C}$ to the ground-truth (which is represented by a clustering, cf. section 2.1.3). There are 2 factors that influence the search for a clustering in cluster analysis: the model and the clustering algorithm. If the found clustering is very similar to the ground-truth, one can assume both that the model is a good representation of the dataset and that the assumptions made by the algorithm (cf. section 2.2) are suitable for the model.

Like for internal indices (cf. section 4.1.2), we compare the results of different clustering algorithms. In this case, there are 2 different influences that have to be considered: the clustering relation and the search strategy of the algorithm. The clustering relation specifies the notions of cohesiveness and separation that are employed by the clustering algorithm (def. 2.17). The search strategy affects which clusterings are considered by the algorithm. For our experiments, we want to focus on the relationship of clusterability and meaningfulness based on notions of cohesiveness and separation. Therefore, we try to keep the effect of the search strategy at a minimum (see below).

In order to compute the similarity of 2 clusterings we employ the Rand index (Rand, 1971) defined by

$$\nu^e_{\text{Rand}}(\mathcal{C}, \mathcal{C}') = \sum_{i=1}^{n} \sum_{j=i+1}^{n} \frac{2 \cdot \delta_{i,j}(\mathcal{C}, \mathcal{C}')}{n \cdot (n-1)}$$

Where $\delta_{i,j}(\mathcal{C}, \mathcal{C}')$ is 1 if and only if both clusterings agree on the placement of objects $\mathbf{x}_i$ and $\mathbf{x}_j$ either in different or a common cluster. More formally,

$$\delta_{i,j}(\mathcal{C}, \mathcal{C}') = \begin{cases} 1 & \text{if } (\exists_{C \in \mathcal{C}}(\mathbf{x}_i \in C \wedge \mathbf{x}_j \in C)) \Leftrightarrow (\exists_{C' \in \mathcal{C}'}(\mathbf{x}_i \in C' \wedge \mathbf{x}_j \in C')) \\ 0 & \text{else} \end{cases}$$

Note that the score of the Rand index lies between 0 (complete disagreement) and 1 (identical clusterings up to a permutation of clusters).

The 2 indices of this type that we consider in our experiments use the clustering algorithms K-Means (γ_{km}) and Single-link (γ_{sl}) respectively:

$$\nu_{\gamma_{\text{km}}}(\mathcal{C}, \mathbf{X}) = \nu^e_{\text{Rand}}(\mathcal{C}, \gamma^k_{\text{km}}(\mathbf{X})) \qquad \nu_{\gamma_{\text{sl}}}(\mathcal{C}, \mathbf{X}) = \nu^e_{\text{Rand}}(\mathcal{C}, \gamma^k_{\text{sl}}(\mathbf{X}))$$

Both clustering algorithms are parameterized to only consider clusterings with the same number of clusters as the ground-truth.

K-Means A short description of K-Means can be found on page 38. More details are provided by MacQueen (1967, chap. 3.6). Identical to the use of the algorithm within η_Δ in section 3.2.1, we repeat for each model evaluation K-Means 1 000 times with different initial centroids and choose the clustering with the smallest residual sum of squares. Additionally, we initialized K-Means with the ground-truth which minimizes the effect of K-Means search strategy on the results as the algorithm is guaranteed to consider the ground-truth. However, the results are nearly identical in both cases. We thus only show the results for the η_Δ-like procedure.

Single-link The Single-link algorithm starts with all objects in an own cluster and iteratively merges those clusters with the smallest dissimilarity as given by $\Psi_b(C, C') = \min_{\mathbf{x} \in C, \mathbf{x}' \in C'} \psi(\mathbf{x}, \mathbf{x}')$ (cf. equation 3.2). The procedure stops when the desired number of clusters are left.

Since Single-link is guaranteed to find the global optimum clustering with respect to its clustering relation, its search strategy can be neglected in the discussion of the results. Lance and Williams (1967) discuss Single-link and related algorithms in more detail. As the definition of the between-cluster dissimilarity suggests, $\nu_{\gamma_{\text{sl}}}$ is related to η_D and ν_{D}.

Table 4.1: Summary of the processed seeds and abalone datasets. All attributes are real-valued. The clusters of the seeds dataset correspond to the wheat types Kama, Rosa and Canadian. For the abalone dataset, the clusters correspond to young, average-aged and old abalone respectively. m is the number of attributes, n is the total number of objects and $|C_i|$ is the number of objects in cluster i.

Dataset	m	n	$\lvert C_1\rvert$	$\lvert C_2\rvert$	$\lvert C_3\rvert$
Seeds	7	210	70	70	70
Abalone	7	2 000	690	625	685

4.2 The Datasets

For the experiments, we selected the seeds and abalone datasets that are available for public access from the UCI Machine Learning Repository (Bache and Lichman, 2013). We normalized both datasets such that the values of every attribute have a mean of 0 and a standard deviation of 1, which weighs every attribute equal. A summary of the datasets is shown in table 4.1.

The seeds dataset was collected for the cluster analysis study by Charytanowicz et al. (2010). They showed that the K-Means clustering algorithm (see above) is able to group seeds based on their wheat type. The three different types which are present in the dataset are correctly distinguished in about 90% of the cases.

The abalone dataset originates from the study by Nash et al. (1994) and is originally not related to machine-learning. Nevertheless, the UCI web page cites 31 publications related to machine-learning that use this dataset.[2] The usual task is to learn a predictor for the age of abalone based on the attribute values. For our experiment we grouped the abalone into 3 clusters of about the same size based on their age: "young" (8 years and younger), "average-age" (9 or 10 years) and "old" (11 years and older). We use this grouping as the ground-truth. Since this categorization is only a discretization of a continuous feature, we expect the clusters of the ground-truth to be not separated from each other in the models, which is indeed the case. This is because the thresholds (9 and 11 years respectively) are rather arbitrary with respect to abalones in general and only motivated by the age frequencies in the dataset. Since we restrict our considerations in this thesis to ratio attributes, we removed the categorical sex attribute. We only use a random sample of 2 000 of the originally 4 177 objects in the experiments in order to speed them up.

[2] `http://archive.ics.uci.edu/ml/datasets/Abalone` (last accessed May 3, 2014)

Table 4.2: Pearson product-moment correlation by attributes. A darker cell corresponds to a higher absolute correlation. The last row shows the correlation of the attributes with the cluster assignment (encoded as attribute values 1, 2 or 3).

Seeds	A	P	C	KL	KW	AC	KGL
Area (A)	-	0.9943	0.6082	0.9499	0.9707	-0.2295	0.8636
Perimeter (P)	0.9943	-	0.5292	0.9724	0.9448	-0.2173	0.8907
Compactness (C)	0.6082	0.5292	-	0.3679	0.7616	-0.3314	0.2268
Kernel length (KL)	0.9499	0.9724	0.3679	-	0.8604	-0.1715	0.9328
Kernel width (KW)	0.9707	0.9448	0.7616	0.8604	-	-0.2580	0.7491
Asymmetry coefficient (AC)	-0.2295	-0.2173	-0.3314	-0.1715	-0.2580	-	-0.0110
Kernel groove length (KGL)	0.8636	0.8907	0.2268	0.9328	0.7491	-0.0110	-
Wheat type (ground-truth)	0.9086	0.9049	0.5907	0.8483	0.8924	-0.3113	0.7529

Abalone	L	D	H	WW	SKW	VW	SLW
Length (L)	-	0.9855	0.7717	0.9252	0.8995	0.9037	0.8933
Diameter (D)	0.9855	-	0.7778	0.9246	0.8935	0.9000	0.9005
Height (H)	0.7717	0.7778	-	0.7650	0.7256	0.7454	0.7595
Whole weight (WW)	0.9252	0.9246	0.7650	-	0.9665	0.9653	0.9573
Shucked weight (SKW)	0.8995	0.8935	0.7256	0.9665	-	0.9292	0.8800
Viscera weigth (VW)	0.9037	0.9000	0.7454	0.9653	0.9292	-	0.9085
Shell weight (SLW)	0.8933	0.9005	0.7595	0.9573	0.8800	0.9085	-
Age (ground-truth)	0.5866	0.6065	0.5164	0.5874	0.4976	0.5721	0.6331

The attributes and their pairwise product-moment correlations (cf. section 4.1) are shown in table 4.2. As can be seen from the table, most attributes have a strong positive correlation. Moreover, it should be noted that the compactness attribute of the seeds dataset is directly related to other attributes (Charytanowicz et al., 2010): $\text{Compactness} = 4 \cdot \pi \cdot \text{Area} \cdot \text{Perimeter}^{-2}$. The only attribute that seems to be unrelated to the others is the asymmetry coefficient in the seeds dataset.

The tables also show that some of the attributes are correlated with the assignment of objects to clusters. For the abalone dataset we used an encoding of 1 = "young" to 3 = "old" as attribute values. Although a natural ordering of wheat seeds seems not obvious, we noticed that indeed most attributes correlate with a specific ordering: 1 = "Canadian," 2 = "Kama" and 3 = "Rosa." The strong correlation with some of the attributes facilitates the clustering of the seeds dataset with clustering algorithms that segment the object space, like K-Means. The strong correlation of some attributes with each other and with the cluster assignment can also be seen in the top left plot of figure 4.1. The top right plot, on the other hand, shows the 2 attributes of the seeds

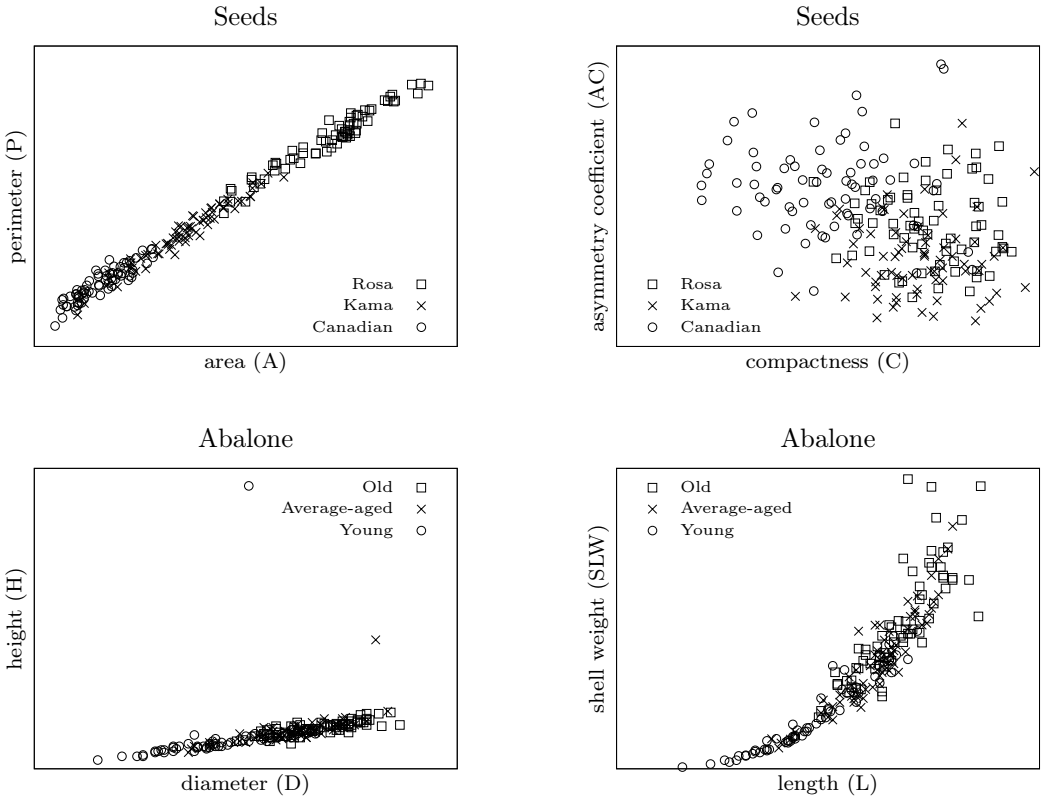

Figure 4.1: Plots of the objects of the 2 datasets for different attribute combinations. In the case of the abalone dataset, the 2 outliers regarding the height and 198 randomly sampled other objects are shown. The assignment to clusters in the ground-truth clustering is depicted by the different symbols.

dataset for which this is not the case. Some attributes for the abalone dataset are shown in the bottom plots. We want to note that the attribute H contains 2 outliers. While these would be removed in a normal cluster analysis, we keep them in order to study their effect on the clusterability indices.

The circumstances detailed above affect the indices differently based on their notion of separation. There is no empty space between the clusters of the ground-truth for any of the considered models. Thus, the ground-truth clusters are not separated when separation is based on the minimum dissimilarity between objects in different clusters (η_D, ν_{D}, $\nu_{\gamma_{\mathrm{sl}}}$). Although the clusters "overlap," they do only marginally so in the case of the seeds but more so in the case of the abalone dataset. This explains the relatively good performance of K-Means reported by Charytanowicz et al. (2010). This evidence suggests that models of the seed dataset might be meaningful with respect to a less strict notion of separation (η_Δ, η_E, $\eta_{H^\square}$, ν_{RSS}, ν_{ED}, $\nu_{\gamma_{\mathrm{km}}}$).

4.3 Results

We first highlight some general observations and then continue with the results on the seeds and abalone datasets. After that, we note the results for models with a fixed number of attributes (4). The results are shown in the scatter plot matrices in figure 4.2 (seeds) and figure 4.3 (abalone) and the tables 4.3 (rank correlation) and 4.4 (rank correlation, only models with 4 attributes).

Number of attributes A general comparison of models with different numbers of attributes is difficult as the indices are not absolute. In the scatter plot matrix of figure 4.2, it can be noted that for some of the indices, models with only 1 attribute (depicted as circles) are separated from the other models. This is especially the case for $\eta_{H^\square}$ (low score) and $\eta_{\hat{m}}$ (high score). It is to some extent also the case for η_E (high score). In the case of $\eta_{\hat{m}}$, the estimated intrinsic dimensionality is close to the real dimensionality of 1, which results in a relatively high score. This shows the natural bias of $\eta_{\hat{m}}$ for models with fewer attributes, which is the reason we also consider $\eta_{\hat{M}}$. On the other hand, the score with respect to $\eta_{H^\square}$ is relatively low for models with only 1 attribute as there are no large empty regions in this case.

Another effect that occurs for low-dimensional models in the case of the abalone dataset is that a significant number of object pairs with a dissimilarity of 0 exist. This is because the measurements on the abalones use only between 3 and 4 significant digits which results in identical object vectors for up to 2-dimensional models. If a certain part of the pairwise dissimilarities are 0, $\eta_{\hat{M}}$ and $\eta_{\hat{m}}$ can not be employed (cf. section 3.2.3). A possible fix to this problem is to add some small random value to the measurements. However, the effect of such a modification on the clusterability scores has not yet been studied and we thus instead simply omit these cases in our considerations.

Although these effects show the difficulty of comparing the (non-absolute) scores on models with different numbers of attributes, still some noticeable correlations of indices exist in these cases. We also verified the results by only considering models with a common number of attributes. We found that the exact number of attributes has some but no large effect on the results for all indices but $\nu_{\gamma_{\mathrm{sl}}}$. However, we assume that this special behaviour by $\nu_{\gamma_{\mathrm{sl}}}$ can be explained by the notion of separation that it employs, which is too strict for the particular models (cf. section 4.2). Because of this, we assume that $\nu_{\gamma_{\mathrm{sl}}}$ is more influenced by random variations in the models than by their meaningfulness (as defined by the clustering relation of $\nu_{\gamma_{\mathrm{sl}}}$, but which is also approximated through ν^e_{Rand}).

Seeds, 2 and more attributes The scatter plot matrix in figure 4.2 gives a visual impression on the correlation of the different indices. Note that the scatter plot also shows the scores for 1-dimensional models.

When models with only 1 attribute are excluded, one can see a rather strong positive correlation of the ν_{RSS} scores with the scores of η_Δ, $\eta_{H^\square}$, η_E and ν_{ED}. The strong correlation of some of the attributes of the seeds dataset (cf. section 4.2) has the effect that the objects lie in a relatively thin corridor when these attributes are employed (cf. fig. 4.1, top left). This explains the high scores with respect to $\eta_{H^\square}$ and η_E. Moreover, the seeds of each wheat type are concentrated in a relatively small region (about 1/3) for some of the attributes. This is the case for the attributes A, P, KL, KW and KGL but not for C and AC. Because of this, the scores for the indices that rely especially on the concentration of objects within each cluster, η_Δ, ν_{RSS} and—to some extent—ν_{ED}, are especially high for models that contain only these attributes. Indeed, the gap between models that contain the attributes AC and C and those that do not is clearly visible in the scatter plots for η_Δ and ν_{RSS}.

While the clusters of the ground-truth are rather cohesive, the cluster of Kama seeds is not that separated from the others in most combinations of attributes (cf. fig. 4.1, top). Because of this, indices that require separated clusters for a high clusterability score, η_D in our case, do not correlate with the ones mentioned above. Another effect of the missing separation is that $\nu_{\gamma_{\text{sl}}}$ produces bad results for nearly all models. Indeed, there are only 2 models for which the Single-link clustering algorithm finds clusterings that are somewhat related to the ground-truth. Even in these cases, we found that the clustering produced by Single-link separates only 1 of the 3 ground-truth clusters from the others, but combines the remaining ones. This shows the effect of the employed concept of cohesiveness and separation on the indices.

Although $\nu_{\gamma_{\text{km}}}$, which optimizes the residual sum of squares and compares the result clustering to the ground-truth, is related to η_Δ and ν_{RSS}, it shows a small negative correlation with respect to these. Thus, also for models that fit the ground-truth with respect to the residual sum of squares (high relative ν_{RSS}), the ground-truth is still not optimal under this measure ($\nu_{\gamma_{\text{km}}} < 1$). We want to highlight this point, as it shows that the different methodologies for measuring the meaningfulness of a model to the ground-truth, represented by the otherwise related methods ν_{RSS} and $\nu_{\gamma_{\text{km}}}$, do not necessarily agree.

The different behaviour of the methodologies for measuring the meaningfulness of a model suggests the question which of them is better suited for the problem. Unfortunately, we have no solid answer to this question, yet. Both methodologies have their weaknesses. For example, ν_{RSS}, ν_{ED} and ν^e_{Rand} are not absolute. However, their main weakness in this regard is with respect to different number of clusters, where this number is fixed in our experiments.

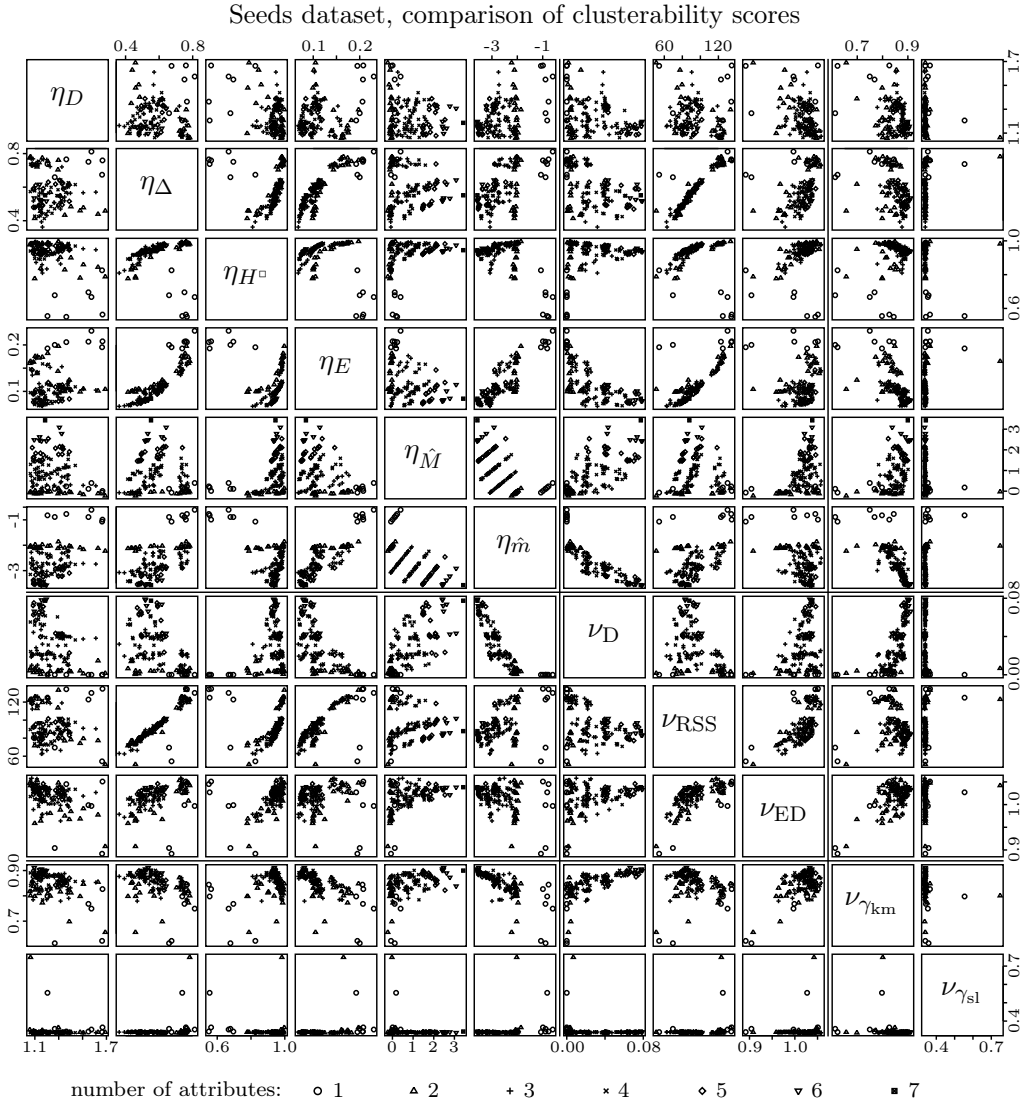

Figure 4.2: Scatter plot matrix of the different scores on models of the seed dataset with different attribute sets. All 127 possible different attribute sets are shown. The shown clusterability scores are with respect to η_D (Dunn/minimum spanning tree), η_Δ (Ostrovsky et al.), $\eta_{H^\square}$ (re-sampled Hopkins and Skellam statistic with estimated sampling window), η_E (normalized Dash et al. entropy), $\eta_{\hat{M}}$ (intrinsic dimensionality difference based on the estimator by Levina and Bickel) and $\eta_{\hat{m}}$ (negative intrinsic dimensionality). Also shows internal evaluation scores of the ground-truth based on ν_{D} (Dunn index/minimum spanning tree), ν_{RSS} (residual sum of squares) and ν_{ED} (expected density by Stein et al.). The last scores are Rand index similarities of the ground-truth to the clustering found by some clustering algorithm: $\nu_{\gamma_{\mathrm{km}}}$ (K-Means) and $\nu_{\gamma_{\mathrm{sl}}}$ (Single-link). A more detailed description of the experiment setup can be found in section 4.1.

Abalone, 3 and more attributes We proceed with an overview of the results for the abalone dataset. A visual representation is given by the scatter plot matrix in figure 4.3. The rank correlation of each pair of attributes of both datasets are shown in table 4.3.

A first significant difference to the seeds dataset is the existence of outliers with respect to 1 of the attributes (H, cf. fig. 4.1 bottom left). There are 3 indices for which the models without H are clearly separated from the models that include H with respect to their scores: η_Δ, η_D and ν_{ED}. The indices η_Δ and ν_{ED} both give a higher score to models without H, while η_D does the opposite. This is because the outliers introduce an otherwise lacking separation into the model, which is required for a high η_D-score. It can also be noted that $\eta_{H^\square}$ and η_E tend to give a higher score to models without H, as well.

In contrast to the results for the seeds dataset, η_Δ and ν_{RSS} are not correlated for the abalone dataset. An explanation for this is that there is nearly no separation between the clusters of the ground-truth for the abalone models (compare fig. 4.1 top left and bottom right). This makes the ground-truth less evident in the models and a clusterability analysis based on η_Δ thus more susceptible to random dissimilarity variations which are unrelated to the ground-truth. Moreover, presumably for the same reason, $\nu_{\gamma_{\mathrm{km}}}$ is uncorrelated to every other index (cf. table 4.3).

Both datasets, 4 attributes We report the results for the 35 models with 4 attributes for each dataset in table 4.4.

The strong correlation we noticed for η_Δ, $\eta_{H^\square}$, η_E and ν_{RSS} for the seeds dataset is even stronger when only considering models with 4 attributes. Additionally, there is a strong correlation with $\eta_{\hat{M}}$-scores, as well. Furthermore, the small negative correlation with the $\nu_{\gamma_{\mathrm{km}}}$-scores is strengthened, too. Indeed, although the notions of separation and cohesiveness on which $\nu_{\gamma_{\mathrm{km}}}$ is based are more similar to those of η_Δ and ν_{RSS} than to the one of ν_{D}, it shows a negative correlation to the former and a positive correlation to the latter.

For the abalone models, on the other hand, $\nu_{\gamma_{\mathrm{km}}}$ is practically uncorrelated to all other indices. Moreover, η_Δ, $\eta_{H^\square}$, η_E and $\eta_{\hat{M}}$ are less pairwise correlated and less correlated with ν_{RSS}. The correlation of these 4 clusterability indices with ν_{ED} is, however, similar for both datasets.

Finally, we want to point out that the correlation of η_D with the other 4 clusterability indices is at least as strong and sometimes considerably stronger for models of the seeds dataset as it is for those of the abalone dataset. Moreover, the correlation is positive in the first case and negative for the abalone dataset.

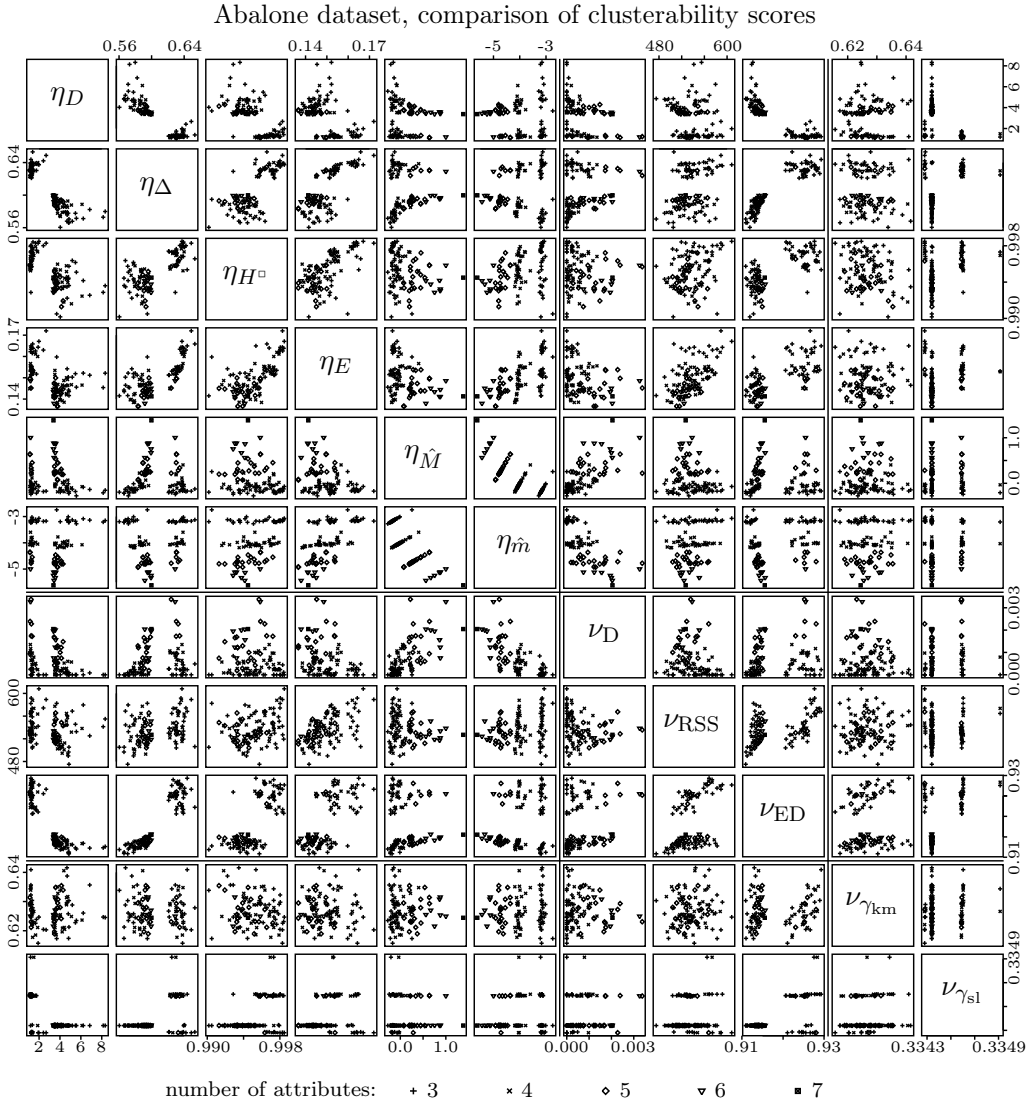

Figure 4.3: Scatter plot matrix of the different scores on models of the abalone dataset with different attribute sets. All possible different attribute sets with at least 3 attributes are shown. The shown clusterability scores are with respect to η_D (Dunn/minimum spanning tree), η_Δ (Ostrovsky et al.), $\eta_{H^\square}$ (re-sampled Hopkins and Skellam statistic with estimated sampling window), η_E (normalized Dash et al. entropy), $\eta_{\hat{M}}$ (intrinsic dimensionality difference based on the estimator by Levina and Bickel) and $\eta_{\hat{m}}$ (negative intrinsic dimensionality). Also shows internal evaluation scores of the ground-truth based on ν_{D} (Dunn index/minimum spanning tree), ν_{RSS} (residual sum of squares) and ν_{ED} (expected density by Stein et al.). The last scores are Rand index similarities of the ground-truth to the clustering found by some clustering algorithm: $\nu_{\gamma_{\mathrm{km}}}$ (K-Means) and $\nu_{\gamma_{\mathrm{sl}}}$ (Single-link). A more detailed description of the experiment setup can be found in section 4.1.

Table 4.3: Rank correlation (Spearman) of different index scores on models with different attributes for both datasets. In the calculation, the 121 models with 2 or more attributes are used for the seeds dataset and the 99 models with 3 or more attributes for the abalone dataset. Darker cells correspond to a higher absolute correlation.

Seeds

	η_D	η_Δ	$\eta_{H^\square}$	η_E	$\eta_{\hat{M}}$	$\eta_{\hat{m}}$	ν_{D}	ν_{RSS}	ν_{ED}	$\nu_{\gamma_{\mathrm{km}}}$	$\nu_{\gamma_{\mathrm{sl}}}$
η_D	-	-0.1683	-0.0903	0.0058	-0.0215	0.2126	-0.1902	-0.1628	-0.3778	-0.2230	-0.2086
η_Δ	-0.1683	-	0.9271	0.8896	0.1670	0.4614	-0.2438	0.9777	0.6929	-0.4782	-0.2801
$\eta_{H^\square}$	-0.0903	0.9271	-	0.8526	0.2642	0.4602	-0.2603	0.9445	0.6129	-0.4707	-0.2837
η_E	0.0058	0.8896	0.8526	-	-0.1402	0.7870	-0.5924	0.8903	0.4286	-0.7346	-0.2630
$\eta_{\hat{M}}$	-0.0215	0.1670	0.2642	-0.1402	-	-0.5686	0.6561	0.1697	0.4344	0.4011	-0.1742
$\eta_{\hat{m}}$	0.2126	0.4614	0.4602	0.7870	-0.5686	-	-0.8817	0.4743	-0.0976	-0.8306	-0.0253
ν_{D}	-0.1902	-0.2438	-0.2603	-0.5924	0.6561	-0.8817	-	-0.2483	0.2426	0.7351	-0.0810
ν_{RSS}	-0.1628	0.9777	0.9445	0.8903	0.1697	0.4743	-0.2483	-	0.7212	-0.4411	-0.2747
ν_{ED}	-0.3778	0.6929	0.6129	0.4286	0.4344	-0.0976	0.2426	0.7212	-	0.1425	-0.1496
$\nu_{\gamma_{\mathrm{km}}}$	-0.2230	-0.4782	-0.4707	-0.7346	0.4011	-0.8306	0.7351	-0.4411	0.1425	-	0.1879
$\nu_{\gamma_{\mathrm{sl}}}$	-0.2086	-0.2801	-0.2837	-0.2630	-0.1742	-0.0253	-0.0810	-0.2747	-0.1496	0.1879	-

Abalone

	η_D	η_Δ	$\eta_{H^\square}$	η_E	$\eta_{\hat{M}}$	$\eta_{\hat{m}}$	ν_{D}	ν_{RSS}	ν_{ED}	$\nu_{\gamma_{\mathrm{km}}}$	$\nu_{\gamma_{\mathrm{sl}}}$
η_D	-	-0.8381	-0.6120	-0.4893	-0.1746	0.0116	-0.4096	-0.1924	-0.8651	-0.0550	-0.4769
η_Δ	-0.8381	-	0.6603	0.7198	0.1432	0.1287	0.2839	0.3042	0.8659	-0.0662	0.2900
$\eta_{H^\square}$	-0.6120	0.6603	-	0.6412	-0.2640	0.3819	-0.1214	0.2837	0.6262	-0.1638	0.2067
η_E	-0.4893	0.7198	0.6412	-	-0.2206	0.6605	-0.2653	0.5524	0.6133	-0.0505	0.1902
$\eta_{\hat{M}}$	-0.1746	0.1432	-0.2640	-0.2206	-	-0.6633	0.7048	-0.0394	0.1200	0.0362	0.0287
$\eta_{\hat{m}}$	0.0116	0.1287	0.3819	0.6605	-0.6633	-	-0.6948	0.2108	0.0428	0.0542	0.0373
ν_{D}	-0.4096	0.2839	-0.1214	-0.2653	0.7048	-0.6948	-	-0.2732	0.2946	0.1084	0.2163
ν_{RSS}	-0.1924	0.3042	0.2837	0.5524	-0.0394	0.2108	-0.2732	-	0.5192	0.0954	-0.0033
ν_{ED}	-0.8651	0.8659	0.6262	0.6133	0.1200	0.0428	0.2946	0.5192	-	0.1873	0.3814
$\nu_{\gamma_{\mathrm{km}}}$	-0.0550	-0.0662	-0.1638	-0.0505	0.0362	0.0542	0.1084	0.0954	0.1873	-	0.2315
$\nu_{\gamma_{\mathrm{sl}}}$	-0.4769	0.2900	0.2067	0.1902	0.0287	0.0373	0.2163	-0.0033	0.3814	0.2315	-

Table 4.4: Rank correlation (Spearman) of different index scores on the 35 models with 4 attributes for both the seeds and abalone datasets. Since the normalized scores of $\eta_{\hat{M}}$ and $\eta_{\hat{m}}$ are identical when only models with a common number of attributes are considered, we omit $\eta_{\hat{m}}$ as it would show the same correlation values as $\eta_{\hat{M}}$. Darker cells correspond to a higher absolute correlation.

Seeds

	η_D	η_Δ	$\eta_{H^\square}$	η_E	$\eta_{\hat{M}}$	ν_{D}	ν_{RSS}	ν_{ED}	$\nu_{\gamma_{\mathrm{km}}}$	$\nu_{\gamma_{\mathrm{sl}}}$
η_D	-	0.1806	0.2918	0.2137	0.4414	-0.2425	0.2098	-0.0075	-0.4422	-0.2445
η_Δ	0.1806	-	0.9537	0.9764	0.8313	-0.4616	0.9691	0.6719	-0.7425	-0.6476
$\eta_{H^\square}$	0.2918	0.9537	-	0.9859	0.9350	-0.5834	0.9789	0.5689	-0.8159	-0.6613
η_E	0.2137	0.9764	0.9859	-	0.8918	-0.5456	0.9913	0.6322	-0.7939	-0.6718
$\eta_{\hat{M}}$	0.4414	0.8313	0.9350	0.8918	-	-0.6910	0.8879	0.4019	-0.8094	-0.5980
ν_{D}	-0.2425	-0.4616	-0.5834	-0.5456	-0.6910	-	-0.5355	0.0266	0.6513	0.3069
ν_{RSS}	0.2098	0.9691	0.9789	0.9913	0.8879	-0.5355	-	0.6647	-0.7622	-0.6560
ν_{ED}	-0.0075	0.6719	0.5689	0.6322	0.4019	0.0266	0.6647	-	-0.2813	-0.2968
$\nu_{\gamma_{\mathrm{km}}}$	-0.4422	-0.7425	-0.8159	-0.7939	-0.8094	0.6513	-0.7622	-0.2813	-	0.7010
$\nu_{\gamma_{\mathrm{sl}}}$	-0.2445	-0.6476	-0.6613	-0.6718	-0.5980	0.3069	-0.6560	-0.2968	0.7010	-

Abalone

	η_D	η_Δ	$\eta_{H^\square}$	η_E	$\eta_{\hat{M}}$	ν_{D}	ν_{RSS}	ν_{ED}	$\nu_{\gamma_{\mathrm{km}}}$	$\nu_{\gamma_{\mathrm{sl}}}$
η_D	-	-0.7070	-0.7134	-0.5498	-0.4142	-0.5937	-0.1529	-0.7288	-0.0854	-0.6532
η_Δ	-0.7070	-	0.6481	0.8840	0.5918	0.5189	0.3568	0.7876	-0.0140	0.5351
$\eta_{H^\square}$	-0.7134	0.6481	-	0.6022	0.2397	0.3611	0.3621	0.6938	-0.0719	0.3723
η_E	-0.5498	0.8840	0.6022	-	0.6403	0.2815	0.5476	0.7400	0.0226	0.3563
$\eta_{\hat{M}}$	-0.4142	0.5918	0.2397	0.6403	-	0.3398	0.0184	0.3638	0.1633	0.3815
ν_{D}	-0.5937	0.5189	0.3611	0.2815	0.3398	-	-0.2109	0.4990	0.2376	0.6426
ν_{RSS}	-0.1529	0.3568	0.3621	0.5476	0.0184	-0.2109	-	0.6462	0.1851	0.0190
ν_{ED}	-0.7288	0.7876	0.6938	0.7400	0.3638	0.4990	0.6462	-	0.3406	0.5771
$\nu_{\gamma_{\mathrm{km}}}$	-0.0854	-0.0140	-0.0719	0.0226	0.1633	0.2376	0.1851	0.3406	-	0.2610
$\nu_{\gamma_{\mathrm{sl}}}$	-0.6532	0.5351	0.3723	0.3563	0.3815	0.6426	0.0190	0.5771	0.2610	-

4.4 Discussion

We have seen some evidence that the different clusterability indices are correlated, with η_D being a special case as it adheres to a more strict notion of separation than the other indices. This difference is very obvious in our experiments as the models contain no completely separated groups.

The different notions of cohesiveness and separation affect the clusterability indices as well as the measures of meaningfulness. There is a strong correlation of the meaningfulness of a model as measured by ν_{RSS} and the clusterability as measured by all clusterability indices besides η_D for the seeds dataset. A similar correlation, but less strong, is also apparent for ν_{ED} and even for the abalone dataset. On the other hand, as the evaluation of the abalone dataset shows, correlation heavily depends on the particular models and the dataset. With respect to the considerations of section 4.2, one can conjecture that, when the models are *somewhat* meaningful representations of a dataset, clusterability analysis can help select the models that fits the dataset best. This seems intuitive, as the condition of a meaningful representation assures that an evident structure exists in the model. One can then assume that out of the meaningful models, the ones with a more evident structure (higher clusterability) are also clearer representations with respect to the ground-truth (higher meaningfulness). When considering cohesiveness and separation, one can expect that indices which employ similar notions have a higher correlation. The experiment on the seeds dataset provides some empirical evidence for this theoretical consideration, but only for measures of meaningfulness that use internal indices.

The experiments show that the different methodologies for the evaluation of model meaningfulness can lead to different results. The evaluation by means of an evaluation index measures the agreement of the ground-truth and the model. The other methodology, on the other hand, compares an "optimal" clustering and the ground-truth. The latter methodology is more strict to the notions of cohesiveness and separation which are implied by the choice of the clustering algorithm. The former methodology allows for some divergence in the notions implied by the evaluation index and the actual structure in the model as it does not focus on the optimum clustering. Therefore, we assume that the reason for the general disagreement of the methodologies is that the particular models are not completely separated representations with respect to the ground-truth.

A further difference between the methodologies is the weighting of objects with dissimilarities that do not fit to the ground-truth. The evaluation by the index of Rand (1971) gives an equal weight to every such object (as long as the clusters have about the same size), while internal evaluation indices

usually weight by the degree to which the dissimilarities disagree with the ground-truth. This observation might also be a part of the explanation why the different methodologies disagree. However, further investigations in this regard are necessary for a solid understanding of the causes and effects.

The experiments on the abalone dataset also show that outliers can have a strong influence on clusterability indices. Some indices (like η_D) can be mislead by outliers to overestimate the clusterability of the model. Other indices (like η_Δ) give a low clusterability score to models that contain outliers. It has to be decided with the particular task in mind if single objects should be able to have a large influence on the score (cf. robustness, def. 3.5).

Chapter 5

Conclusion

Clusterability is the assessment of the clustered structure in the dissimilarities of models. In chapter 2, we showed how clusterability fits into the broader context of cluster analysis as a factor in model evaluation and selection. While classical cluster evaluation asks if particular clusterings agree with the dissimilarities of a model, clusterability asks if the model is amenable to cluster analysis at all. While the former is of interest when a clustering has to be selected from a set of possible clusterings, the latter is helpful when different models are at disposal. Chapter 3 introduced the different clusterability indices and formal properties of these. Although one can approach the problem of clusterability with ideas based on cluster evaluation and clustering algorithms, other ideas based on statistical tests and the concentration of dissimilarities exist, as well. In the experiments of chapter 4, we analyzed the relationship of the clusterability of models and their meaningfulness with respect to the ground-truth of 2 real-world datasets. If multiple models are all *somewhat* meaningful representations of the dataset, the more clusterable models of them are also likely to be even better representations as they contain a more evident structure. However, when the considered models are not meaningful, the choice of a more clusterable model might actually mislead clustering algorithms as they pursue evident structure that is unrelated to the ground-truth.

In chapter 4, we analyzed clusterability for models of 2 different publicly available datasets. While the models for one of the datasets are somewhat meaningful, the models of the other dataset are less so. The clusterability indices indeed correlate with measures of model meaningfulness for the former dataset, but not for all notions of cluster-separation or methodologies for model-meaningfulness evaluation. A further analysis of the requirements for the correlation is an interesting task for future publications.

We showed that properties of evaluation indices can apply to clusterability indices, as well. Unfortunately, we were not yet able to define a property that

captures the effect of straightforward changes of models on the clusterability score. We believe that the formalization of such a definition of consistency is an important task for future work. Currently, there are many different ideas on how clusterability can be measured. A sound formalization of consistency can provide a sanity-check for these methods and is likely to be a good starting point for the development of new and specialized clusterability indices. Especially when models with thousands of dimensions are considered, which are far outside the reach of human intuition, a mathematical proof of consistency can still provide a solid justification for the application of clusterability indices.

We discussed multiple different clusterability indices and compared them on synthetic datasets that showcase their behaviour. It is interesting to note that only 1 of the analyzed indices assigned a higher clusterability score to models that have a more fine-grained structure with more clusters, but which are on the other hand less separated. It might be interesting to analyze in future work if a property can be formalized that captures this behaviour.

Further interesting possibilities for future work include the relationship of clusterability and cluster stability or constraint-set coherence. Are clusterable models stable as defined by Lange et al. (2004)? Stability implies that the structure of the model does not change for sub-models sampled from it. Intuitively, this corresponds to an evident, and maybe clusterable, structure. Another possible relationship is that of clusterability and the coherence of constraint sets. Davidson et al. (2006) noted that instance-based constraints—which specify additional knowledge in the form of which objects definitely belong to the same cluster and which do not—can mislead clustering algorithms even if they agree with the ground-truth. As such constraints modify the model, an evaluation of constraints with respect to clusterability might provide further insight on which constraints are beneficial and why.

Finally, we want to note that, for a practical application of clusterability, an assessment of the running time and memory requirements of clusterability indices is necessary. Especially for large datasets, even a runtime that grows by the square of the number of objects might be prohibitive. However, sampling strategies similar to the statistic by Hopkins and Skellam (1954) seem to be promising in this regard.

Glossary

absolute The result of a measurement or evaluation can be interpreted on its own. Can be a property of *evaluation indices* (p. 10 (def. 2.15)) or *clusterability indices* (p. 33 (def. 3.7))

attribute Some measurement taken on all objects. Each attribute defines a domain of possible values as well as a semantic meaning of these values. We assume *ratio* attributes in this thesis. 18

- **ratio** Values are real numbers ($\mathbb{R}$) and semantically scaled by multiplication with a scalar.

cluster (C_i) A subset of the *dataset* X. When the term cluster is used, it is usually assumed that the subset C is *cohesive* and *separated* from objects $\mathbf{x} \notin C$. The index i ($1 \leqslant i \leqslant k$) identifies the cluster when the associated *clustering* $\mathcal{C}$ is clear. 4 (def. 2.4)

cluster analysis The task of identifying *clusters* in a *dataset*. It encompasses the choice of a *model*, the assessment of *clusterability*, the use of a *clustering algorithm*, and the validation by *cluster evaluation*. 4 (def. 2.4)

cluster evaluation The problem of measuring the quality of a *clustering* $\mathcal{C}$ either in an *absolute* or *relative* manner. 5 (def. 2.7)

- **external** A *ground-truth* $\mathcal{C}_{\text{gt}}$ is used to define high-quality clusterings. Deviations of $\mathcal{C}$ from $\mathcal{C}_{\text{gt}}$ are associated with a loss in quality.
- **internal** The quality is measured with respect to the dissimilarities in a *model*. The model is assumed to be *meaningful* with respect to the *dataset*. A high-quality $\mathcal{C}$ contains *cohesive* and *separated* clusters.

cluster evaluation index (ν) A quality measure of a *clustering* $\mathcal{C}$. We mainly consider indices for *internal cluster evaluation*, $\nu(\mathcal{C}, \mathbf{X})$, which are *consistent* and *permutation invariant* and can be *absolute*, *distribution normalized*, *scale invariant* and *robust*. 7 (def. 2.10)

clusterability (analysis) The problem of assessing the extent to which *clusters* are evident in a *model*. 5 (def. 2.8), 22, 33

clusterability index (η) A *permutation invariant* mapping from the space of *models* to the space of real values. A higher value/score indicates a more clusterable model $\mathbf{X}$. Some indices are also *absolute*, *distribution*

normalized, *scale invariant* or *robust*. 30 (def. 3.2), 30 (def. 3.2)

clustering ($\mathcal{C}$) A mapping from the *objects* $\mathbf{x}$ of a *dataset* X to a set of *clusters* which satisfy *completeness* ($\forall_{\mathbf{x}\in X}(\exists_{C\in\mathcal{C}}(\mathbf{x}\in C))$) and *strictness* ($\forall_{C\neq C'\in\mathcal{C}}(C\cap C'=\{\})$). The number of clusters is referred to as k. 4 (def. 2.4)

locally optimal ($\hat{\mathcal{C}}_{\text{opt}}$) The *clustering algorithm* prefers it over all other $\mathcal{C}$ it compares it to. 16 (def. 2.17), 38

optimal ($\mathcal{C}_{\text{opt}}$) The *clustering algorithm* prefers it over all other $\mathcal{C}$ it can be compared to. 16 (def. 2.17)

clustering algorithm (γ) A function that takes a *model* and returns a *clustering* of the underlying *dataset*. 4 (def. 2.5), 16

clustering relation ($\leqslant_\gamma$) A partial order relation used by a *clustering algorithm* γ in order to decide which of multiple clusterings to choose. Depending on the search algorithm and cluster assumptions of γ, also not total relations can be possible. 16 (def. 2.17)

cohesive The parts (objects in a cluster) form a consistent whole. 6 (def. 2.9)

consistent A straightforward change in the input variables has no unintuitive effect on a measurement. Is defined for *evaluation indices* based on *clusterings* of the model (p. 8 (def. 2.12)).

cumulative distribution function ((c)cdf) A function that gives for a value y the probability $\Pr_Y[Y\leqslant y]$ for observations from a distribution, Y. Similarly, the complementary cumulative distribution function is defined as $\text{ccdf}(y)=\Pr_Y[Y\geqslant y]$. 14, 27

dataset (X) The set of the *objects* to be clustered. The number of objects in the dataset X is given by $|X|$ and is often referred to as n. 3 (def. 2.1)

dissimilarity function (ψ) A function that takes two *objects* $\mathbf{x}$, $\mathbf{x}'$ of the *dataset* X and returns a non-negative real number, $\psi: X\times X\mapsto\mathbb{R}^+$. Higher values indicate more dissimilar objects. Required properties are $\psi(\mathbf{x},\mathbf{x})=0$ and $\forall_{\mathbf{x},\mathbf{x}'}(\psi(\mathbf{x},\mathbf{x}')=\psi(\mathbf{x}',\mathbf{x}))$ (symmetry). 4 (def. 2.2)

distribution normalized For *models* $\mathbf{X}$ from a (specific) *distribution*, the expected value of a measurement as well as its standard deviation are analytically known. Can be a property of *evaluation indices* (p. 10 (def. 2.16)) and *clusterability indices* (p. 33 (def. 3.6)).

ground-truth A natural categorization of a *dataset* which is independent of any *model*. 3 (def. 2.1), 14, 58, 59

minimum spanning tree The smallest tree that connects all objects with the size of a tree being the sum of its edge weights. 20, 27, 35

model ($\mathbf{X}$) A representation of the *dataset* X which includes at least a *dissimilarity function* ψ defined over the *objects*. 4 (def. 2.3), 18

clusterable A model that contains evident *clusters*.

meaningful The dissimilarities reflect the *ground-truth* of the dataset.

synthetic A model that is not an actual representation of a real-world dataset, but instead drawn from a *model distribution*.

vector A type of model in which the $\mathbf{x}$ are from a object space $\mathbb{X}$, most often $\mathbb{X} = \mathbb{R}^m$. The dimensions often correspond to *attributes*.

model distribution ($\mathscr{X}$) For a *synthetic* random *model* $\mathbf{X}$ sampled from $\mathscr{X}$, $\mathbf{X} \leftarrow \mathscr{X}$, $\mathscr{X}$ specifies the distribution from which the dissimilarities for $\mathbf{X}$ are sampled. 5 (def. 2.6), 42

noise (object) An *object* that does not clearly belong to a specific *cluster*. 20 (def. 2.19)

object ($\mathbf{x}_i$) One entity with a semantic real world meaning outside the mathematical context. The index i ($1 \leqslant i \leqslant n$) identifies the object in the *dataset*. *Models* define dissimilarities between the objects. Special objects are *noise* objects and *outliers*. 3 (def. 2.1), 19

outlier An *object* that has a high dissimilarity to (nearly) every other object. 9, 19 (def. 2.18), 32

permutation invariant The ordering of the input variables is irrelevant for a measurement. Applies to all *evaluation indices* (ordering of *objects* and *clusters*, p. 8 (def. 2.11)) and *clusterability indices* (ordering of objects, p. 30 (def. 3.3)).

relative The result of a measurement or evaluation can be interpreted with corresponding results on similar entities (*models*, *clusterings*). Not *absolute*.

robust The effect on a measurement of a change in a single dissimilarity (weak) or dissimilarities of one *object* (strong robustness) of a *model* is limited by the model size. Property of some *evaluation indices* (p. 9 (def. 2.14)) and *clusterability indices* (p. 32 (def. 3.5)).

sampling window (S) For a model distribution of *vector models*, the sampling window is the subspace that contains all objects that could possibly be sampled. 42 (def. 3.8), 51

scale invariant The measurement is constant under an uniform scaling of the input variables. Property of some *evaluation indices* (p. 9 (def. 2.13)) and *clusterability indices* (p. 31 (def. 3.4)) with respect to all dissimilarities in a *model*.

separated The addition of other elements (objects not in the cluster) would have a significant negative impact on the *cohesiveness*. 6 (def. 2.9)

Bibliography

M. Ackerman and S. Ben-David. Measures of Clustering Quality: A Working Set of Axioms for Clustering. In *Proceedings of Neural Information Processing Systems (NIPS)*, pages 121–128, 2008.

M. Ackerman and S. Ben-David. Clusterability: A Theoretical Study. In *International Conference on Artificial Intelligence and Statistics*, pages 1–8, 2009.

C. C. Aggarwal. Re-designing Distance Functions and Distance-based Applications for High Dimensional Data. *SIGMOD Record*, 30(1):13–18, 2001a.

C. C. Aggarwal. On the Effects of Dimensionality Reduction on High Dimensional Similarity Search. In *Proceedings of the Twentieth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '01, pages 256–266, New York, NY, USA, 2001b. ACM.

C. C. Aggarwal and P. S. Yu. Finding Generalized Projected Clusters in High Dimensional Spaces. *SIGMOD Record*, 29(2):70–81, 2000.

C. C. Aggarwal, A. Hinneburg, and D. A. Keim. On the Surprising Behavior of Distance Metrics in High Dimensional Spaces. In *Proceedings of the 8th International Conference on Database Theory*, ICDT '01, pages 420–434, London, UK, 2001. Springer-Verlag.

R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications. In *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data*, SIGMOD '98, pages 94–105, New York, NY, USA, 1998. ACM.

K. Bache and M. Lichman. UCI Machine Learning Repository, 2013. URL `http://archive.ics.uci.edu/ml`.

A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall. Learning a Mahalanobis Metric from Equivalence Constraints. *Journal of Machine Learning Research*, 6:937–965, 2005.

K. P. Bennett, U. Fayyad, and D. Geiger. Density-based Indexing for Approximate Nearest-neighbor Queries. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '99, pages 233–243, New York, NY, USA, 1999. ACM.

K. S. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When Is "Nearest Neighbor" Meaningful? In *Proceedings of the 7th International Conference on Database Theory*, ICDT '99, pages 217–235, London, UK, 1999. Springer-Verlag.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

M. Charytanowicz, J. Niewczas, P. Kulczycki, P. A. Kowalski, S. Łukasik, and S. Żak. Complete Gradient Clustering Algorithm for Features Analysis of X-Ray Images. In *Information Technologies in Biomedicine*, Advances in Intelligent and Soft Computing, pages 15–24. Springer Berlin Heidelberg, 2010.

E. Chávez, G. Navarro, R. Baeza-Yates, and J. L. Marroquín. Searching in Metric Spaces. *ACM Computing Surveys*, 33(3):273–321, 2001.

T. F. Cox and T. Lewis. A Conditioned Distance Ratio Method for Analyzing Spatial Patterns. *Biometrika*, 63(3):483–491, 1976.

M. Dash, H. Liu, and J. Yao. Dimensionality Reduction for Unsupervised Data. In *Ninth IEEE International Conference on Tools with AI, ICTAI'97*, pages 532–539, 1997.

I. Davidson, K. Wagstaff, and S. Basu. Measuring Constraint-Set Utility for Partitional Clustering Algorithms. In *Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD-2006)*, Lecture Notes in Computer Science, pages 115–126, Berlin, Germany, 2006. Springer.

J. C. Dunn. Well-Separated Clusters and Optimal Fuzzy Partitions. *Journal of Cybernetics*, 4:95–104, 1974.

J. G. Dy and C. E. Brodley. Feature Subset Selection and Order Identification for Unsupervised Learning. In *Proceedings of the 17th International Conference on Machine Learning*, pages 247–254, 2000.

A. El-Hamdouchi and P. Willett. Techniques for the Measurement of Clustering Tendency in Document Retrieval Systems. *Journal of Information Science*, 13(6):361–365, 1987.

M. Ester, H. Kriegel, J. Sander, and X. Xu. A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *KDD*, pages 226–231, 1996.

B. S. Everitt. *The Cambridge Dictionary of Statistics*. Cambridge University Press, 2 edition, 2002.

D. Francois, V. Wertz, and M. Verleysen. The Concentration of Fractional Distances. *IEEE Transactions on Knowledge and Data Engineering*, 19(7): 873–886, 2007.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*, volume 2. Springer, 2009.

T. C. Havens, J. C. Bezdek, J. M. Keller, and M. Popescu. Clustering in Ordered Dissimilarity Data. *International Journal of Intelligent Systems*, 24(5):504–528, 2009.

B. Hopkins and J. G. Skellam. A New Method for Determining the Type of Distribution of Plant Individuals. *Annals of Botany*, 18(2):213–227, 1954.

M. E. Houle, H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek. Can Shared-neighbor Distances Defeat the Curse of Dimensionality? In *Proceedings of the 22nd International Conference on Scientific and Statistical Database Management*, SSDBM'10, pages 482–500, Berlin, Heidelberg, 2010. Springer-Verlag.

L. Hubert and P. Arabie. Comparing Partitions. *Journal of Classification*, 2: 193–218, 1985.

A. K. Jain. Data Clustering: 50 Years Beyond K-means. *Pattern Recognition Letters*, 31(8):651–666, 2010.

A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1988.

A. K. Jain, A. Topchy, M. H. C. Law, and J. M. Buhmann. Landscape of Clustering Algorithms. In *Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 1 - Volume 01*, ICPR '04, pages 260–263, Washington, DC, USA, 2004. IEEE Computer Society.

S. D. Kamvar, D. Klein, and C. D. Manning. Interpreting and Extending Classical Agglomerative Clustering Algorithms Using a Model-Based Approach. In *Proceedings of the Nineteenth International Conference on Machine Learning*, ICML '02, pages 283–290, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc.

D. Klein, S. D. Kamvar, and C. D. Manning. From Instance-level Constraints to Space-level Constraints: Making the Most of Prior Knowledge in Data Clustering. In C. Sammut and A. G. Hoffmann, editors, *Machine Learning, Proceedings of the Nineteenth Internationl Conference (ICML 2002)*, pages 307–314. Morgan Kaufmann, 2002.

M. Köppen. The Curse of Dimensionality. In *Proceedings of the 5th Online World Conference on Soft Computing in Industrial Applications (WSC5)*, 2000.

F. Korn, B. Pagel, and C. Faloutsos. On the "Dimensionality Curse" and the "Self-Similarity Blessing". *IEEE Transactions on Knowledge and Data Engineering*, 13(1):96–111, 2001.

R. Krishnapuram and J. M. Keller. A Possibilistic Approach to Clustering. *Fuzzy Systems, IEEE Transactions on*, 1(2):98–110, 1993.

G. N. Lance and W. T. Williams. A General Theory of Classificatory Sorting Strategies: I. Hierarchical Systems. *The Computer Journal*, 9:373–380, 1967.

T. Lange, V. Roth, M. L. Braun, and J. M. Buhmann. Stability-based Validation of Clustering Solutions. *Neural computation*, 16(6):1299–1323, 2004.

M. H. C. Law, M. A. T. Figueiredo, and A. K. Jain. Simultaneous Feature Selection and Clustering Using Mixture Models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(9):1154–1166, 2004.

E. Levina and P. J. Bickel. Maximum Likelihood Estimation of Intrinsic Dimension. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 777–784. 2004.

R. F. Ling and G. G. Killough. Probability Tables for Cluster Analysis Based on a Theory of Random Graphs. *Journal of the American Statistical Association*, 71(354):293–300, 1976.

J. B. MacQueen. Some Methods for Classification and Analysis of Multivariate Observations. In L. M. Le Cam and J. Neyman, editors, *Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297. University of California Press, 1967.

G. J. McLachlan. On Bootstrapping the Likelihood Ratio Test Stastistic for the Number of Components in a Normal Mixture. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 36(3):pp. 318–324, 1987.

W. J. Nash, T. L. Sellers, S. R. Talbot, A. J. Cawthorn, and W. B. Ford. The Population Biology of Abalone (Haliotis Species) in Tasmania. I. Blacklip Abalone (H. Rubra) from the North Coast and the Islands of Bass Strait. Technical Report No. 48, Sea Fisheries Division, 1994.

R. Ostrovsky, Y. Rabani, L. J. Schulman, and C. Swamy. The Effectiveness of Lloyd-type Methods for the K-means Problem. In *47th Annual IEEE Symposium on Foundations of Computer Science, 2006. FOCS'06*, pages 165–176. IEEE, 2006.

E. Panayirci and R. C. Dubes. A Test for Multidimensional Clustering Tendency. *Pattern Recognition*, 16(4):433–444, 1983.

S. M. Peres and M. L. De A. Netto. A Fractal Fuzzy Approach to Clustering Tendency Analysis. In A. L. C. Bazzan and S. Labidi, editors, *SBIA*, Lecture Notes in Computer Science, pages 395–404. Springer, 2004.

R. C. Prim. Shortest connection networks and some generalization. *Bell System Technical Journal*, 36(6):1389–1401, 1957.

J. Puzicha, T. Hofmann, and J. M. Buhmann. A Theory of Proximity Based Clustering: Structure Detection by Optimization. *Pattern Recognition*, 33 (4):617–634, 2000.

W. M. Rand. Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

S. T. Roweis and L. K. Saul. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 290:2323–2326, 2000.

J. W. Sammon, Jr. A Nonlinear Mapping for Data Structure Analysis. *Computers, IEEE Transactions on*, 100(5):401–409, 1969.

O. Shamir and N. Tishby. Cluster Stability for Finite Samples. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1297–1304. 2007.

N. Smirnov. Table for Estimating the Goodness of Fit of Empirical Distributions. *The Annals of Mathematical Statistics*, 19(2):279–281, 1948.

B. Stein and O. Niggemann. On the Nature of Structure and its Identification. In P. Widmayer, G. Neyer, and S. Eidenbenz, editors, *Graph-Theoretic Concepts in Computer Science*, Lecture Notes in Computer Science, pages 122–134, Berlin Heidelberg New York, 1999. Springer.

Benno Stein, Sven Meyer zu Eißen, and Frank Wißbrock. On Cluster Validity and the Information Need of Users. In M. H. Hanza, editor, *3rd International Conference on Artificial Intelligence and Applications (AIA 03)*, pages 216–221, Anaheim, Calgary, Zurich, Switzerland, 2003. acta pRess.

M. Steinbach, L. Ertöz, and V. Kumar. The Challenges of Clustering High Dimensional Data. In L. T. Wille, editor, *New Vistas in Statistical Physics – Applications in Econophysics, Bioinformatics, and Pattern Recognition*. Springer-Verlag, 2003.

R. Tibshirani, G. Walther, and T. Hastie. Estimating the Number of Clusters in a Data Set via the Gap Statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

V. Vinay, I. J. Cox, N. Milic-Frayling, and K. Wood. Measuring the Complexity of a Collection of Documents. In M. Lalmas, A. MacFarlane, S. Rüger, A. Tombros, T. Tsikrika, and A. Yavlinsky, editors, *Advances in Information Retrieval*, Lecture Notes in Computer Science, pages 107–118. Springer Berlin Heidelberg, 2006.

K. Wagstaff and C. Cardie. Clustering with Instance-level Constraints. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 1103–1110, 2000.

S. S. Wilks. The Large-sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62, 1938.

W. E. Wright. A Formalization of Cluster Analysis. *Pattern Recognition*, 5 (3):273–282, 1973.