Advances in Correlation Clustering

Dissertation von Daniyal Kazempour

München 2022

Advances in Correlation Clustering

Dissertation
zur Erlangung des Grades eines Doktors der Informatik
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

vorgelegt von
Daniyal Kazempour

aus
Soest

Erstgutachter:	Prof. Dr. Thomas Seidl
Zweitgutachter:	Prof. Dr. Arthur Zimek
Drittgutachter:	Prof. Michael Houle, Ph.D.
Tag der Einreichung:	22.11.2021
Tag der mündlichen Prüfung:	01.03.2022

Eidesstattliche Versicherung

Hiermit erkläre ich, Daniyal Kazempour, an Eides statt, dass die vorliegende Dissertation ohne unerlaubte Hilfe gemäß Promotionsordnung vom 12.07.2011, § 8, Abs. 2 Pkt. 5, angefertigt worden ist.

München, 19.11.2021

.....
Daniyal Kazempour

Abstract

The task of clustering is to partition a given dataset in such a way that objects within a cluster are similar to each other while being dissimilar to objects from other clusters. One challenge to this task arises when dealing with datasets where the objects are characterized by an increased number of features. Objects within a cluster may exhibit correlations among a subset of features. In order to detect such clusters, within the past two decades significant contributions have been made which yielded a wealth of literature presenting algorithms for detecting clusters in arbitrarily oriented subspaces. Each of them approaches the correlation clustering task differently, by relying on different underlying models and techniques. Building on the current progress made, this work addresses the following aspects: First, it is dedicated to the research question of how to actually measure and therefore evaluate the quality of a correlation clustering. As an initial endeavor, it is investigated how far objectives for internal evaluation criteria can be derived from existing correlation clustering algorithms. The results from this approach, however, exhibited limitations rendering the derived internal evaluation measures not suitable. As a consequence endeavors have been made to identify commonalities among correlation clustering algorithms leading to a cost function that is introduced as an internal evaluation measure. Experiments illustrate its capability to assess clusterings based on aspects that are inherent to all correlation clustering algorithms studied so far. Second, among the existing correlation clustering algorithms, one takes a unique approach. Clusters are detected in a space spanned by the parameters of a given function, known as Hough space. The detection itself is achieved by finding so-called regions of interest (ROI) in Hough space. While the detection of ROIs in the existing algorithm performs well in most cases, there are conditions under which the runtime deteriorates, especially in data sets with high amounts of noise. In this work, two different novel strategies are proposed for ROI detection in Hough space, where it is elaborated on their individual strengths and weaknesses. Besides the aspect of ROI detection, endeavors are made to go beyond linearity by proposing approaches for detecting quadratic and periodic correlated clusters using Hough transform. Third, while there exist different views, like local and global correlated clusters, explorations are made in this work with the question in mind, in how far both views can be unified under a single concept. Finally, approaches are proposed and investigated that enhance the resilience of correlation clustering methods against outliers.

Zusammenfassung

Die Aufgabe von Clustering besteht darin einen gegebenen Datensatz so zu partitionieren dass Objekte innerhalb eines Clusters ähnlich zueinander sind, während diese unähnlich zu Objekten aus anderen Clustern sind. Eine Herausforderung bei dieser Aufgabe kommt auf, wenn man mit Daten umgeht, die sich durch eine erhöhte Anzahl an Merkmalen auszeichnen. Objekte innerhalb eines Clusters können Korrelationen zwischen Teilmengen von Merkmalen aufweisen. Um solche Cluster erkennen zu können, wurden innerhalb der vergangenen zwei Dekaden signifikante Beiträge geleistet. Darin werden Algorithmen vorgestellt, mit denen Cluster in beliebig ausgerichteten Unterräumen erkannt werden können. Jedes der Verfahren verfolgt zur Lösung der Correlation Clustering Aufgabenstellung unterschiedliche Ansätze indem sie sich auf unterschiedliche zugrunde liegende Modelle und Techniken stützen. Aufbauend auf die bislang gemachten Fortschritte, adressiert diese Arbeit die folgenden Aspekte: Zunächst wurde sich der Forschungsfrage gewidmet wie die Güte eines Correlation Clustering Ergebnisses bestimmt werden kann. In einer ersten Bestrebung wurde ermittelt in wie fern Ziele für interne Evaluationskriterien von bereits bestehenden Correlation Clustering Algorithmen abgeleitet werden können. Die Ergebnisse von dieser Vorgehensweise offenbarten Limitationen die einen Einsatz als interne Evaluationsmaße ungeeignet erschienen ließen. Als Konsequenz wurden Bestrebungen unternommen Gemeinsamkeiten zwischen Correlation Clustering Algorithmen zu identifizieren, welche zu einer Kostenfunktion führten die als ein internes Evaluationsmaß eingeführt wurde. Die Experimente illustrieren die Fähigkeit, Clusterings auf Basis von Aspekten die inherent in allen bislang studierten Correlation Clustering Algorithmen vorliegen zu bewerten. Als einen zweiten Punkt nimmt ein Correlation Clustering Verfahren unter den bislang existierenden Methoden eine Sonderstellung ein. Die Cluster werden in einem Raum erkannt welches von den parametern einer gegebenen Funktion aufgespannt werden welches als Hough Raum bekannt ist. Die Erkennung selbst wird durch das Finden von sogenannten "Regions of Interest" (ROI) im Hough Raum erreicht. Während die Erkennung von ROIs in dem bestehenden Verfahren in den meisten Fällen gut verläuft, gibt es Bedingungen, unter welchen die Laufzeit sich verschlechtert, insbesondere bei Datensätzen mit großen Mengen von Rauschen. In dieser Arbeit werden zwei verschiedene neue Strategien für die ROI Erkennung im Hough Raum vorgeschlagen, wobei auf die individuellen Stärken und Schwächen eingegangen wird. Neben dem Aspekt der ROI Erkennung sind Forschungen unternommen worden um über die Linearität der Correlation Cluster hinaus zu gehen, indem Verfahren entwickelt wurden, mit denen quadratisch- und periodisch korrelierte Cluster mittels Hough Transform erkannt werden können. Der dritte Aspekt dieser Arbeit widmet sich den sogenannten "views". Während es verschiedene views gibt wie z.B. bei lokal oder global korrelierten Clustern, wurden Forschungen unternommen mit der Fragestellung, in wie fern beide Ansichten unter einem einzigen gemeinsamen Konzept vereinigt werden können. Zuletzt sind Ansätze vorgeschlagen und untersucht worden welche die Resilienz von Correlation Clustering Methoden hinsichtlich Ausreißer erhöhen.

Acknowledgements

From elementary school, over to high-school, the bachelor studies, the master studies. At all the stages there have been people of different kind, at different times, with different words and thoughts. Yet, they all have certain things in common: they made me curious, they made me grow, they made me stand up again and continue, they wanted to share time with me, they contributed to the entity I am now, writing these lines. This stage of my life - pursuing the PhD - is no exception to the rule. As in the previous stages there have been people in the role of a mentor, so I had the luck and honor to have here my doctorial supervisor, Prof. Thomas Seidl. Thank you for motivating me, guiding me, to encourage me in moments of self-doubt, challenging me to grow, to improve. I can say for sure, I did not only learn during my time as a PhD student to grow with regards to my knowledge and ideas, but also to grow my scientific skills, and my own entity. Before coming that far, to pursue my PhD, there were other people who sparked the flame of curiosity and motivation to follow my academic career in this domain of science. At this point my gratitude to Prof. Arthur Zimek and Prof. Erich Schubert, their introduction in data mining with the KDD 1 lecture made me pursue this path. Also my thanks to Prof. Peer Kröger, who guided me to take my first tiny steps towards PhD with the master thesis I wrote under his supervision. The fascination for detecting clusters in high-dimensional data has been sparkled by Prof. Matthias Schubert holding the KDD 2 lecture, making my first contact with the topic of Hough space and Hough transform possible. My gratitude goes also to the students who I supervised during my time here. I enjoyed the discussions with you, and also to see you grow with the tasks and challenges you faced. My deepest hope is that I could spark in you a passion and excitement for research and science. Besides the mentors, there are all my friends and colleagues. My gratitude to all the people of Prof. Seidl's chair. The atmosphere, the scientific exchanges and the time I spend with you will remain in my memories. Many thanks to Julian Busch for the scientific exchange and time we spend together, but also for the Hough-mug. The mug with the slogan "When times are tough - get a Hough" still got a special place on my desk. Many thanks also to Anna Beer, for the publications we wrote, for the time we had together in the same office, for the conference trips, and for sustaining sometimes my stubborn mind. Florian Richter, you and Julian and Janina Sontheim made a great beginning at this group when I started, and a great time over all the past years, thank you for that. I am also grateful to have met two more people: Franz Krojer, you made really sure that I never had any problems with the tech, and I enjoyed the time talking with you. Susanne Grienberger, you were there when I stood clueless in front of bureaucratic challenges. You also did not only teach me how to fill out forms, but to also improve myself in terms of precision, discipline and rigor. Besides the environment of the university, there is also a component in my life which was, is and will always be there: my gratitude to my mother Nadia Janani and my father Hassan Kazempour. Thank you for your patience, your comforting words, your advising words, sharing your experiences and your love. Also thanks to my grandmother for all

the prayers and best wishes. Then, there are people who have been at my side for years, experiencing me from my good and my bad sides, from my cheerful and my most sad moments, who sustained me, over all these years. Great thanks to Magnus Deininger and Nadja Deininger, Daniel Schmitt and Christian Duta. You were not only there for me, but also have a curiosity and passion for what you are pursuing, a passion which inspired me. My great admiration goes to Dr. Diana Desirée Batzer-Kaufmann and Dr. Leonie Chiara Martens. Desirée, your energy and passion for research is remarkable. My last and great gratitude is dedicated to Melanie Oelker. Melanie, thank you for all the years, the memories, your passion for science, for accepting me the way I am, for making myself reflecting and becoming a better being.

See the vast space
with all its directions
for knowledge we chase
drowning in infinite combinations

Seek out, for what matters most
look for those who belong together
features being the knowledge's host
and those who shall bond, but never

Take a glimpse from far distance
or a look close to dense local places
the binary view is inconsistent
see both, local and global faces

While you see the iterations pass
hyperplanes piece-by-piece begin to cover
the manifolds surface like lays of glass
like water wetting the surface of a flower

You see over time in the alleged entropy
the patterns that begin to show
eager to study the emerging entity
on the paths to knowledge you follow

Daniyal Kazempour (15.09.2020)

Contents

Chapter 1
Introduction

Background When we are dealing with the topic of machine learning, we face a distinction between two, not strictly disjoint, high-level categories, namely the supervised setting, for when labels are available, serving as a "ground truth" and the unsupervised setting when we have no labels at our disposal. There are different research areas in the unsupervised setting, such as e.g., frequent itemset mining or clustering. For the latter, the task is to partition a given dataset in such a way that objects within a cluster are similar to each other while objects from different clusters are dissimilar to other clusters. While clustering algorithms such as DBSCAN [10] or kmeans [29] compute clusterings based on all dimensions, questions arise like: do we need all dimensions? There are cases where scientists want to see if subsets of features exist among which groups of objects within a dataset exhibit a high similarity. Here the classic clustering task is extended by detecting subsets of features among which partitions of objects exhibit certain patterns, i.e., being highly dense. This field of research within the clustering theme addressing also high-dimensionality is coined in the literature [27] with the term *subspace clustering*. There it is distinguished between (a) axis-parallel and (b) arbitrarily oriented subspace clustering, which is also known as correlation clustering. In the case of (a), the clusters reside in subspaces where each of its features is independent of each other. In the case of (b) the features of a subspace hint at a potentially linear dependence [28]. Regarding the complexity of both areas it is notable that while the axis-parallel approach requires without heuristics a full enumeration of all possible 1 to $d-1$ tuples of subspaces, which is exponential, the arbitrarily oriented approach has an uncountable infinite number of possible subspaces. This raises inevitably the need for heuristics.

To provide a better understanding of the concept behind correlation clustering we refer to Figure 1.1 that is based on [27] . On the left hand side we can observe an affine subspace $S+a$ with $S \subset \mathbb{R}^d$ and the translation $a \in \mathbb{R}^d$. This affine subspace (sometimes also named as *relevant* or *cluster* subspace) is regarded as *interesting* if a set of objects cluster within this subspace. This cluster becomes visible when the objects are *projected* onto $S+a$ (Figure 1.1 right). Among the perpendicular subspace $(\mathbb{R}^d \setminus S) + a$ the objects may exhibit a high variance, $(\mathbb{R}^d \setminus S)+a$ is also coined with the term *correlation subspace* in the literature.

Here the term *correlation* may give rise to the question of when we are dealing with a *single* correlation within a given dataset and when we are observing *multiple* clusters that may hint at a possible correlation within the partition. The number of clusters, and therefore of potential correlations, is mediated either directly or indirectly through the choice of the parameters in a correlation clustering algorithm. In algorithms like ORCLUS [3] the number of clusters is set through the parameter k by the user. 4C [6], on the

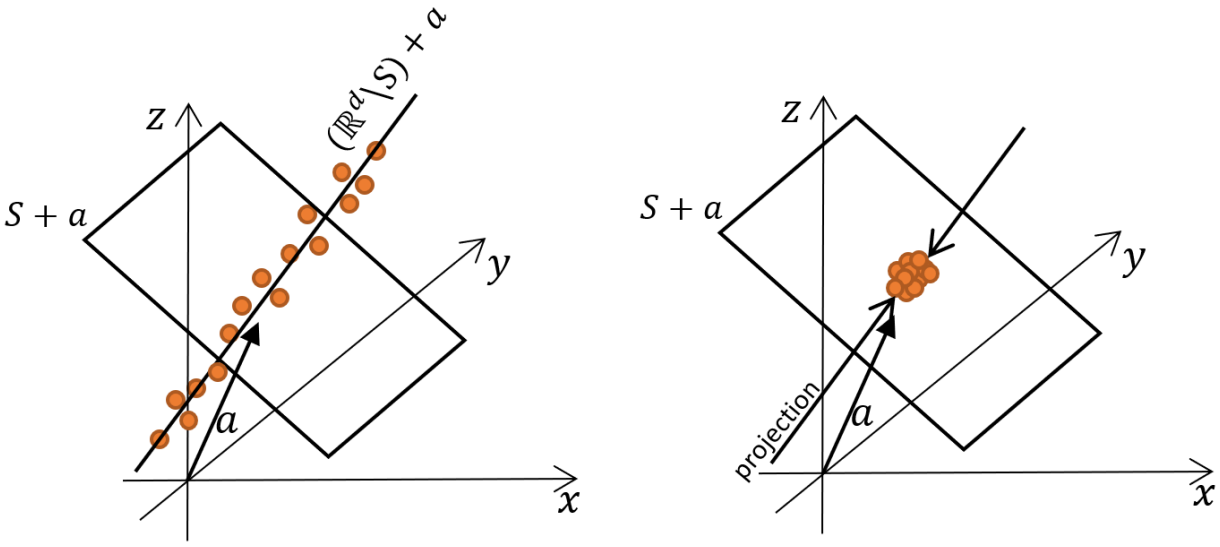

Figure 1.1: Left: illustration of the affine/cluster subspace and its orthogonal correlation subspace. Right: projection of objects onto the cluster subspace. This illustration is based on [27]

contrary, does not provide direct control of the number of clusters, but provides means to indirectly take influence on the number of clusters by specifying for example what can be considered as a sufficiently dense cluster through $minPts$ and ε parameters.

This dissertation has its focus on the sub-field of arbitrarily oriented subspace clustering. Within this sub-field, the thesis is dedicated to three categories which can be seen in Fig. 1.2, namely (i) Hough transform-based methods (Chapters 2-5), (ii) the evaluation of correlation clusterings (Chapter 6, Chapter 7), and (iii) the unification of local and global views (Chapter 8) and the impact of outliers on the detected subspaces and how to mitigate their effect (Chapter 9). Within the category of Hough transform-based methods a division is introduced into two sub-categories namely (a) detecting non-linear correlation clusters relying on Hough transform (Chapter 2, Chapter 3) and (b) detecting regions of interest (ROI) in Hough space (Chapter 4, Chapter 5). In the upcoming pages we provide a brief overview on the single chapters and, where applicable, their relations to each other.

1.1 Hough Transform based Approaches

So far, many of the existing correlation clustering algorithms are based on the Principal Component Analysis (PCA), and as such rely on the locality assumption. According to [34], the assumption is that subspaces of clusters can be derived from their local neighbor-

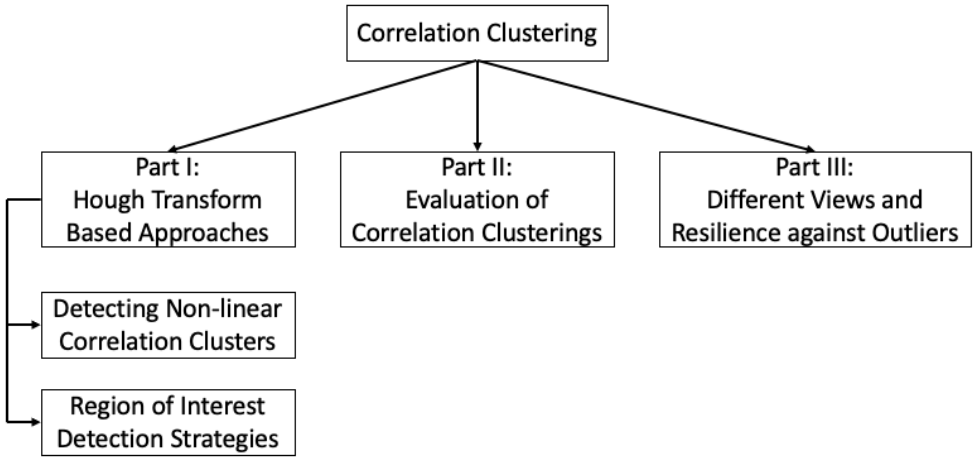

Figure 1.2: Categories of the dissertation.

hood in its full-dimensional data space[1], either through a cluster prototype or through the cluster members. As a consequence one may obtain with PCA-based correlation clustering algorithms several locally dense correlation clusters that could be described through the same linear model, i.e. through the same rotation and translation. For cases where one wants to discover correlation clusters regardless of any local density, a correlation clustering algorithm is required that is not bound to the locality assumption. One algorithm that satisfies this requirement is CASH [1] that relies on the Hough transform [16]. The idea of this approach is that objects are represented as functions in a space that is spanned by the parameters of a chosen parametrization function, known as Hough space (or among the literature also as parameter space). Identifying dense regions within the Hough space corresponds to detecting subsets of objects that share a common subspace, i.e. for the linear case, that are located on or around a hyperplane. Here the term *dense* means that a minimum of parametrization functions $minPts$ do intersect within a certain boundary within that space. Both information, the minimum number of parametrization functions intersecting, and the aforementioned boundary are user-provided through parameters. In the following we elaborate in more detail on the concept of Hough transform and to introduce the necessary definitions and properties that are adopted from [1]:

Definition 1. *Spherical Coordinates. Let $e_i, 1 \leq i \leq d$, be an orthonormal basis in a d-dimensional data space. Let $x = (x_1, ..., x_d)^T$ be a d-dimensional vector on the hypersphere of radius r with center at the origin. For the $d-1$ independent angles $\alpha_1, ..., \alpha_{d-1}$, let $\alpha_i, 1 \leq i \leq d-1$, be the angle between x and e_i. Then the generalized spherical coordinates of vector x are defined by:*

$$x_i = r \cdot \left(\prod_{j=1}^{i-1} \sin \alpha_j \right) \cdot \cos \alpha_i$$

where $\alpha_d = 0$ and $\alpha_i \in [0, \pi)$.

[1]In the literature, data space is also named as feature space or embedding space or ambient space.

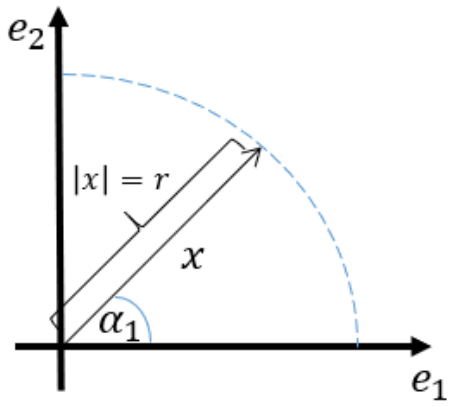

Figure 1.3: Simplified two-dimensional illustration for Definition 1.

For the following suppose we are given an object $p \in \mathcal{D} \subseteq \mathbb{R}^d$. For p there exists an infinite number of hyperplanes in data space that contain p. To be more specific, any of the hyperplanes containing p are characterized through the object p, and the $d-1$ angles $\alpha_1, ..., \alpha_{d-1}$ with $\alpha_i \in [0, \pi)$. The normal vector containing the angles $\alpha_1, ..., \alpha_{d-1}$ and a given object p can be mapped to the distance of the corresponding hyperplane to the origin through the following function named parametrization function [1]:

Definition 2. *Parametrization Function.*[2] *Let $p = (p_1, ..., p_d)^T \in \mathcal{D} \subseteq \mathbb{R}^d$ be a d-dimensional vector of a dataset $\mathcal{D}$. Further let $n = (n_1, ..., n_d)^T$ be a d-dimensional unit vector specified by $d-1$ angles $\alpha_1, ..., \alpha_{d-1}$ as stated in Definition 1. The parametrization function $f_p : \mathbb{R}^{d-1} \mapsto \mathbb{R}$ of a vector p denotes the distance of the hyperplane defined by an object p and the normal vector n to the origin:*

$$f_p(\alpha_1, ..., \alpha_{d-1}) = \langle p, n \rangle = \sum_{i=1}^{d} p_i \cdot \left(\prod_{j=1}^{i-1} \sin \alpha_j \right) \cdot \cos \alpha_i$$

With Definition 2 we have now the means to map any object $p \in \mathbb{R}^d$ to its corresponding parameter function $f_p(\alpha_1, ..., \alpha_{d-1})$ that resides in a d-dimensional parameter space $\mathcal{P}$ with the angles α_i being the arguments of that function. The semantic behind the parameter function is that it represents all possible hyperplanes that contain the object p. The parameter space $\mathcal{P}$ is spanned by the $d-1$ angles $\alpha_1, ..., \alpha_{d-1}$ of the normal vector n and their distances $\delta = f_p(\alpha_1, ..., \alpha_{d-1})$.

Having elaborated on an object p and its mapping through a parametrization function f_p we elaborate on an interesting observation which we coin with the term *Embedding-space - Hough-space Duality*. For this, we introduce four properties as stated in [1]:

[2] In some of the contributions of this dissertation parametrization function is also coined with the term *object* function.

Property 1. *An object* $p = (p_1, ..., p_d)^T \in \mathcal{D} \subseteq \mathbb{R}^d$ *in data space is represented by a sinusoidal curve* $f_p : \mathbb{R}^d \mapsto \mathbb{R}$ *in parameter space* $\mathcal{P}$*.*

By setting a fixed object p from the dataset into the parametrization function of Definition 2, we obtain a function that represents all possible hyperplanes, characterized through their different angles α_i and the distance δ, that contain that particular object p, which is reflected by Property 1.

Property 2. *An object* $(\alpha_1, ..., \alpha_{d-1}, \delta) \in \mathcal{P}$ *in parameter space corresponds to a* $(d-1)$*-dimensional hyperplane in data space.*

If we consider now to pick an arbitrary point in Hough space, that would be represented by a tuple of the form $(\alpha_1, ..., \alpha_{d-1}, \delta)$, we have in fact chosen a normal vector that represents a hyperplane in data space, as resembled by Property 2.

In summary Property 1 and 2 state the duality regarding single objects, namely that a single object in data space corresponds to a parametrization function in Hough space and that a single object in Hough space corresponds to a hyperplane in data space. The properties for multiple objects in either data space or Hough space are expressed by the following two properties taken from [1]:

Property 3. *Objects that are located on a common* $(d-1)$*-dimensional hyperplane in data space correspond to sinusoidal curves through a common point in parameter space.*

From Definition 1 we know that objects are mapped to parametrization functions in Hough space. Now suppose that a subset of parametrization functions intersect in Hough space at a specific point $(\alpha_1, ..., \alpha_{d-1}, \delta)$. From Property 2 we know that a specific point in Hough space corresponds to a specific hyperplane in data space. For the aforementioned subset of parametrization functions, it means that if they intersect at a specific point in Hough space, that their corresponding objects in data space are located on a *common* hyperplane. In the opposite direction, it means that if a subset of objects in data space is located on a *common* $(d-1)$-dimensional hyperplane, then their corresponding parametrization functions intersect at a specific point in Hough space, namely that one representing the common hyperplane in data space they are located on, which is the message of Property 3.

Property 4. *Objects located on the same sinusoidal curve in parameter space represent* $(d-1)$*-dimensional hyperplanes through the same point in data space.*

Imagine now that we map an object p to Hough space obtaining its parametrization function. If we pick any point on this parametrization function in Hough space, it means that we obtain a tuple of the form $(\alpha_1, ..., \alpha_{d-1}, \delta)$ in Hough space that corresponds to a hyperplane in data space. Since the selected point was located among the parametrization function of p it means that the hyperplane in data space contains p which is the statement of Property 4.

In the introductory lines of this paragraph, it was mentioned that this Hough transform-based approach is used with the purpose to locate dense regions within Hough space. The

understanding of *dense* is here that a subset of parametrization functions intersect in Hough space at (a) exactly one common point or (b) within a boundary of angles α_i and distance δ. The former means that the corresponding objects of the parametrization functions are located perfectly on a hyperplane in data space, while the latter means that the objects are located within a certain vicinity around the hyperplane. In the following of this dissertation, such dense regions are coined with the term *Region of Interest* (ROI). This concept of duality between data space and Hough space is the foundation on which also the non-linear Hough transform-based variants (Chapter 2 and 3) rely.

Detecting Non-linear Correlation Clusters Datasets can be generated by multiple underlying processes. The unknown generative functions from which the data is sampled can be of non-linear nature. One explicit use-case in which the detection of non-linear correlations within individual clusters is of interest can be seen in [30] where correlations between isotopes are investigated in archaeological context. In order to detect non-linear subspaces, and therefore clusters that exhibit a shape hinting to non-linear relations among their attributes, in Part I, two Hough transform-based methods have been proposed. One of them is capable of detecting clusters residing on low-dimensional *paraboloids* of revolution (axis-parallel) subspaces (Chapter 2) and the other method is capable of detecting clusters that reside on two-dimensional *periodic* subspaces within the given dataset (Chapter 3). Contrary to the semantic of correlation clustering in the linear case, namely to detect clusters that reside in arbitrarily oriented subspaces, we use in Part I the term correlation clustering synonymous for clusters where their objects reside on or around non-linear subspaces. The two proposed methods detect clusters on non-linear subspaces that are not arbitrarily-oriented, but axis-parallel, i.e. a parabola is not arbitrarily rotated but aligned towards one of the axes. One question in both works is how to parametrize the non-linear functions, by which the corresponding Hough space is defined. For the quadratic Hough space we introduce here the following formalization exemplary for two-dimensional data:

Definition 3. *Quadratic Parametrization Function. Let $x = (x_1, x_2)^T \in \mathcal{D} \subseteq \mathbb{R}^2$ be a 2-dimensional vector of a dataset $\mathcal{D}$. Further let $v = (v_1, v_2)^T$ be a 2-dimensional vector representing the vertex of a parabola. The parametrization function $f_x : \mathbb{R}^2 \mapsto \mathbb{R}$ of a vector x denotes the opening of a parabola defined by an object x and its vertex v:*

$$f_x(v_1, v_2) = \frac{(x_2 - v_2)^2}{4(x_1 - v_1)}$$

With the given definition we can map any two-dimensional object $x \in \mathbb{R}^2$ to its corresponding parameter function $f_x(v_1, v_2)$ that resides in a 3-dimensional parameter space $\mathcal{P}$, where two dimensions are spanned by the vertex v and one dimension by the opening p of the parabola. The semantic of the parameter function is that it represents all possible parabolas (of different vertices (v_1, v_2) and openings p) that contain an object x.

Having elaborated on an object x and its mapping through a parametrization function f_x we continue with the *Embedding-space - Hough-space Duality*, by introducing, like for the linear Hough space, four properties:

Property 5. *An object $x = (x_1, x_2)^T \in \mathcal{D} \subseteq \mathbb{R}^2$ in data space is represented by a quadratic curve $f_x : \mathbb{R}^2 \mapsto \mathbb{R}$ in parameter space $\mathcal{P}$.*

Similar to the previously mentioned linear case for a specific object x its corresponding parametrization function represents the infinite number of quadratic curves with different vertices and different openings of the form described in Definition 3 that would include the object x.

Property 6. *An object $(v_1, v_2, p) \in \mathcal{P}$ in parameter space corresponds to a 2-dimensional parabola in data space.*

In the opposite direction Property 6 means that any picked point in the quadratic Hough space corresponds to a parabola in data space characterized by its vertex through v_1 and v_2 and its opening p.

For the case of considering multiple objects in either spaces the following properties are provided that are adopted from[1]:

Property 7. *Objects that are located on a common 2-dimensional parabola in data space correspond to quadratic curves through a common point in parameter space.*

Again, suppose we are given a subset of objects in data space and their corresponding parametrization functions. Further we observe that the parametrization functions intersect at a specific point (v_1, v_2, p) in Hough space. This means that there exists a common parabola with the vertex v_1, v_2 and an opening p in data space on which the subset of objects are located.

Property 8. *Objects located on the same quadratic curve in parameter space represent 2-dimensional parabolas through the same point in data space.*

From the parametrization function of an object x (Definition 3) we obtain a function that represents all possible parabolas with all possible vertices v_1, v_2 and openings p that contain x. If we select in Hough space arbitrarily any points that are located on the parametrization function of x, these points represent parabolas in data space that contain the object x.

Moving from the quadratic Hough transform to periodic correlation clustering we provide here a formalization for a two-dimensional case in data space:

Definition 4. *Periodic Parametrization Function. Let $x = (x_1, x_2)^T \in \mathcal{D} \subseteq \mathbb{R}^2$ be a 2-dimensional vector of a dataset $\mathcal{D}$. Further let $p = (p_1, p_2, p_3)^T$ be a 3-dimensional vector representing with p_1 the amplitude, with p_2 the frequency and with p_3 the phase shift of a periodic (sinusoidal) function. The parametrization function $f_x : \mathbb{R}^3 \mapsto \mathbb{R}$ of a vector x denotes the vertical shift of a periodic function defined by an object x and its parameters p:*

$$f_x(p_1, p_2, p_3) = x_2 - p_1 \cdot \sin(p_2 \cdot x_1 - p_3)$$

With the given definition we can map any two-dimensional object $x \in \mathbb{R}^2$ to its corresponding parameter function $f_x(p_1, p_2, p_3)$ that resides in a 4-dimensional parameter space $\mathcal{P}$, where three dimensions are spanned by the parameters in p (spanned by amplitude, frequency and phase shift) and one dimension by the vertical shift d of the periodic function. The semantic of the parameter function is that it represents all possible periodic functions (of different amplitude, frequency, phase shift, and vertical shift) that contain an object x.

Having elaborated on an object x and its mapping through a parametrization function f_x we continue with the *Embedding-space - Hough-space Duality*, by introducing, like for the linear and quadratic Hough space, four properties as stated in [1]:

Property 9. *An object $x = (x_1, x_2)^T \in \mathcal{D} \subseteq \mathbb{R}^2$ in data space is represented by a sinusoidal curve $f_x : \mathbb{R}^3 \mapsto \mathbb{R}$ in parameter space $\mathcal{P}$.*

For the case of Property 9 imagine that we are given an object x. This object is mapped to a parametrization function in Hough space as provided in Definition 4. Since the arguments of the function are the amplitude p_1, frequency p_2 and phase shift p_3 the parametrization function represents all possible periodic functions containing x.

Property 10. *An object $(p_1, p_2, p_3, d) \in \mathcal{P}$ in parameter space corresponds to a 2-dimensional periodic function in data space.*

If we choose any point in the periodic Hough space we obtain a tuple (p_1, p_2, p_3, d) since that space is spanned by these four scalars representing the parameters of a periodic function as provided in Definition 4. The scalars in this tuple represent a specific periodic function in data space of amplitude p_1, frequency p_2, phase shift p_3 and vertical shift d.

Next, we dedicate to the properties for multiple objects in either spaces being expressed by the following two properties adopted from [1]:

Property 11. *Objects that are located on a 2-dimensional periodic function in data space correspond to sinusoidal curves through a common point in parameter space.*

Given the case that for a subset of objects their corresponding parametrization functions intersect in Hough space at a common point, this point is represented by a tuple (p_1, p_2, p_3, d). This means that the subset of parametrization functions share a common amplitude p_1, frequency p_2, phase shift p_3 and vertical shift d. Therefore the subset of objects are located on a *common* 2-dimensional periodic function with a common amplitude, frequency, phase shift and vertical shift in data space.

Property 12. *Objects located on the same sinusoidal curve in parameter space represent 2-dimensional periodic functions through the same point in data space.*

Again we map an object x to Hough space obtaining its parametrization function. Any sampled points in Hough space that are located on that parametrization function represent in data space 2-dimensional periodic functions that contain the object x.

The goal of the work in the Chapters 2 and 3 was to investigate if and in how far the Hough transform-based concept from CASH [1] can be transferred to the detection of clusters with their objects residing on non-linear subspaces. As a first step, it was necessary to identify the parameters and the parametrization function that provide the means to characterize the quadratic and periodic Hough space. Building on the previous step, it turned out to be challenging to identify ROIs in Hough space through the adaptive grid technique used in CASH[1]. In particular, formulating queries for checking if $minPts$ number of parameter functions intersect within a region in Hough space revealed that it is necessary to cover different cases. More specifically, as one case the parametrization function in Definition 3 exhibits a point of singularity if $x_1 = v_1$ and as such requires attention when formulating ROI queries. For the periodic case, we observed that data of a cluster within a dataset may be covered by not only one, but multiple periodic models equally well. In this case, we had to introduce heuristics to prefer those periodic models that provide a lower frequency, since an arbitrarily high frequency turned out to lead to an overfitting on the data. The conducted experiments in both cases (Chapter 2 and 3) showed that the developed methods can detect clusters residing on quadratic or periodic subspaces whereas other existing methods like CASH[1] or DBSCAN[10] are, as expected, incapable of detecting those clusters.

Another observation that we made was that the detection of clusters in which the object resides on quadratic subspaces is computationally expensive. To mitigate this, in Chapter 2 a first simple parallelization approach is proposed. One challenge was to develop a concept where the computations in Hough space are distributed on multiple CPU cores while at the same time being mostly mutual independent of other computations. The experiments illustrated that this simple approach scales with an increasing number of CPU cores. However, further work is required to develop more sophisticated approaches, relying on different frameworks such as Map-Reduce[8] for example. Another aspect that addresses all three presented correlation clustering techniques (linear, quadratic, periodic) is the question of how the methods deal with cases where actually no linear, quadratic, or periodic patterns are present within the data. This depends on the decisions made by the domain experts using the methods. To be more specific, by setting the minimum support ($minPts$) of objects that need to be on or around a (linear, quadratic, or periodic) subspace the users determine when actually a pattern is present in a given dataset.

Lastly, one feature that comes with CASH[1], and also with the other Hough transform-based methods in Chapters 2 and 3 for "free" is that for each cluster one obtains a quantitative model of the underlying subspace. As one use case, the obtained models for each cluster can be utilized to generate their own artificial datasets for training machine learning systems, which is of special interest in cases where the amount of collected data is sparse. Endeavors to derive quantitative models from PCA-based correlation clustering methods have been made in [2]. In conclusion, we provide the following two hypotheses for Chapters 2 and 3:

Hypothesis 1: The concepts of Hough transform-based correlation clustering including the adaptive grid-based search strategy as proposed in CASH[1] can be extended to detect clusters with their objects residing on quadratic axis-parallel subspaces in lower-dimensional settings.

Hypothesis 2: The concepts of Hough transform-based correlation clustering including the adaptive grid-based search strategy as proposed in CASH[1] can be extended to detect clusters with their objects residing on periodic subspaces in two-dimensional settings.

1.2 Strategies for Region of Interest Detection in Hough Space

While the Hough transform-based approaches enable the detection of global correlation clusters and are unbound from any locality assumption, they suffer from three factors when it comes to the detection of regions of interest (ROI) in Hough space: (a) the runtime deteriorates to a full enumeration of all paths of a binary tree and therefore to an exponential runtime, where a tendency to such behavior is notable if the data contains large amounts of noise (b) the computation of regions of interest is expensive, or more precisely: computing if a multivariate parametrization function intersects with a hypercuboid in Hough space is costly and (c) it is not possible to apply 'classical' well-known clustering algorithms such as DBSCAN [10] or k-means [29] in Hough space.

Historically, the first approach to determine ROIs in Hough space was to deploy a static grid and to count the number of intersecting parametrization functions per grid cell. This approach is coined in the literature [9] with the term "accumulator". The handling with accumulators leads to two observations: (1) it is non-trivial to identify an 'adequate' grid resolution and (2) with increasing dimensionality the runtime increases exponentially which is also additionally affected by increasing grid resolution. The correlation clustering algorithm CASH[1] deploys a depth-first search (DFS) strategy in which all axes in Hough space are split by a pre-defined order. If a minimum number of parametrization functions does not intersect with a partition (hypercuboid) in Hough space, the entire partition is pruned. This approach, however, can deteriorate to an exponential runtime as mentioned before.

To mitigate this, in Part I, the algorithm D-MASC is proposed (Chapter 4) which addresses the problem by first applying a MeanShift algorithm [12] on the dataset to reduce the number of parametrization functions down to the number of modes from the MeanShift approach. At this point, we provide a formal introduction to the MeanShift algorithm. The following description of the MeanShift method is based on [7] and briefly summarized as follows:

MeanShift is an algorithm by which the density within a given d-dimensional feature space can be estimated. This is achieved by using a kernel $K : \mathbb{R}^d \mapsto \mathbb{R}$, e.g. a radially symmetric kernel and a shifting of cluster centers through a gradient ascent approach. By computing the gradient ascent of the kernel density estimate (KDE) $\hat{f}_{h,K}$ we obtain the following expression based on [7]:

$$\nabla \hat{f}_{h,K}(x) = \frac{2c_{k,d}}{nh^{d+2}} \underbrace{\left[\sum_{i=1}^{n} g\left(\left\|\frac{x - x_i}{h}\right\|^2\right)\right]}_{\text{term 1}} \underbrace{\left[\frac{\sum_{i=1}^{n} x_i g\left(\left\|\frac{x-x_i}{h}\right\|^2\right)}{\sum_{i=1}^{n} g\left(\left\|\frac{x-x_i}{h}\right\|^2\right)} - x\right]}_{\text{term 2}}$$

Term 1 corresponds to the density estimation at x. The second term is known as the *MeanShift* vector m which shows in the direction of the maximum increase in density and corresponds to the density gradient estimate at a given object x using a kernel K. Hence in the literature MeanShift is also known as a *mode seeking algorithm* since the objects shift towards their center of mass (mode). The MeanShift algorithm for a given object x_i performs the three following steps: (1) Compute the MeanShift vector $m(x_i^t)$ at an iteration t (2) Translate the density estimation window (bandwidth): $x_i^{t+1} = x_i^t + m(x_i^t)$ and (3) Iterate step (1) and (2) until convergence is reached: $\nabla \hat{f}_{h,K}(x_i) = 0$, or in more simple terms: until the objects barely *shift*.

In D-MASC, the step of applying MeanShift on the dataset is followed by rasterization of single parametrization functions of the modes into small cells in Hough space. Those regions where increasing number of intersecting cells can be observed are considered as ROIs. The conducted experiments indicate that D-MASC is in high-noise scenarios capable to detect ROIs in lower runtime compared to the adaptive grid-based approach in CASH. This benefit in runtime can be traced back to the involved MeanShift step that acts as an aggregation of objects, leading in consequence to a reduction of the number of parametrization functions from the size of the dataset down to the number of detected modes.

The aforementioned accumulator strategy for ROI detection, as illustrated in Figure 1.4 (a) has been replaced in CASH by an adaptive grid-based approach where only regions that satisfy the criterion of $minPts$ parametrization functions intersecting are further scanned at higher grid resolutions. An illustration for the adaptive grid strategy can be seen in 1.4 (b). What can be observed here is that regions that do not meet the $minPts$ threshold are not further refined, reducing the number of queries. For scenarios where the data contains large amounts of noise, the strategy in D-MASC first detects the modes within a given dataset. The corresponding parametrization functions of those modes are then rasterized into small cells. Those regions with a high count of overlapping cells are considered as ROIs. The rasterization (green cells) of the parametrization functions of the modes (red) can be seen in 1.4 (c). In all the so far mentioned strategies (Accumulator, adaptive grid, and D-MASC) the queries if a parametrization function intersects with a region turn out to be computationally expensive. This raised the questions (a) if it is possible to detect ROIs in Hough space without handling with parametrization functions and (b) if then well-

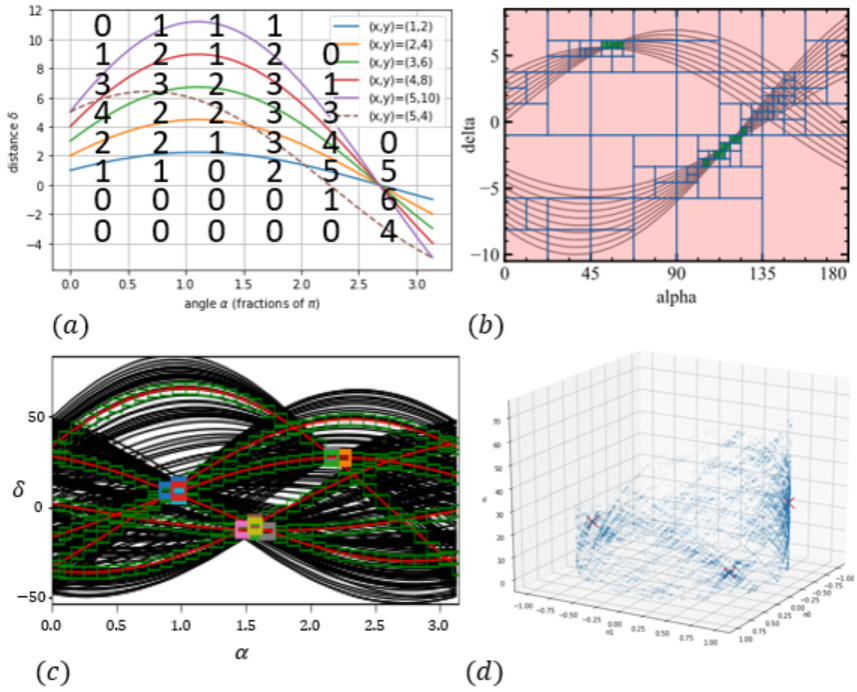

Figure 1.4: Different ROI detection strategies: (a) Accumulator-based, (b) adaptive grid-based (c) D-MASC, (d) d-tuples sampling method (Part I, Chapter5)

known clustering algorithms can be applied to discover the ROIs. In Chapter 5 it turned out that handling with parametrization functions can be circumvented by taking a different approach: Instead of mapping each object in data space to a parametrization function in Hough space, we take in a d-dimensional datasets in data space multiple samples of size d and compute the parameters of the hyperplane they would be located on. The computed hyperplane can be represented as a single point in Hough space. As a consequence, all computations are performed on a set of normal vectors in Hough space. This enables the use of clustering algorithms like DBSCAN[10] or k-means[29] to detect ROIs in Hough space. In the conducted experiments of Chapter 5, we observed that one can find the ROIs with this strategy eliminating the need to handle with parametrization functions. As one illustrative example, the detected ROIs with the strategy of Chapter 5 can be seen in Figure 1.4 (d). In summary, we come with the following hypothesis regarding Chapters 4 and 5:

Hypothesis 3: The runtime of Hough transform-based methods on datasets with large amounts of noise can be reduced through a pre-aggregation in data space and a rasterization approach in Hough space.

Hypothesis 4: Handling with parametrization functions in Hough space for ROI detection can be circumvented by a different model where instead it is operated with objects in Hough space permitting the ROI detecting with well-known clustering algorithms.

1.3 Evaluation of Correlation Clusterings

We leave now the field of Hough transform-related algorithms and dedicate to a topic that concerns correlation clustering algorithms in general. While there exist several correlation clustering algorithms, one question that has not been asked so far with this observation, yet being important, is how to *evaluate* the clusterings that are yielded by the respective algorithms. This is of special interest in cases where we do not have any "ground truth" labels, rendering the application of external measures impossible. In this context, we elaborate on our understanding of the terms *internal* and *external* since they will accompany the reader throughout Chapters 6 and 7.

We understand under the term *external* validation a measure which compares a given (observed) clustering result against a reference (expected) result which is sometimes referred to among the literature as a *"ground truth"* as it is stated in [14]. This *a priori* knowledge (to which class an object belongs) is not part of the actual dataset on which the clustering has been performed, but has been provided *after* the data was obtained by e.g. domain experts. Accordingly, in the work of [32] external validation is meant to be based on previous knowledge about the provided data, which corresponds to the idea of a *"ground truth"* (in form of class labels). Supporting [14] and [32], according to [33] so-called *external criteria* are those which evaluate a cluster regarding of an independently drawn structure imposed on the dataset, where such a structure would be class labels for example.

With the term *internal* validation we understand that a clustering result is evaluated based on intrinsic information of the data itself [14],[32],[33] (Ch. 16.1, P.864) (e.g. the structure of detected clusters). Further, since no *external* information (in form of labels) are provided, internal evaluation measures require assumptions about what can be considered as a favorable property of a detected cluster [14]. Being aware that the underlying *model assumption* can also be regarded as some sort of *a priori* and therefore *external* knowledge that is imposed a given clustering, in this dissertation it is adhered to the previously mentioned and common understanding of the term *internal* among the literature.

Lastly, we elaborate on the term *model agnostic* in the context of Chapter 7. Among clustering algorithms, there exist different *archetypes*. As an example, we may have a density-based archetype encompassing clustering algorithms such as DBSCAN [10] or OPTICS [4]. In our case we have the correlation clustering archetype of clustering algorithms. *Within* the archetype of correlation clustering algorithms, the proposed evaluation measure in Chapter 7 is *model agnostic* in a sense that it does not favor any particular correlation clustering algorithm based on an algorithm-specific property (like density in correlation subspace), but evaluates correlation clusterings based on three *properties* which are, to the best of our knowledge, inherent to all correlation clustering algorithms, namely: all correlation clustering algorithms can be evaluated based on (1) their reconstruction error which is obtained when e.g. projecting objects of a cluster to their respective subspaces

and back (2) the number of clusters and thus subspaces (3) the number of dimensions their subspaces have. The evaluation can be made based on these three criteria since all correlation clustering algorithms yield a clustering which is characterized through the number of clusters (and therefore also subspaces), number of dimensions of their subspaces, and of how "close" objects of a cluster are located on or around their respective subspace. Transferring what has been stated before in the paragraph on internal validation, in Chapter 7 for correlation clustering it relies on the *assumption* that it is a favorable property if a given dataset can be represented with a low reconstruction error (Property 1) while at the same time exhibiting a low model complexity (Properties 2 and 3). Conclusively, the term *model agnostic* refers in this work to the aspect that the internal evaluation measure is agnostic to the algorithms *within* an archetype of clustering algorithms (here: correlation clustering) but not necessarily to other archetypes.

While there exists a variety of internal evaluation measures which are tailored at different underlying clustering models and assumptions, none of them, to the best of our knowledge are specialized for arbitrarily oriented subspace clustering. For that reason we have proposed in Part II, Chapter 6 [23] the first two internal evaluation measures, namely the Normalized Projected Energy (NPE) and the Subspace Compactness Score (SCS). NPE is derived from the objective function of the ORCLUS[3] clustering algorithm while SCS is derived from CASH[1]. During the investigations, the expected insight confirmed the assumption that ORCLUS and CASH optimize for the same objective. However, both of them have algorithm-specific biases and come with certain shortcomings. To be more specific, NPE does not account explicitly for the dimensionality of a subspace, meaning that a better NPE score is achieved if the subspaces have a higher dimensionality since the deviations of objects to a subspace are reduced with increasing dimensionality. To remedy such shortcomings, we have derived an evaluation measure that accounts for the reconstruction error on the one hand, and the model complexity, on the other hand, coined with the term *S*um of subspace *R*econstruction *E*rrors (SRE)). SRE is, in contrast to NPE and SCS, model agnostic in a sense as it has been stated in the previous paragraph. In summary, we have come up for Chapters 6 and 7 with the following hypothesis:

Hypothesis 5: Among some correlation clustering algorithms exist common properties, such as common objectives by which the algorithms compute their clusterings, or commonalities that can be observed with the clustering results, that can be used as a basis for a common internal evaluation measure.

1.4 Different Views of Correlation Clusterings

Among the plethora of correlation clustering algorithms, different methods support different views. To commit an intuition for the meaning of the term "view" we refer to Figure 1.5. Here we can see a toy example where we observe a set of objects in which two attributes per object are captured, namely (1) age in years and (2) chocolate consumption in gram

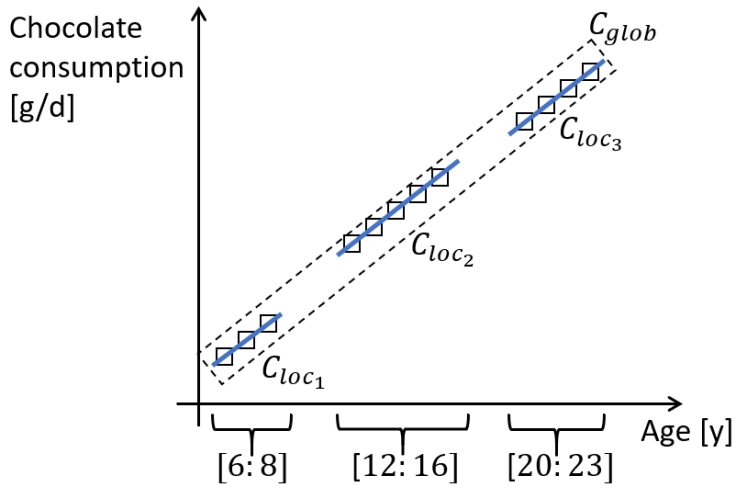

Figure 1.5: Toy example with local and global correlation clusters

per day, which denote the two axes of our coordinate system. At a first sight we may see three *locally* dense correlation clusters, denoted with C_{loc_1}, C_{loc_2} and C_{loc_3}. Each of these clusters encloses an age interval.

Having only the *local* view communicates to the scientists that there exist three age groups. Additionally, they observe that there exist "age gaps" i.e., between 9 and 11 years old people. This information can be valuable for example to initiate the development of marketing programs that specifically target 9 to 11-year-old consumers. On the contrary, in the *global* view, data scientists can observe that all three clusters, follow a *global* correlation or trend. This information reveals that among all three age groups a common trend is visible, from which potentially the trends for higher age groups can be extrapolated, given the case that no market analysis data is available. By applying algorithms that support solely either a local or a global view, data scientists see only "half of the picture".

One challenge in pursuit of obtaining both views is to identify local and global correlation clusters that belong to the same common subspace. A simple application of a local and a global correlation clustering algorithm on a given dataset would not necessarily lead to the expected result of finding local and global correlation clusters of the *same* common subspace for the following reasons: First, it would require to probe ranges of parameter settings in both (local and global) algorithms in order to find local and global clusters belonging to common subspaces. Depending on how thorough the parameter ranges have been probed, the detection of local and global correlation clusters with a common subspace may be missed. The second reason is founded on a statement in [34] which addresses an aspect related to the locality assumption, namely: "Outliers in the neighborhoods, that do not belong to the corresponding cluster prevent the algorithms from finding suitable subspaces, ..." [34]. Indeed we could independently observe that outliers do have an impact on the detection of subspaces with respect to their orientation and translation as we

have investigated on the example of PCA, a method that is predominantly used in local correlation clustering algorithms (s. Subsection 1.5 and Chapter 9). Outliers may then render it impossible for local correlation clustering algorithms to find local clusters with subspaces being similar to those yielded by global correlation clustering algorithms. These reasons motivated the development of the following method:

In this proposed framework first with a density-based clustering algorithm like DBSCAN[10] dense regions in the full-dimensional data space are detected. On the detected local regions a global correlation clustering algorithm (here CASH) is applied to detect within the full-dimensional dense, and therefore local regions, correlation clusters. Since CASH is a global correlation clustering method, the detected clusters and their *subspaces* should not be affected by any outliers within the locally dense cluster. The obtained local correlation clusters are in a bottom-up fashion subsumed to global clusters if they do not exceed mutually the user-defined thresholds for orientation and translation dissimilarity. Further in Part III, Chapter 8 we elaborate on the impacts of different hyperparameter settings with regards to the properties of the obtained local and global correlation clusters. In summary, we come up with the following hypothesis for Chapter 8:

Hypothesis 6: It is possible to obtain local and global correlation clusters that belong to the same common subspace by using a full-dimensional density-based clustering algorithm and a global correlation clustering method.

1.5 Resilience against Outliers

So far we have elaborated in this introductory part on different aspects of correlation clustering like views, non-linearity, internal evaluation measures, and ROI detection in Hough space. There exists one aspect that affects correlation clusterings when it comes to the computation of their respective subspaces: outliers. Depending on the chosen hyperparameters and on the underlying model, outliers can heavily skew the orientation and translation of arbitrarily oriented subspaces of each cluster and as such impact view, the reconstruction error, and therefore also the resulting clustering quality.

At this point we have to address the following questions first, namely (a) what an outlier is and (b) how it can be identified. If we think of the subspaces, we compute the distances of objects orthogonal to a subspace. However, it remains so far open which distance is an indicator for an outlier. There exist many outlier models as it has been elaborated on in the survey of [35]. We refer here to the so-called *deviation-based* outlier model which states that: "Deviation-based outlier detection groups objects, captures some characteristics of the group, and considers those objects outliers that deviate considerably from the general characteristics of the groups." [35]. The *general characteristics* in our case is that objects belonging to the group should exhibit a low distance to the subspaces, or in other terms,

outlier show a significantly higher distance to the subspace. In this context we cite the definition of an outlier by Hawkins [15]: "An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism.". Referring to Hawkins outlier definition an open question in this context is: what is understood by "so much" deviating from other observations? One simple approach would be to observe the distribution of distances of objects to a subspace and determine a cut-off threshold, by which all distances that are larger than the specified threshold are considered as outliers.

While in Hough transform-based approaches outliers do not influence the arbitrarily oriented subspaces, the PCA-based methods are potentially susceptible to outliers, especially if no outlier detection routines are incorporated. At this point we provide a more formal introduction of PCA based on [2]:

Given a d-dimensional dataset $\mathcal{D} \subseteq \mathbb{R}^d$ consisting of n vectors. Further let $\mu_\mathcal{D}$ denote the mean of all objects $x \in \mathcal{D}$. The *covariance matrix* $\Sigma_\mathcal{D}$ of $\mathcal{D}$ is defined as:

$$\Sigma_\mathcal{D} = \frac{1}{|\mathcal{D}|} \cdot \sum_{x \in \mathcal{D}} (x - \mu_\mathcal{D}) \cdot (x - \mu_\mathcal{D})^T$$

The decomposition of the covariance matrix $\Sigma_\mathcal{D}$ yields:

$$\Sigma_\mathcal{D} = V_\mathcal{D} \cdot E_\mathcal{D} \cdot V_\mathcal{D}^T$$

where $V_\mathcal{D}$ denotes the Eigenvector matrix which is an orthonormal matrix containing the Eigenvectors of $\Sigma_\mathcal{D}$ and $E_\mathcal{D}$ denotes the Eigenvalue matrix that is a diagonal matrix with the corresponding Eigenvalues being located in decreasing order in its diagonal. In fact, every Eigenvector $\lambda_\mathcal{D}$ is accompanied by a corresponding Eigenvalue that expresses *how strong* the variance is in that particular direction. Typically the *strong* Eigenvectors, namely those capturing the direction of highest variance are identified by sorting the Eigenvectors in descending order by their corresponding Eigenvalues. The number of Eigenvectors required to meet an explained variance of at least α, where (according to [2]) α is typically chosen between 0.8 and 0.9, serves as a threshold for determining the dimensionality of the subspace spanned by the strong Eigenvectors.

Single or multiple outlier objects primarily influence the covariance matrix, or to be more precise: affect the variance and covariance entries in the covariance matrix. This causes the Eigenvectors which are obtained from solving the Eigenproblem to be skewed with respect to orientation or translation. In Part III, Chapter 9 [21], the conducted experiments illustrate the effects of outliers on the principal components. There exists a rich body of literature dealing with robustness for PCA, which approaches this issue from different perspectives, using different techniques as stated in [21]. However, the impressions from observing the different techniques are, that they also increase in complexity, with respect to the complexity of the underlying mathematical framework. This circumstance challenged us to approach the following question: Is it possible to make PCA more robust

by introducing a minimum of modifications to the current method? To approach this endeavor, we remembered at which point outliers have an impact on PCA. In the covariance matrix, the variance with respect to the mean is computed. From a statistical point of view, the mean itself is prone to outliers as the following toy example shall illustrate: given a sequence of numbers: $S = [1, 4, 4, 4, 40]$, the mean of S, is $\mu(S) = 10.6$, however, the median is $med(S) = 4$. As it can be observed on this tiny example, the median is, as expected, more robust compared to the mean. In the work of [11], the comedian (or co-MAD, for co-Median Absolute Deviation) matrix is introduced which instead of relying on the mean and the variance, relies on the median and the absolute deviation to the median. In fact, in Part III, Chapter 9, the sole change that has been introduced to the PCA is to replace the covariance matrix with a co-MAD matrix. The first conducted experiments support our expectations that the resulting principal components are only marginally affected by outliers, in comparison to a classical covariance-based PCA. In summary, the hypothesis in Chapter 9 is as follows:

Hypothesis 7: The susceptibility of PCA towards outliers can be mitigated by introducing a minimum of modification to the PCA method itself.

In the upcoming pages, three parts of this dissertation are presented, namely Part I, dealing with Hough transform-based approaches, Part II that is dedicated to the evaluation of correlation clusterings, and Part III that consists of chapters addressing different views and resilience against outliers. This dissertation is concluded with chapter 10 that provides a wrap-up and proposes potential targets for future work.

Part I

Hough Transform Based Approaches

"We can only see a short distance ahead,
but we can see plenty there that needs to be done."

Alan Turing

Chapter 2

Detecting global hyperparaboloid correlated clusters: a Hough-transform based multicore algorithm

Published in:
Kazempour, Daniyal; Mauder, Markus; Kröger, Peer and Seidl, Thomas. "Detecting global hyperparaboloid correlated clusters: a Hough-transform based multicore algorithm." Distributed and Parallel Databases (DAPD) 37.1 2019: 39-72. https://doi.org/10.1007/s10619-018-7246-0

In the first chapter of the dissertation the linear Hough transform [16] and its generalization to an arbitrary number of dimensions [1] has been introduced which is utilized for detecting clusters that reside in arbitrarily oriented subspaces. In the following chapter, an endeavor is made towards detecting clusters that reside on or around subspaces of parabolic shape hinting at features that may exhibit quadratic relations.

The detected clusters of parabolic shape are axis-parallel, where the one rotation axis is provided by the user. The decision to restrict the user to focus to one particular rotation axis was made since even in low-dimensional settings capturing clusters that reside on paraboloid subspaces with different rotational axes would significantly increase the complexity of the computations.

Besides the focus on the axis-parallel setting, the proposed method is, contrary to [1] which is generalized to arbitrary dimensionality, constructed and investigated towards low-dimensional settings. While in the non-linear cases a risk of overfitting may occur with increasing degree of polynomials, the proposed method in this chapter is not affected by this issue since it is limited to polynomials of degree two. In addition to overfitting over the polynomial degree also overfitting by increasing the dimensionality may be possible which is not further discussed in this chapter, since this work does not discuss paraboloids in high-dimensional settings.

Chapter 3

Detecting Global Periodic Correlated Clusters in Event Series based on Parameter Space Transform

Published in:
Kazempour, Daniyal; Emmerig, Kilian; Kröger, Peer and Seidl, Thomas. "Detecting Global Periodic Correlated Clusters in Event Series based on Parameter Space Transform." Proceedings of the 31st International Conference on Scientific and Statistical Database Management (SSDBM). 2019. https://doi.org/10.1145/3335783.3335803

In the previous chapter, we proposed a first endeavor for detecting clusters that reside in subspaces that exhibit the shape of a paraboloid. In the context of time series, subsets of data may be generated from one (or multiple) underlying periodic process(es). To detect clusters that are located on or around subspaces in which their features hint towards a periodic relationship we propose a Hough transform-based approach which is focused on two-dimensional time series.

An observation that emerged during the development of this method were cases where single clusters could be characterized by multiple periodic models. A closer look revealed that with increasing frequency even objects that were supposed to be outliers (by the "ground truth") have been overfitted to the underlying model. Investigating the cause that lead to this observation directed to the frequency. More specifically, by just increasing the frequency of the underlying periodic model any object of the data may be fitted to the model. Due to the compact character of a short paper, the understanding of this type of overfitting may not be easy to follow, and as a consequence, we provide the reader a more concrete case. For this purpose consider the dataset in Figure 3.1 (left) where five objects can be observed that are located on a periodic function (with the parameter of amplitude $a = 4$, frequency $b = 5$, phase shift $c = \frac{\pi}{2}$ and vertical shift $d = 3$) and one object as an outlier. Through an increase of the frequency by a factor of 6 ($b = 30$) we obtain the function as seen in Figure 3.1 (right) where the outlier is now also located on the sinusoidal curve.

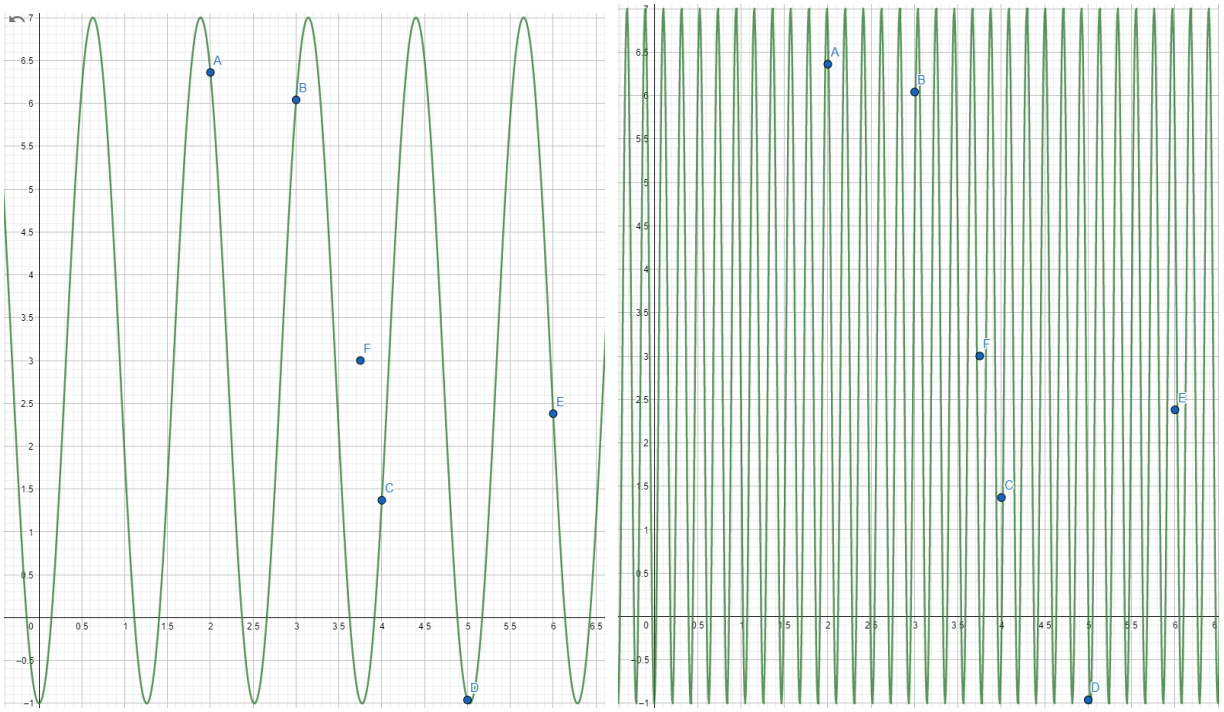

Figure 3.1: Left: Data fitted to a periodic model. Right: Data fitted to a periodic model with increased frequency

By increasing the frequency to arbitrarily high values one can potentially overfit on any dataset. The proposed method in this chapter provides a simple heuristic by choosing in cases of equally suitable models, those that exhibit the lowest frequency. What remains for future work are thorough studies of how overfitting can be detected by the method.

Chapter 4

D-MASC: A Novel Search Strategy for Detecting Regions of Interest in Linear Parameter Space

Published in:
Kazempour, Daniyal; Bein, Kevin; Kröger, Peer and Seidl, Thomas. "D-MASC: A Novel Search Strategy for Detecting Regions of Interest in Linear Parameter Space." 11th International Conference on Similarity Search and Applications (SISAP). Springer, Cham, 2018: 163-176. https://doi.org/10.1007/978-3-030-02224-2_13

While the previous two chapters dealt with the topic of non-linearity in the context of the Hough transform, this chapter is dedicated to the process of detecting regions of interest (ROI) in Hough space. Besides the original accumulator-based approach and the strategy introduced in [1], a different technique is proposed in this chapter where the conducted experiments indicate a certain benefit of D-MASC in high-noise scenarios. D-MASC itself relies in an initial step on a pre-aggregation of the data into modes using the MeanShift algorithm which has been introduced in chapter 1. To ensure self-containedness, further information regarding the dataset on which the experiments were conducted in Fig. 11 are provided: The dataset used there consists of 500 objects assigned to the observable clusters and 500 objects as noise.

Chapter 5
A Galaxy of Correlations

Published in:
Kazempour, Daniyal; Krombholz, Lisa; Kröger, Peer and Seidl, Thomas. "A Galaxy of Correlations - Detecting Linear Correlated Clusters through k-Tuples Sampling using Parameter Space Transform." 22nd International Conference on Extending Database Technology (EDBT). 2019: 702-705. https://doi.org/10.5441/002/edbt.2019.91

The D-MASC approach from the previous chapter relies on a region of interest (ROI) detection which is tailored at finding intersecting parametrization functions in Hough space. A question that motivated the method proposed in this chapter was: is it possible to detect clusters with their corresponding subspaces hinting towards potentially linear relations between their features without dealing with parametrization functions in Hough space? To recall the semantics behind the parametrization functions as introduced in chapter 1: An object in embedding space is mapped to a parametrization function in Hough space representing all possible lines in embedding space containing the particular object. To turn away from parametrization functions in Hough space, we pursued a different idea: By sampling d tuples (where d refers to the dimensionality of the dataset) and computing the linear model (line or plane on which the d objects are located) we obtain a normal vector and a distance from the origin orthogonal to the plane. Both information can be resembled as a point in Hough space. In fact, we no longer deal with parametrization functions but with points in Hough space, each of them representing planes in embedding space. This chosen approach is promising since we can now detect ROI without any queries if parametrization functions intersect within a certain area in Hough space and without grid refinements which can deteriorate in worst-case according to [1] to an exponential runtime. Instead, we can utilize clustering algorithm, such as OPTICS [4]. The conducted preliminary experiments show that with this approach we can detect in Hough space ROIs that represent linear subspace clusters in embedding space.

Part II

Evaluation of Correlation Clusterings

"The most important property of a program is whether it accomplishes the intention of its user."

C.A.R. Hoare

Chapter 6

Towards an Internal Evaluation Measure for Arbitrarily Oriented Subspace Clustering

Published in:
Kazempour, Daniyal; Kröger, Peer and Seidl, Thomas. "Towards an Internal Evaluation Measure for Arbitrarily Oriented Subspace Clustering". In 2020 International Conference on Data Mining Workshops (ICDMW) (pp. 300-307). IEEE. https://doi.org/10.1109/ICDMW51313.2020.00049

This second part of the dissertation deals with the aspect of evaluating correlation clusterings without any labels. In this context, we refer to the terminologies of internal and external evaluation measures as stated in chapter 1. One challenge was to identify a potentially common objective among some correlation clustering algorithms. As a first attempt, in this chapter evaluation criteria were derived from two existing correlation clustering algorithms namely CASH [1] and ORCLUS [3]. The two derived internal evaluation measures (Subspace Compactness Score [SCS] and Normalized Projected Energy [NPE]) are further investigated in this chapter regarding their objectives by which they evaluate a correlation clustering. The conducted experiments showed that while the proposed measures take into account the reconstruction error, more specifically, the deviation of objects to their respective subspace, they do not account for the model complexity. The latter is an aspect that is further addressed in the following chapter.

Further, there are certain aspects that need to be addressed for clarification purposes: Within this chapter, it is discussed in II.C how a quantitative model can be obtained from CASH clustering results by applying PCA on them. This idea has been originally mentioned in [1] (Sec. 3.7) and was not further pursued since, as the authors of [1] state, CASH provides already an implicit quantitative model capturing the orientation, translation, and through the perimeters of the hypercuboids in Hough space also the deviations to the detected hyperplane.

Besides the elaboration of how a quantitative model could be obtained for a CASH clustering result, it is also elaborated on how CASH can emulate properties of ORCLUS. The opposite direction has not been discussed and is advised as a target for future work. Further, in the context of mutual emulation capabilities, Definition 4 should be regarded as a proposition.

Apart from the differences on which it has been elaborated on in this chapter, one major difference remains to be investigated in more detail in future work, namely the impact of different shapes of noise distributions on the two clustering algorithms.

Chapter 7

I fold you so! An internal evaluation measure for arbitrary oriented subspace clustering

Published in:
Kazempour, Daniyal; Beer, Anna; Kröger, Peer and Seidl, Thomas. "I fold you so! An internal evaluation measure for arbitrary oriented subspace clustering". In 2020 International Conference on Data Mining Workshops (ICDMW) (pp. 316-323). IEEE. https://doi.org/10.1109/ICDMW51313.2020.00051

The evaluation measures from the previous chapter did not account for model complexity (number of clusters/subspaces and dimensionality). More specifically, one could observe that a clustering with more number of clusters is valued as equally good compared to a clustering with fewer clusters but the same reconstruction error. While the internal measures from the previous chapter behave as expected, namely to evaluate clusterings based on the reconstruction error, they do not account for the model complexity. This gets in particular problematic in terms of overfitting, since in a d-dimensional embedding space potentially any d objects can be assigned to a cluster spanning their own $d-1$-dimensional subspace with a reconstruction error of zero. In this sketched extreme case, the reconstruction error is minimized by increasing the model complexity through an increased number of clusters and therefore subspaces. In the pursuit of a better evaluation measure that also accounts for the model complexity a question that emerged was: what are properties that are common to correlation clustering results? It turns out that correlation clustering algorithms can on the one hand be characterized by the deviation of the cluster members to their respective subspaces (reconstruction error) and on the other hand by their model complexity in terms of number of clusters (therefore subspaces) and dimensionality of subspaces. From this observation emerged the evaluation measure named Sum of subspace Reconstruction Errors (SRE). The design decision was to construct SRE in such a way that (a) even with a low reconstruction error, the clustering is penalized if the model complexity is increased (i.e. through an increased number of clusters and therefore subspaces) and (b) that even with a low model complexity the clustering is penalized if the reconstruction error increases. The chosen approach is promising since it takes both aspects (reconstruction error and model complexity) into account yielding a bad score if the clustering result exhibits tendencies towards overfitting or an increasing reconstruction error is observed. The mentioned expected behaviors could be observed in the conducted experiments, making SRE a potential candidate for an internal evaluation measure.

The proposed internal evaluation measure of this chapter shares similarities to a quantitative model for correlation clustering introduced in [2] on a conceptual level. The similarities to [2] are the following: (1) Like in [2] first a correlation clustering algorithm is applied

on a given dataset followed by the computation of the Eigenvectors derived from the covariance matrix of each cluster (2) Those Eigenvectors with the λ-largest corresponding Eigenvalues (strong Eigenvectors) are considered as those directions spanning the correlation hyperplane. Further for each cluster the deviation of objects to their respective hyperplane is computed by projecting the objects down to the subspace spanned by the strong Eigenvectors and back to the full-dimensional embedding space.

Part III

On Different Views and Resilience against Outliers of Correlation Clusterings

"There is no unique picture of reality."

Stephen Hawking

Chapter 8

You see a set of wagons - I see one train: Towards an unified view of local and global arbitrarily oriented subspace clusters

Published in:
Kazempour, Daniyal; Kröger, Peer; Yan, Long Matthias and Seidl, Thomas. "You see a set of wagons-I see one train: Towards a unified view of local and global arbitrarily oriented subspace clusters". In 2020 International Conference on Data Mining Workshops (ICDMW) (pp. 308-315). IEEE. https://doi.org/10.1109/ICDMW51313.2020.00050

The so far presented chapters either dealt with Hough transform-based approaches or with the internal evaluation of correlation clusterings. This chapter deals with the aspect of views, more concrete with the endeavor to bring the so-called local and global views together. The design decision in this chapter was to first identify locally dense regions and to derive their respective subspaces followed by a merging of clusters based on the similarity of their subspaces. The conducted experiments showed that the proposed approach while providing local *and* global views, does not forfeit w.r.t. runtime and cluster quality compared to the global correlation clustering algorithm CASH.

While the chapter provides notions for local and global correlation clusters, we introduce at this point a notion for density-reachable (as stated in [10]) in the context of Definitions 3 [density-connected] and 5 [Local Arbitrarily Oriented Subspace Cluster] to ensure self-containedness in this chapter:

Definition. *Density-reachable A point p is density-reachable from a point q wrt. ϵ and $MinPts$ if there is a chain of points $p_1, ..., p_n, p_1 = q, p_n = p$ such that p_{i+1} is directly density-reachable from p_i*

Besides the definition of density-reachability, we address in the following an aspect that is not handled in the chapter. While the proposed method works on the experiments provided, one can think of cases that may not be captured by it. As stated at the beginning of this preface, the proposed approach first identifies locally dense regions within the dataset. If we now consider that a density-based cluster may not be located only in a single common subspace but reside in two or more different subspaces we are confronted with a situation as shown in Figure 8.1 (left). Here the *single* dense region contains subsets of objects that may reside in *two* different superimposed subspaces with different orientation as it can be observed in Figure 8.1 (right). Determining the subspace of the entire dense region would mean forcefully fitting all objects into one common subspace, neglecting local directions of variance and therefore the superimposed subspaces.

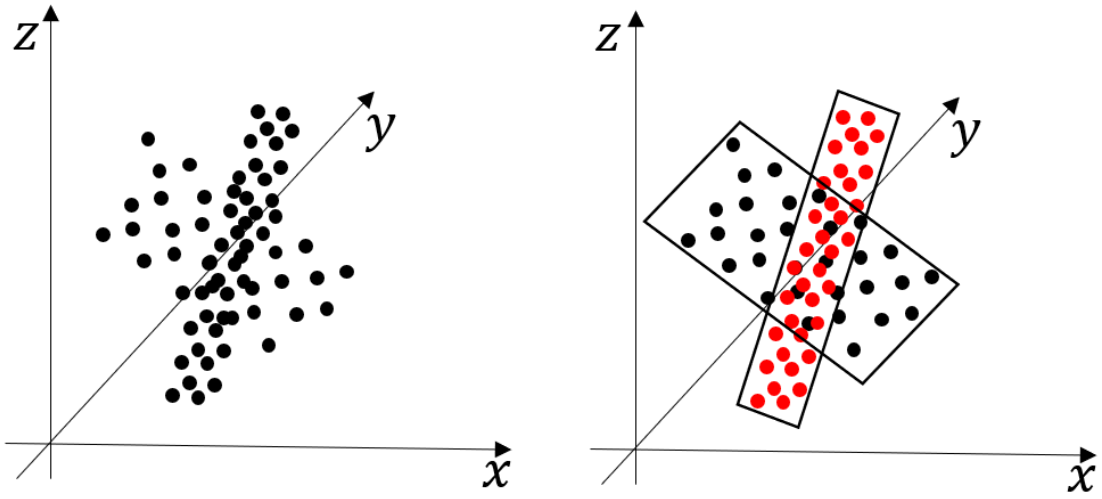

Figure 8.1: Left: a two-dimensional dataset in embedding space exhibiting one dense region. Right: two subspaces to which the one dense region can be distributed to.

As for now, the risk of force-fitting a single dense region into one common subspace can be circumvented by reducing the $minPts$ parameter of the CASH algorithm that is applied to the dense regions. This however bears the risk of a strong fragmentation leading to many small clusters within a dense region that exhibit different orientations and translations of the subspaces. The fragmentation can further be influenced through the parameter j and s that permit the control of the variance within a cluster detected by CASH.

Another behavior of the proposed method that can be observed is that by its order to detect local clusters first followed by a merging to its global cluster, objects that are not belonging to any of the locally dense regions need to be re-evaluated if they belong to the global common subspace. As a future work, it may be considered to circumvent this re-evaluation by starting with the detection of the global subspace first followed by a determination of its local clusters. Lastly, we address the parameters of this framework. While their setting seems intuitive in lower-dimensional cases, it becomes significantly less intuitive in higher dimensionalities.

Chapter 9
On coMADs and Principal Component Analysis

Published in:
Kazempour, Daniyal; Hünemörder, M.A.X and Seidl, Thomas. "On coMADs and Principal Component Analysis." 12th International Conference on Similarity Search and Applications (SISAP). Springer, Cham, 2019: 273-280. https://doi.org/10.1007/978-3-030-32047-8_24

This chapter of the dissertation is dedicated to the topic of resilience against outliers. Certain correlation clustering algorithms do not rely on PCA and have an outlier handling like CASH [1]. Other algorithms of that archetype rely on PCA and have an outlier handling by their underlying model, such as 4C [6]. On the contrary, algorithms like local PCA [17] and ORCLUS [3] rely on PCA and do not have in their native form any outlier handling. This combination exposes their vulnerability towards outliers regarding the computation of the subspaces of a correlation cluster. The endeavor made in this chapter leads to a method that introduces a minimum of changes to the existing Principal Components Analysis while at the same time increasing its robustness towards outliers.

While in this chapter the influence of single outliers or micro-groups of outliers is discussed, it is also of interest for future work to study the impact of neighboring objects, that belong to other clusters, on the orientation and translation of subspaces. More specifically it is of interest to investigate (a) which factors promote or reduce such an influence from neighboring clusters and (b) how the magnitude of their impact can be measured.

Chapter 10
Conclusion and Future Work

It has been about twelve years ago when the work in [34] laid the foundation for correlation clustering. In the work of [34] algorithms were proposed which rely on different views (local, global, hierarchical), and different underlying models (density-based, Hough transform-based, etc.). Further in [34] common and distinctive properties to axis-parallel subspace clustering were investigated and the algorithms were studied concerning their algorithmic properties (i.e., grid-based, top-down, bottom-up). Building on the insights of [34] and looking at the advances made in the field of correlation clustering, we investigated in this dissertation different aspects. First, we dedicated our research in Part I to the Hough transform-based approaches. Within this category, we explicitly researched for methods that are capable of detecting clusters where the relation of the attributes within a cluster is potentially non-linear, as proposed in Chapters 2 and 3. While any arbitrary (non-linear) relation between features can be detected by finding the parameters of a non-linear function, one issue that remains is that it requires scientists a-priori to know the type of non-linearity they are looking for. This issue is also inherent to all kernel-based methods. As another aspect, we investigated strategies for improving the region of interest (ROI) detection within Hough space (Part I, Chapters 4 and 5). While D-MASC [19] is effective on datasets that contain substantial amounts of noise, it is inferior to CASH [1] concerning runtime, when the data contains low amounts of noise. This is owed to the fact that D-MASC has a single fixed resolution for the rasterization procedure. An adaptive resolution may mitigate the runtime performance issue in such a scenario.

The second part of this dissertation was dedicated to the evaluation of correlation clusterings without a given "ground truth". For that purpose, endeavours have been made to derive evaluation measures from the objective functions of existing correlation clustering algorithms (Part II, Chapter 6). Both of the evaluation criteria (NPE and SCS) rely partially on algorithm-specific properties. One drawback of both measures is, that they do not fully penalize model complexity meaning that they minimize the deviation of objects to their subspaces regardless of e.g. how many dimensions or clusters (and thus subspaces) are needed. While SCS penalizes the deviation of objects from their respective subspace with increasing dimensionality of the subspaces, NPE does not penalize increasing dimensionality at all.

The insights from Part II, Chapter 6 raised the question if there exists a common theme among correlation clustering algorithms that can be utilized for an evaluation measure. It turned out that the so far observed correlation clustering algorithms (1) yield clusterings which have (obviously) a number of clusters k (and thus subspaces), (2) where the subspaces of each cluster have a dimensionality l and (3) where the linear subspaces are "fitted"

into each cluster such that the deviation of the objects within a cluster to that subspace is minimized concerning the algorithms internal objective. The assumption on which this evaluation measure relies is that a correlation clustering is considered as favourable if it provides a low reconstruction error [Property (3)] while at the same time maintaining a low model complexity in terms of the required number of clusters (and therefore subspaces) and dimensionality of the subspaces [Properties (1) and (2)]. These three properties are considered for constructing an evaluation measure as proposed in Part III, Chapter 7.

Among the different properties and characteristics is the aspect of the different *views* in correlation clustering (local, global, hierarchical). While in the previous work by [34] algorithms have been proposed which were able to detect local, global, hierarchical or global and hierarchical at the same time, there exists, to the best of our knowledge, no correlation clustering algorithm so far which aims at yielding global *and* local correlation clusters at the same time. Besides this aspect, our contribution in Part III, Chapter 8 is further to provide a formal definition for local and global correlation clusters. We also briefly elaborated on the matter of why it would be not an option to simply run on a given dataset first a local correlation clustering algorithm followed by a global one (or the other way around).

Finally, we investigated the aspect of resilience against outliers (Part III, Chapter 9). Motivated by the observation that either single objects or micro-groups of objects can massively impact the orientation and translation of arbitrarily oriented subspaces, we investigated if there exist, besides the so far present methods, techniques that add a minimum of change to the existing PCA method while significantly contribute to the resilience towards outliers. As such we proposed in Chapter 9 a PCA variant in which the sole modification was to change the covariance matrix to a co-MAD (co-Median Absolute Deviation) matrix, where the co-MAD relies on the idea of the deviation from the median while the covariance relies on the deviation from the mean.

Looking back at the progress made in this work, the question may arise: where to go next from here?

From the findings in this dissertation, the following research opportunities emerge, when revisiting the field of correlation clustering: A different direction to pursue is to investigate if and in how far correlation clustering can be expressed in terms of relational algebra. So far we have an algorithmic view, through the existing methods, and a view of correlation clustering in terms of an optimization problem. For the relational algebra approach, one potential research question is: In how far is the relational description of a correlation clustering method similar to the objective functions by which the clustering is optimized? What are the differences? What are the potential benefits of a relational representation with regards to i.e., runtime? According to [31] the transformation of a machine learning problem into a database problem can be beneficial for both sides, the machine learning side, by improving the efficiency of the methods and the databases side

by extending the influence of database research. Another endeavour should be made in the direction of re-formulating correlation clustering algorithms as a matrix factorization task. Just as k-means can be re-formulated as a matrix factorization problem [5], so, we are convinced, can also correlation clustering algorithms be transformed to such a setting. Here we may also face potential challenges, among them questions such as: Is it possible to perform the optimization by clusters and subspaces within *one* computational step or is an alternating optimization required, i.e., first by the clusters and then by the computed subspaces? We recommend as a prototype algorithm to consider local PCA since this algorithm is quite straightforward and shares similarities to the well known PCA. Further one may ask what the differences and similarities are concerning the objective functions or the relational representations? Are there any differences observable?

Another major aspect to pursue in future work is related to the differences between clusters residing on a single subspace (similarly to PCA) and multiple subspaces as in correlation clustering. A single subspace may have the advantage that inter-cluster relationships can be 'easier' studied, as claimed by the authors in [13]. On the other hand, it may exhibit a poor performance, if the underlying process(es) and thus the subspace(s) around which objects are located are inherently non-linear. In such a case one single linear subspace fails to capture the non-linearities and therefore leads to a larger reconstruction error. In contrast, correlation clustering methods such as 4C and CASH achieve through a piecewise linear approximation, high-quality subspace clusters, but may render it difficult to compare between clusters of a clustering. For future work, we propose to perform multi-level subspace correlation clustering and to compute hierarchies of subspace similarities for investigating inter-cluster relationships. In more detail: On each hierarchy-level subspaces that are most similar are merged to a common cluster and a new 'single subspace' is computed from the newly merged subset of similar subspaces. The aspired benefits are as follows: (1) we can have the 'best of both worlds' (single subspace vs. multiple subspaces) (2) scientists can decide at which similarity level in the dendrogram they wish to investigate the relationships between clusters, and then merge the subspaces, which leads at the end of the dendrogram to a common single optimal subspace.

As further targets for future work, but with minor impact, a Hough-transform based ensemble can be pursued in which a dataset is scanned for a set of distinctive non-linear correlations. The non-linear model which provides the lowest reconstruction error from that non-linear manifold is considered, concerning the model complexity of the function, as the better representative model. In this context it is vital to be cautious regarding the risk of overfitting: While a more complex non-linear model may lower the reconstruction error, its non-linear complexity and the dimensionality of the model severely lead to an overfitting and as such need to be regularized when considering such ensembles.

Regarding the aforementioned matrix factorization formulation of the correlation clustering task, it opens opportunities for parallelization. The so-far existing correlation clustering algorithms are, to the best of our knowledge, not developed or extended for dis-

tributed settings. While there exist axis-parallel subspace clustering algorithms tailored at the MapReduce [8] framework, no methods have been developed for running a correlation clustering algorithm in a distributed setting. A very simple approach was proposed in our previous works [24]. However, a matrix factorization task is predestined to be parallelized on the graphic processing unit (GPU). As such, based on the advances made in matrix-based GPU computations, we think that a matrix factorization reformulation of the correlation clustering task allows developing methods that provide a significant speedup, which is so far harnessed by neural-based approaches. Moving to the aspect of resilience of subspaces towards noise (Chapter 9) the effectiveness of coMAD has been studies to single scattered outliers. It is of interest to investigate how different *distributions* of noise impact the subspaces obtained from a decomposition of the coMAD matrix.

Finally, an open question is to investigate the semantics of clusters in Hough space, which are obtained i.e., like in [22]. What is the meaning of detecting a dense region in Hough space, or a mode, or a cluster hierarchy? What are the semantics behind detecting an axis-parallel subspace cluster in Hough space? Pursuing to find answers to these questions may open up an additional field in the research area of correlation clustering. Lastly, the search for suitable hyperparameter settings in the framework proposed in Chapter 7 is currently done by imposing a hyperparameter grid of a fixed resolution. Depending on the chosen grid resolution finding suitable settings may either be time-consuming or potentially good settings may be missed, as in the case of a too coarse resolution. A future direction is to impose a grid with an adaptive resolution. Areas of the grid which exhibit a significant increase of a score are further refined by an increased grid resolution on that particular sub-area. At this point, it is also of interest to investigate how far this approach shares, from an intuitive point of view, commonalities with the gradient descent approach of neural methods and in how far it can benefit from current state of the art reinforcement techniques.

Coming to an end of this dissertation, and by looking at the increasing numbers of neural-based publications, a question that arises is: what is this 'classical' correlation clustering actually good for? Does it still stand a chance among the plethora of neural-based methods? In our opinion correlation clustering does play a role beyond academic interest. For this we would like to refer to the aspect of explainability. In [25], the issue of current neural methods has been addressed, stating that they lack the capability to *explain* the users how the actual results were obtained. While it may be argued that it is not always of interest to understand how something has been detected or computed, there are however cases (crime detection, recruitment procedures, productivity assessments etc.) where it is of special interest and importance of what actually was learned and how decisions by a machine were made. With the concepts proposed in this work, scientists can indeed follow which piecewise (non-)linear embeddings were learned. They can follow the different views (local and global), and they can comprehend, based on the internal measures, the quality of the clusterings. As a conclusion to the initial question, we can see that the aspect of explainability gives these classic correlation clustering algorithms still a purpose. Just like

in the work of [34], also this work is not meant to be an "end" regarding the research in the field of correlation clustering, but another stage in the journey, as I am convinced that there are many exciting discoveries ahead of us.

Statement of Originality

The research in this dissertation has been conceptualized, implemented and evaluated by the author of this doctoral thesis in cooperation with the supervisor Prof. Dr. Thomas Seidl, researchers and bachelor as well as master students from the Department for Database Systems and Data Mining at the LMU Munich and with a researcher from the Department of Chemistry at the Umeå-University of Sweden.

The work in Part I, Chapter 2, on "Detecting global hyperparaboloid correlated clusters: a Hough-transform based multicore algorithm" as published in [24] has been written in cooperation with Prof. Dr. Peer Kröger, who contributed the introduction and with Dr. Markus Mauder, with whom I had valuable discussions and revisions of, as well as contributions to the source code. The original concept of detecting quadratic correlation clusters using Hough transform, and the conducted experiments as well as the written text of [24] were contributed by the author of this dissertation. Prof. Dr. Thomas Seidl stimulated ideas for a first parallelization of the HCC algorithm.

The following work in Part I, Chapter 3, "Detecting Global Periodic Correlated Clusters in Event Series based on Parameter Space Transform" [20] was written in cooperation with the bachelor student Kilian Emmerig. He conducted experiments which were contributed to [20]. Prof. Dr. Thomas Seidl stimulated in a discussion aspects for the motivation and need for detecting periodic correlated clusters. Prof. Dr. Peer Kröger stimulated the investigation of detecting adequate frequency phase shifts. The conceptualization, writing of the paper and the formalizations were conducted by the author of this work. The code is partially by the author of this work and partially contributed by Kilian Emmerig.

The work on improving the ROI detection in Hough space as proposed in Part I, Chapter 4, "D-MASC: A Novel Search Strategy for Detecting Regions of Interest in Linear Parameter Space" [19] was written in cooperation with the bachelor student Kevin Bein. He performed the implementation of the algorithm, conducted the experiments, provided figures for the experiments and the visualization of the single steps of the algorithm. He also contributed partially in the paper to the investigations on different effects of the parameter settings. Prof. Dr. Thomas Seidl provided the initial idea for the D-MASC approach by introducing the author to the Bresenham algorithm for curve rasterization which gave rise to the method proposed in [19]. Prof. Dr. Peer Kröger stimulated in a discussion to investigate and highlight the conditions under which the CASH algorithm exposes its weaknesses. The conceptualization and design of the experiments, and the writing of major parts of the paper was conducted by the author of this work.

In Part I, Chapter 5, "A Galaxy of Correlations" [22], the publication was made in cooperation with the bachelor student Lisa Krombholz. She performed the implementation and conducted the experiments and contributed the majority of figures in [22]. Prof. Dr. Thomas Seidl stimulated a discussion in which the author became aware that this contribution is not only a modification of search strategy in Hough space, but a different perspective and way of transformation. Prof. Dr. Peer Kröger stimulated in a discussion

to test different "standard" clustering algorithms in Hough space. The authors contribution is the conceptualization and formalization of the idea as well as the writing of the paper.

The work in Part II, Chapter 6, "Towards an Internal Evaluation Measure for Arbitrarily Oriented Subspace Clustering" [23], was conceptualized, implemented, the experiments conducted and written by the author. Prof. Dr. Thomas Seidl contributed by the discussion of which properties internal evaluation measures should satisfy in context of correlation clustering, and to investigate the differences and common properties between ORCLUS and CASH. Prof. Dr. Peer Kröger stimulated the investigation of CASH in terms of an optimization problem.

For the work in Part II, Chapter 7, "I fold you so! An internal evaluation measure for arbitrary oriented subspace clustering" [18] the publication was made in cooperation with the colleague of our chair Anna Beer. She contributed to the formalization part of the paper. Further she provided invaluable contributions in discussions regarding the choice of experiments. Further she reviewed the code and assisted in eliminating errors in the implementation. Prof. Dr. Thomas Seidl provided recommendations and strategies for the experiments, including the message they should convey. Prof. Dr. Peer Kröger stimulated the investigations for further properties which are required for an internal evaluation measure. The author of this work had the initial idea of regarding correlation clustering as an autoencoding problem and to utilize this re-formulation of the problem statement for the evaluation of correlation clustering results. Further the author wrote major parts of the paper, implemented the framework, conducted and discussed the experiments and made the figures.

The work in Part III, Chapter 8, "You see a set of wagons - I see one train: Towards a unified view of local and global arbitrarily oriented subspace clusters" [26] was made in cooperation with the master student Long Matthias Yan. He implemented the proposed algorithm, conducted the majority of the experiments and created the majority of the figures. The author of this dissertation conceptualized this work, wrote the paper, extended parts of the implementation and conducted real-world experiments. Prof. Dr. Thomas Seidl stimulated the investigation of different views under the guiding question in how far both views can be consistently obtained by applying a local and a global correlation clustering algorithm. Prof. Dr. Peer Kröger stimulated the distinctive approaches of top-down and bottom-up strategies for obtaining local *and* global views.

Finally the work in Part III, Chapter 9, "On coMADs and Principal Component Analysis" [21] has been performed in a cooperation with a former colleague of this chair, Maximilian Archimedes Xaver Hünemörder. He conducted the experiments, uncovered an error in the original implementation, created the figures and contributed to the discussions of the results. The author of this work contributed the original idea of using coMAD instead of co-variance matrices and conceptualized this idea to investigate the impact of PCA under the aspect of resilience against outliers.

List of Figures

Bibliography

[1] Achtert, E., Böhm, C., David, J., Kröger, P., and Zimek, A. Global correlation clustering based on the hough transform. *Statistical Analysis and Data Mining: The ASA Data Science Journal 1*, 3 (2008), 111–127.

[2] Achtert, E., Böhm, C., Kriegel, H.-P., Kröger, P., and Zimek, A. Deriving quantitative models for correlation clusters. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* (2006), pp. 4–13.

[3] Aggarwal, C. C., and Yu, P. S. Finding generalized projected clusters in high dimensional spaces. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data* (2000), pp. 70–81.

[4] Ankerst, M., Breunig, M. M., Kriegel, H.-P., and Sander, J. Optics: ordering points to identify the clustering structure. *ACM Sigmod record 28*, 2 (1999), 49–60.

[5] Bauckhage, C. K-means clustering is matrix factorization. *arXiv preprint arXiv:1512.07548* (2015).

[6] Böhm, C., Kailing, K., Kröger, P., and Zimek, A. Computing clusters of correlation connected objects. In *Proceedings of the 2004 ACM SIGMOD international conference on Management of data* (2004), pp. 455–466.

[7] Comaniciu, D., and Meer, P. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on pattern analysis and machine intelligence 24*, 5 (2002), 603–619.

[8] Dean, J., and Ghemawat, S. Mapreduce: simplified data processing on large clusters. *Communications of the ACM 51*, 1 (2008), 107–113.

[9] Duda, R. O., and Hart, P. E. Use of the hough transformation to detect lines and curves in pictures. *Communications of the ACM 15*, 1 (1972), 11–15.

[10] Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd* (1996), vol. 96, pp. 226–231.

[11] Falk, M. On mad and comedians. *Annals of the Institute of Statistical Mathematics 49*, 4 (1997), 615–644.

[12] Fukunaga, K., and Hostetler, L. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on information theory 21*, 1 (1975), 32–40.

[13] Goebl, S., He, X., Plant, C., and Böhm, C. Finding the optimal subspace for clustering. In *2014 IEEE International Conference on Data Mining* (2014), IEEE, pp. 130–139.

[14] Hassani, M., and Seidl, T. Using internal evaluation measures to validate the quality of diverse stream clustering algorithms. *Vietnam Journal of Computer Science 4*, 3 (2017), 171–183.

[15] Hawkins, D. M. *Identification of outliers*, vol. 11. Springer, 1980.

[16] Hough, P. V. Method and means for recognizing complex patterns, Dec. 18 1962. US Patent 3,069,654.

[17] Kambhatla, N., and Leen, T. K. Dimension reduction by local principal component analysis. *Neural computation 9*, 7 (1997), 1493–1516.

[18] Kazempour, D., Beer, A., Kröger, P., and Seidl, T. I fold you so! an internal evaluation measure for arbitrary oriented subspace clustering. In *2020 International Conference on Data Mining Workshops (ICDMW), accepted, to be published* (2020), IEEE, pp. 00–00.

[19] Kazempour, D., Bein, K., Kröger, P., and Seidl, T. D-masc: A novel search strategy for detecting regions of interest in linear parameter space. In *International Conference on Similarity Search and Applications* (2018), Springer, pp. 163–176.

[20] Kazempour, D., Emmerig, K., Kröger, P., and Seidl, T. Detecting global periodic correlated clusters in event series based on parameter space transform. In *Proceedings of the 31st International Conference on Scientific and Statistical Database Management* (2019), pp. 222–225.

[21] Kazempour, D., Hünemörder, M., and Seidl, T. On comads and principal component analysis. In *International Conference on Similarity Search and Applications* (2019), Springer, pp. 273–280.

[22] Kazempour, D., Krombholz, L., Kröger, P., and Seidl, T. A galaxy of correlations - detecting linear correlated clusters through k-tuples sampling using parameter space transform. In *EDBT* (2019), pp. 702–705.

[23] Kazempour, D., Kröger, P., and Seidl, T. Towards an internal evaluation measure for arbitrarily oriented subspace clustering. In *2020 International Conference on Data Mining Workshops (ICDMW), accepted, to be published* (2020), IEEE, pp. 00–00.

[24] Kazempour, D., Mauder, M., Kröger, P., and Seidl, T. Detecting global hyperparaboloid correlated clusters: a hough-transform based multicore algorithm. *Distributed and Parallel Databases 37*, 1 (2019), 39–72.

[25] Kazempour, D., and Seidl, T. Insights into a running clockwork: On interactive process-aware clustering. In *EDBT* (2019), pp. 706–709.

[26] Kazempour, D., Yan, L. M., Kröger, P., and Seidl, T. You see a set of wagons - i see one train: Towards a unified view of local and global arbitrarily oriented subspace clusters. In *2020 International Conference on Data Mining Workshops (ICDMW), accepted, to be published* (2020), IEEE, pp. 00–00.

[27] Kriegel, H.-P., Kröger, P., and Zimek, A. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Transactions on Knowledge Discovery from Data (TKDD) 3*, 1 (2009), 1–58.

[28] Kriegel, H.-P., Kröger, P., and Zimek, A. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Transactions on Knowledge Discovery from Data (TKDD) 3*, 1 (2009), 1–58.

[29] Lloyd, S. Least squares quantization in pcm. *IEEE transactions on information theory 28*, 2 (1982), 129–137.

[30] Mauder, M. Analyzing complex data using domain constraints, June 2017.

[31] Olteanu, D. The relational data borg is learning. *arXiv preprint arXiv:2008.07864* (2020).

[32] Rendón, E., Abundez, I., Arizmendi, A., and Quiroz, E. M. Internal versus external cluster validation indexes. *International Journal of computers and communications 5*, 1 (2011), 27–34.

[33] Theodoridis, S., and Koutroumbas, K. Pattern recognition, 4th edn academic press, 2009.

[34] Zimek, A. *Correlation Clustering*. PhD thesis, Ludwig-Maximilians-Universität München, 2008.

[35] Zimek, A., and Filzmoser, P. There and back again: Outlier detection between statistical reasoning and data mining algorithms. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 8*, 6 (2018), e1280.