Good Clusterings Have Large Volume

Steffen Borgwardt\textsuperscript{a}, Felix Happach\textsuperscript{b}

\textsuperscript{a}Department of Mathematical and Statistical Sciences, University of Colorado Denver, USA
\textsuperscript{b}Fakultäten für Mathematik und Wirtschaftswissenschaften, Technische Universität München, Germany

Abstract

The clustering of a data set into disjoint clusters is one of the core tasks in data analytics. Many clustering algorithms exhibit a strong contrast between a favorable performance in practice and bad theoretical worst-cases. Prime examples are least-squares assignments and the popular \textit{k}-means algorithm.

We are interested in this contrast and approach it using methods of polyhedral theory. Several popular clustering algorithms are readily connected to finding a vertex of the so-called bounded-shape partition polytopes. The vertices correspond to clusterings with extraordinary separation properties, in particular allowing the construction of a separating power diagram such that each cluster has its own cell.

The geometric structure of these polytopes reveals useful information: First, we are able to quantitatively measure the space of all sites that allow construction of a separating power diagrams for a clustering by the volume of its normal cone. This gives rise to a new quality criterion for clusterings, and explains why good clusterings are also the most likely to be found by some classical algorithms.

Second, we characterize the edges of the bounded-shape partition polytopes. Through this, we obtain an explicit description of the normal cones. This allows us to compute measures with respect to the new quality criterion, and even compute “most stable” sites for the separation of clusters. The hardness of these computations depends on the number of edges incident to a vertex, which may be exponential. However, this computational effort is rewarded with a wealth of information that can be gained from the results, which we highlight using a toy example.

Keywords: clustering; linear programming; power diagram; polyhedron; normal cone; stability

1. Introduction

Informed decision-making based on large data sets is one of the big challenges in operations research. We are interested in one of the fundamental tasks in data analytics: The clustering of a data set into disjoint clusters. Data is often represented as a finite set \( X \subset \mathbb{R}^d \) in \( d \)-dimensional Euclidean space. A clustering \( C = (C_1, \ldots, C_k) \) then is a partition of the set \( X \) into parts \( C_i \subset X \), such that \( \bigcup_{i = 1}^k C_i = X \) and \( C_i \cap C_j = \emptyset \).

There is a wealth of literature on clustering methods. We refer to three surveys \cite{9, 25, 37}. For most clustering algorithms, there is a strong contrast between an extremely favorable behavior in practice and the lack of provable guarantees on their performance in theory. A prime example is the popular \textit{k}-means algorithm \cite{27, 28}. It exhibits a stunning discrepancy between its excellent behaviour in practice and its known worst-case behaviour: In practice, it typically terminates in just a few iterations and produces human-interpretable results. In a theoretical worst-case, it may take exponentially many iterations even for two-dimensional data \cite{36}, and it is easy to construct artificial examples for which its results do not capture the structure of the underlying data at all. In the present paper, we use methods of polyhedral theory to better understand this difference between the practical performance and the theoretical worst-case behaviour.

The studies of polyhedra arising for applications have been a popular approach in operations research. There are numerous cases where the combinatorial properties of these polyhedra revealed deeper insight into the underlying application \cite{2, 13, 15, 16, 21, 22, 26, 31, 33}. For an introduction to polyhedral theory, we recommend \cite{20, 32, 38}. Further,
Figure 1: A separating power diagram for the blue, red, and black cluster. The data points of each cluster lie in the interior of their respective cells. The small dots indicate the sites.

we refer the reader to the book review [34] of Ziegler's classical textbook Lectures on Polytopes [38] for an overview and further literature on the power of polyhedral theory for applications.

The so-called assignment polytopes are closely related to our setting and have been studied well [6, 19, 20]. It is possible to represent the partition of data points \( X = \{x_1, \ldots, x_n\} \) into clusters \( C_1, \ldots, C_k \) by using decision variables \( y_{ij} \) that indicate whether data point \( x_j \) is assigned to cluster \( C_i (y_{ij} = 1) \) or not \( (y_{ij} = 0) \). Further, many applications specify lower bounds \( s^-_i \) and upper bounds \( s^+_i \) on the number of points that may be assigned to cluster \( C_i \). This gives rise to a simple set of linear constraints that describe all clusterings:

\[
\begin{align*}
    s^-_i & \leq \sum_{j=1}^n y_{ij} \leq s_i \quad (i \leq k) \\
    \sum_{i=1}^k y_{ij} & = 1 \quad (j \leq n) \\
    y_{ij} & \geq 0 \quad (i \leq k, j \leq n).
\end{align*}
\]

The first set of constraints makes sure the prescribed cluster size bounds are respected, the second set of constraints guarantees that each data point is assigned to a cluster. With the relaxed constraints \( y_{ij} \geq 0 \), we obtain a polytope \( P \). The left-hand sides of the constraints form a totally unimodular matrix, and the right-hand sides are integral, so the vertices of this polytope are 0, 1-vectors, i.e. all \( y_{ij} \) satisfy \( y_{ij} \in \{0, 1\} \). Each vertex describes a clustering, and vice versa.

The polytope \( P \) is a useful way to describe all clusterings, but we are particularly interested in “good” clusterings. To this end, instead of considering the above polytope, we study a special type of projection first introduced in [7, 24], the so-called bounded-shape partition polytope. (For a formal definition, see Section 2.3.) The vertices of bounded-shape partition polytopes exhibit several favorable properties [7], such as being a minimizer of the least-squares functional among all clusterings of the same cluster sizes. But most importantly, these vertices have strong separation properties: They allow for the construction of a separating power diagram, a generalized Voronoi diagram in \( \mathbb{R}^d \) with one polyhedral cell for the data points for each cluster [11]. In machine learning and other parts of operations research, this separation property is sometimes called piecewise-linear separability [8]. See Figure 1 for a small example.

Contributions and Outline

In Section 2, we introduce some notation and review related work. We then use the known vertex characterization of [7] as a starting point for the new contributions in this paper. We briefly mention these contributions in the following paragraphs. Section 3 is a complete and self-contained presentation of our main results. Section 4 contains the necessary proofs. We conclude our discussion with a short outlook in Section 5. Some of the presented work is based on the authors’ dissertation [11] and master’s thesis [23].

Section 3.1. First, we devise a new measure for the quality of a clustering, which is quite different from the stability measures used in the literature [35]. We call it the volume of a clustering, since it is directly related to the volume of the normal cone of the vertex encoding this particular clustering. A large volume corresponds with a clustering of high quality, which provides an informal explanation why some clustering algorithms work well in practice: For example, the computation of a least-squares assignment for fixed cluster sizes is in one-to-one correspondence to linear
optimization over a single-shape partition polytope [11]. When choosing random sites, it is likely to find a clustering of large volume. The $k$-means algorithm can be interpreted as the repeated computation of a least-squares assignments with changing sites in each iteration. Informally, it is likely to terminate with a clustering of large volume — the good clusterings are found most likely.

Section 3.2. Second, we discuss the structure of the normal cone. The key requirement for this is to have an explicit representation of the normal cone. As there is no explicit description of the facets of the bounded-shape partition polytope — it is only defined as the projection of another polytope — this is a difficult task. Nonetheless, we are able to find an explicit representation: In a generalization of previous results [11] [17], we characterize the edges of all bounded-shape partition polytopes which, informally, correspond to so-called movements or cyclical movements of items between clusters. This characterization allows us to construct the inner cone of a vertex, which is the polar of the normal cone. Herewith we obtain the desired explicit representation of the normal cone which allows us to investigate the structure of the site vectors located in the interior of the normal cone. This allows us to identify certain convex areas which contain the sites of all representatives of all site vectors inducing this clustering — up to scaling, which does not change the clustering. We provide some proof-of-concept computations and a running example, in which we compare clusterings of different volumes.

Section 3.3. Finally, we introduce a new stability criterion for sites for a clustering and provide an algorithm for the computation of optimal sites in the sense of this stability criterion. These sites are maximally stable with respect to perturbation, i.e. all sites can be perturbed in any direction with a largest possible amount while staying in the normal cone. We use a classical approach from computational geometry to find such sites: We roll a $p$-norm unit ball into the normal cone, with the origin of the cone as gravity source. We then compute where it gets “stuck”. The center of the ball gives the desired sites. This computation is readily expressed as a mathematical program. Hardness of this computation comes from the fact that there can be exponentially many edges. Of course, this hardness is not surprising — there are related problems, like the $k$-means problem, for which the complexity of finding a global optimum (globally optimal sites) is known to be $\text{NP}$-hard, even for $k = 2$ [1] or for data in the Euclidean plane [29].

2. Preliminaries

We begin with some standard notation.

2.1. Clusterings, Least Squares Assignments and Power Diagrams

Let throughout this paper $n,d,k \in \mathbb{N} := \{1,2,\ldots\}$ be fixed. Let $X := \{x_1,\ldots,x_n\} \subseteq \mathbb{R}^d$ be a set of $n$ distinct non-zero data points and for $m \in \mathbb{N}$ define $[m] := \{1,\ldots,m\}$. We call a partition $C := (C_1,\ldots,C_k)$ of $X$ a clustering and denote by $|C| := (|C_1|,\ldots,|C_k|)$ its shape. For $i \in [k]$, we call $C_i$ the $i$-th cluster of $C$ and $|C_i|$ its size. Let $s^- := (s_1^-,\ldots,s_k^-)$, $s^+ := (s_1^+,\ldots,s_k^+)$ $\in \mathbb{R}^k$ such that $0 \leq s_i^- \leq s_i^+ \leq n$ for all $i \in [k]$ be the lower and upper bounds on the cluster sizes.

A clustering $C$ is said to be feasible if it satisfies $s^- \leq |C| \leq s^+$ componentwisely. We will only consider feasible clusterings in this paper. $C$ is called separable if all pairs of clusters are linearly separable, i.e. for all $i,j \in [k]$, $i \neq j$ there is $a_{ij} \in \mathbb{R}^d$ and $\gamma_{ij} \in \mathbb{R}$ such that $a_{ij}^Tx \leq \gamma_{ij} \leq a_{ij}^Ty$ for all $x \in C_i$, $y \in C_j$. The hyperplane separating the clusters is denoted by $H_{(a_{ij},\gamma_{ij})} := \{x \in \mathbb{R}^d | a_{ij}^Tx = \gamma_{ij}\}$. Analogously, we define $H_{(a_{ij},\gamma_{ij})}^\circ$, $H_{(a_{ij},\gamma_{ij})}^\varepsilon$ and $H_{(a_{ij},\gamma_{ij})}^\ell$ to be the respective (open) half-spaces. A constrained least squares assignment (LSA) to certain sites $a_1,\ldots,a_k \in \mathbb{R}^d$ is a clustering $C = (C_1,\ldots,C_k)$ minimizing

$$\sum_{i=1}^k \sum_{x \in C_i} \|x - a_i\|^2$$

(1)

over all clusterings with the same shape as $C$. If $C$ minimizes (1) over all feasible clusterings, we say $C$ is a general LSA. We call $a := (a_1^T,\ldots,a_k^T)^T \in \mathbb{R}^{d \times k}$ the site vector. Note that there is a close connection between (constrained) least squares assignments and so-called power diagrams [4][11]. Recall that power diagrams are a generalization of the well-known Voronoi diagrams [6]. There are several equivalent ways to define a power diagram. We briefly recall the definition that is most convenient for our purposes [12]. Let $a := (a_1^T,\ldots,a_k^T)^T \in \mathbb{R}^{d \times k}$ be a site vector with distinct sites $a_1,\ldots,a_k$ and let $\alpha_1,\ldots,\alpha_k \in \mathbb{R}$. For $i \in [k]$, we call

$$P_i := \{x \in \mathbb{R}^d | (a_j - a_i)^T x \leq \alpha_i - \alpha_j \text{ for all } j \in [k] \setminus \{i\}\}$$

Page 3
the \(i\)-th (power) cell of the power diagram \((P_1, \ldots, P_k)\). For a convex set \(A \subseteq \mathbb{R}^d\), we call \(\text{int}(A)\) the interior and \(\text{bd}(A)\) the boundary of \(A\). Aurenhammer et al. [4] showed the following connection between constrained LSAs and power diagrams: If \(C\) is a constrained LSA to the site vector \(a\), then there is a power diagram with site vector \(a\) satisfying \(C_i \subseteq \text{int}(P_i)\) for all \(i \in [k]\). On the other hand, if a power diagram \((P_1, \ldots, P_k)\) with site vector \(a\) satisfies \(C_i \subseteq P_i\) for all \(i \in [k]\), then \(C\) is a constrained LSA to the site vector \(a\). If \(C_i \subseteq P_i\) \(\forall i \in [k]\), then we say the power diagram (strongly) induces the clustering and we speak of a separating power diagram. We want to stress the strength of this separation property, which is stronger than just the existence of separating hyperplanes. The constructed cells also partition the underlying space.

2.2. Movements Between Clusterings

In order to compare two clusterings \(C := (C_1, \ldots, C_k), C' := (C'_1, \ldots, C'_k)\), we use the clustering difference graph (CDG), c.f. [11], which is the labeled directed multigraph \(\text{CDG}(C, C') := (V, E)\) with node set \(V := [k]\) and edge set \(E\) constructed as follows: For each \(x_j \in C_i \cap C'_i\) with \(i, l \in [k]\) and \(i \neq l\), there is an edge \((i, l) \in E\) with label \(x_j\). W.l.o.g. we delete isolated nodes in the CDG. These would correspond to clusters that are identical in \(C\) and \(C'\). Recall that a node is isolated, if it is not incident to any edge. We can derive the clustering \(C'\) from \(C\) by applying operations corresponding to the edges of \(\text{CDG}(C, C')\):

Let \((i_1, i_2) - (i_2, i_3) - \cdots - (i_t, i_{t+1})\) be an edge path in \(\text{CDG}(C, C')\) with labels \(x_{j_1}, \ldots, x_{j_t}\). Applying the movement

\[
M : \quad C_{i_1} \xrightarrow{x_{j_1}} C_{i_2} \xrightarrow{x_{j_2}} \cdots \xrightarrow{x_{j_t}} C_{i_{t+1}}
\]

to \(C\) means deriving the clustering \(\tilde{C} := (\tilde{C}_1, \ldots, \tilde{C}_k)\) by setting \(\tilde{C}_l := (C_{i_l} \setminus \{x_{j_l}\}) \cup \{x_{j_{l-1}}\}\) for all \(l \in \{2, \ldots, t\}\), \(\tilde{C}_1 := C_{i_1} \setminus \{x_{j_1}\}\), \(\tilde{C}_{i_{t+1}} := C_{i_{t+1}} \cup \{x_{j_t}\}\) and \(\tilde{C}_l := C_{i_l}\) for all \(r \in [k]\) \(\setminus\{i_1, \ldots, i_{t+1}\}\),

If \(i_{t+1} = i_1\), i.e. in case of a cycle, we speak of a cyclical movement. We then obtain \(\tilde{C}_{i_1} := (C_{i_1} \setminus \{x_{j_1}\}) \cup \{x_{j_t}\}\) and all cluster sizes remain the same. The inverse (cyclical) movement \(M^{-1}\) is defined via the corresponding path (cycle) \((i_{t+1}, i_1) - (i_1, i_{t-1}) - \cdots - (i_2, i_1)\) in \(\text{CDG}(C', C)\). Clearly, one can obtain any clustering from any other clustering by (greedily) decomposing their clustering difference graph into paths and cycles and applying the corresponding (cyclical) movements to \(C\). If both clusterings have the same shape, then their CDG decomposes into cycles, i.e. only cyclical movements are required to transform one into the other [11].

The following figures show a clustering of twelve data points in \(\mathbb{R}^2\) with \(C_1\) black, \(C_2\) blue, \(C_3\) red, as well as two clusterings which can be derived from it by applying a movement (Figure 2) and a cyclical movement (Figure 3), respectively. The convex hull of each cluster is shaded in the respective color in order to highlight changes.

2.3. Bounded-Shape and Single-Shape Partition Polytopes

Let \(N_P(v) := \{a \in \mathbb{R}^d \mid a^T v \geq a^T x \forall x \in P\}\) be the normal cone of a polytope \(P\) at \(v \in P\). Let \(\text{lin}(A)\) denote the linear subspace spanned by a set \(A \subseteq \mathbb{R}^d\). Note that, if \(F\) is a face of the normal cone \(N_P(v)\), then \(F \cap \text{lin}(P - v)\) is a face of \(N_P(v) \cap \text{lin}(P - v)\).

The polytope we are investigating was introduced in [7] and further analyzed in [24]. For a clustering \(C := (C_1, \ldots, C_k)\) and \(i \in [k]\), let \(\sigma_i := \sum_{x \in C_i} x \in \mathbb{R}^d\). The clustering vector of \(C\) is \(w(C) := (\sigma_1^T, \ldots, \sigma_k^T)^T \in \mathbb{R}^{dk}\) and \(\mathbb{R}^d (X, \bar{s}^-, \bar{s}^+) := \text{conv}\{\{w(C) \mid C \text{ is feasible }\}\}\) is called the bounded-shape partition polytope. If \(s = \bar{s}^- = \bar{s}^+\), then we speak of the

![Figure 2: Application of a movement (left) and the CDG (right).](image)
The bounded-shape partition polytope $P^- = P^\pm (X, k, s, s)$, The bounds will typically be clear from the context. In this case, we use the simpler notation $\mathcal{P}^\pm = \mathcal{P}^\pm (X, k, s^-, s^+)$ and $\mathcal{P}^\pm = \mathcal{P}^-$. Note that $\mathcal{P}^\pm$ is a projection of the generalized assignment polytope investigated in [20]. We want to stress that, since the bounded-shape and single-shape partition polytopes are defined as a convex hull, we do not have explicit information on (facet-defining) valid inequalities. This will be important in our discussion later on, as we have to study the edge structure of the polytopes (which may have exponential size) in order to construct the normal cones. The single-shape partition polytopes are of special interest for our purposes, since computation of an optimal constrained LSA corresponds precisely to linear optimization over the corresponding single-shape partition polytope $\mathcal{P}^-$. Another interesting special case is the all-shape partition polytope $\mathcal{P}^\pm$ investigated in [17] which is obtained by choosing $s^- = (0, \ldots, 0)$ and $s^+ = (n, \ldots, n)$. We obtain the following connection between single-shape and bounded-shape partition polytopes [23].

**Lemma 1** The bounded-shape partition polytope is equal to the convex hull of single-shape partition polytopes with feasible shapes, i.e.

\[ \mathcal{P}^\pm (X, k, s^-, s^+) = \text{conv}\left( \bigcup_{s^- \leq s \leq s^+} \mathcal{P}^- \right). \]

**Proof.** Denote by $U$ the union of all single-shape partition polytopes, i.e.

\[ U = \bigcup_{s^- \leq s \leq s^+} \mathcal{P}^- = \text{conv}(U). \]

Let $w(C) \in \mathcal{P}^\pm$ be a clustering vector. Then $w(C) \in \mathcal{P}^\pm (\mathcal{C})$ and, since $C$ is a feasible clustering, $w(C) \in U$. This holds for all clustering vectors and, as $\mathcal{P}^\pm$ is the convex hull of all clustering vectors, we obtain $\mathcal{P}^\pm \subseteq \text{conv}(U)$. On the other hand, clearly $\mathcal{P}^- \subseteq \mathcal{P}^\pm$ for all $s^- \leq s \leq s^+$. Thus, $U \subseteq \mathcal{P}^\pm$. Taking the convex hull on both sides yields $\text{conv}(U) \subseteq \text{conv}(\mathcal{P}^\pm) = \mathcal{P}^\pm$ where the last equality is due to the fact that the bounded-shape partition polytope is convex. \)

Note that Lemma 1 implies that every vertex $w(C)$ of $\mathcal{P}^\pm$ is a vertex of $\mathcal{P}^\pm (\mathcal{C})$ and that the normal cone of a vector in the bounded-shape partition polytope is contained in the one of the single-shape partition polytope, i.e. $N_{\mathcal{P}^\pm} (w(C)) \subseteq N_{\mathcal{P}^-} (w(C))$. In 1992, Barnes et al. [17] gave a concrete characterization of the vertices of $\mathcal{P}^\pm$.

**Proposition 2 (Barnes, Hoffman, Rothblum 1992)** The clustering vector $w(C)$ of a clustering $C$ is a vertex of $\mathcal{P}^\pm$, if and only if there are $a := (a_1, \ldots, a_k)^T \in \mathbb{R}^{d \times k}$ and $\alpha_1, \ldots, \alpha_k \in \mathbb{R}$ satisfying the following statements.

(i) If $|C_i| > s_i^-$ for $i \in [k]$, then $\alpha_i \leq 0$.

(ii) If $|C_i| < s_i^+$ for $i \in [k]$, then $\alpha_i \geq 0$.

(iii) If $x_l \in C_i$ for $l \in [n], i \in [k]$, then for all $j \in [k] \setminus \{i\}$

\[ (a_j - a_i)^T x_l < \alpha_i - \alpha_j. \]
A proof is given in [7] and with more technical details in [24]. We call a clustering $C$ that corresponds to a vertex a **vertex clustering**. Condition (iii) states linear separability of the clusters of a vertex clustering with separation directions $a_{ij} := a_j - a_i \in \mathbb{R}^d$ and right hand sides $s_j := a_\alpha - a_j$ for all $i, j \in [k]$. This implies the existence of a separating power diagram. If the scalars $a_\alpha$,...,$a_k$ additionally satisfy conditions (i) and (ii) then the resulting separation fulfills some further extraordinary properties: For example, if two clusters satisfy $s^- < |C_i|, |C_j| < s^+$, then $C_i$ and $C_j$ are “0-separable” [5], in particular they can be separated by a hyperplane containing the origin. Any vector $a \in \text{int}(N_{\mathcal{D}^\pm}(w(C)))$ can be chosen to construct suitable $a_\alpha$,...,$a_k$ such that the properties of Proposition 2 are satisfied [7]. This gives the following Corollary 3 [11],[23].

**Corollary 3** Let $C := \{C_1, \ldots, C_k\}$ such that $w(C)$ is a vertex of $\mathcal{D}^\pm$ and let $a := (a_1^T \ldots, a_k^T)^T \in N_{\mathcal{D}^\pm}(w(C)) \subseteq \mathbb{R}^{dk}$. Then there is a separating power diagram $(P_1, \ldots, P_k)$ with site vector $a \in \mathbb{R}^{dk}$ such that $C_i \subseteq P_i$ for all $i \in [k]$. If $a \in \text{int}(N_{\mathcal{D}^\pm}(w(C)))$, then this power diagram satisfies $C_i \subseteq \text{int}(P_i)$ for all $i \in [k]$. If we choose $a \in \text{bd}(N_{\mathcal{D}^\pm}(w(C))) \subseteq \mathbb{R}^{dk}$, then there is an index $i \in [k]$ such that $C_i \cap \text{bd}(P_i) \neq \emptyset$.

Further, the normal cone of a vertex clustering of the single-shape partition polytope encodes exactly all site vectors that allow construction of a separating power diagram [11]. Clearly, any positive scaling $\lambda a$ of a site vector $a$ in the normal cone stays in the normal cone. Thus $a$ and $\lambda a$ yield the same constrained LSA. This was first proven by Aurenhammer et al. [4] (without using polyhedral theory).

**Theorem 4 (Aurenhammer, Hoffmann, Aronov 1998)** Let $a \in \mathbb{R}^{dk}$ be site vector of a constrained LSA. For all $\lambda > 0$, the site vectors $\lambda a$ yield the same constrained LSA.

Before we state our main results, note that our assumption that the zero vector is not contained in $X$ is not a restriction, since the overall structure of a data set is not changed when translating the whole set by the same vector. This will make some of our arguments easier in the following. Moreover, we can interpret any movement as the translation of the clustering vertex. Let $C, C'$ such that $CDG(C, C')$ is a path or cycle corresponding to a (cyclical) movement $M$. Then the difference of the clustering vectors $w(M) := w(C') - w(C)$ is called the **vector of the movement** $M$. Note that, if $w(M)$ is the vector of a movement, then the vector of the inverse movement is given by $w(M^{-1}) = -w(M)$.

### 3. Main results

We begin each of the three section with a brief overview.

**3.1. Volume of Clusterings**

**Overview:** A vertex of a bounded-shape partition polytope can be computed efficiently by linear programming. The linear objective vector $a = (a_1^T \ldots, a_k^T)^T \in \mathbb{R}^{dk}$ itself lists a set of sites $a_i \in \mathbb{R}^d$ for the construction of a separating power diagram for the clusters [7][11][24]. In fact, the normal cone of the vertex with respect to the single-shape partition polytope represents precisely all the sites that can be used for the construction of a separating power diagram for this clustering. Sites in the normal cone of a vertex of the bounded-shape partition polytope satisfy even stronger separation properties. Power diagrams are invariant under scaling in the sense that, a power diagram constructed for sites $a_1, \ldots, a_k$ can equivalently be expressed as a power diagram constructed for any sites $\lambda \cdot a_1, \ldots, \lambda \cdot a_k$ for all $\lambda > 0$. Combining all these properties allows us to **quantitatively measure the space of all sites** that allow construction of a separating power diagram for a given clustering by the **volume of its normal cone**. This gives rise to a quality measure that we call the **volume of a clustering**.

Instead of looking at each site vector $a$ individually, we consider its equivalence class $[a] := \{ \lambda a | \lambda > 0 \}$ and choose the unit vector $\frac{1}{\|a\|_2} a$ as a representative. This allows us to introduce a notion of the “distance of sites”. Let $L(\gamma)$ be the length of a curve $\gamma$ and $\mathbb{S}^{d-1} := \{ x \in \mathbb{R}^d | \|x\|_2 = 1 \}$ be the Euclidean unit sphere.

**Definition 5 (Distance of Sites)** Let $a, a' \in \mathbb{S}^{d-1}$ be two site vectors. The distance of the equivalence classes $[a]$ and $[a']$ is defined as the distance of the site vectors on the unit sphere, i.e.

$$d(a, a') := \inf \{ L(\gamma) | \gamma : [0;1] \mapsto \mathbb{R}^d, \gamma(0) = a, \gamma(1) = a', \gamma(t) \in \mathbb{S}^{d-1} \forall t \in [0;1] \}. $$
Note that $d : \mathbb{S}^d \times \mathbb{S}^d \to \mathbb{R}$ is a metric, takes values between 0 and 1 and that an infimum always exists, because $L(\gamma) \geq 0$ for all $\gamma : [0; 1] \to \mathbb{R}^d$.

If a vertex clustering $C$ of $\mathcal{P}^{\pm}$ or $\mathcal{P}^\ast$ has a large normal cone, then it is likely that a randomly chosen site vector lies in its cone, so by previous observations, the chosen site vector defines a separating power diagram inducing $C$. We are interested in measuring the volume of the normal cones of the bounded-shape and single-shape partition polytopes in order to characterize “good” clusterings.

We follow the notation of Bonifas et al. [10] who used the volume of normal cones for the investigation of polytope diameters. For a cone $K \subseteq \mathbb{R}^d$, we call $B(K) := K \cap \mathbb{S}^d$ the base of $K$. The volume of $K$ is defined as the $(d \cdot k - 1)$-dimensional volume of $B(K)$ and is denoted by $\text{vol}(K)$. A set $A \subseteq \mathbb{S}^d$ is called spherically convex, if for all $x, y \in A$ the geodesic $\gamma : [0; 1] \to \mathbb{S}^d$ connecting $x$ and $y$ with $\gamma(0) = x$ and $\gamma(1) = y$ is also contained in $A$. Recall that a geodesic is a curve on the sphere with shortest length. Note that the base of a cone is itself spherically convex. These notions allow use to introduce a new term: the volume of a clustering.

**Definition 6 (Volume of a Clustering)** Let $C$ be a feasible clustering and let $N^{\pm}(w(C))$ and $N^\ast(w(C))$ be its normal cones of $\mathcal{P}^{\pm}$ and $\mathcal{P}^\ast$, respectively. We define

$$\mu_\pm(C) := \frac{\text{vol}(N^{\pm}(w(C)))}{\text{vol}(\mathbb{R}^d)}$$

**to be the BHR volume of $C$ (as a tribute to the first authors discussing this polytope [7]) and**

$$\mu^\ast(C) := \frac{\text{vol}(N^\ast(w(C)))}{\text{vol}(\mathbb{R}^d)}$$

**to be the LSA volume of $C$.**

The volumes of a clustering put the volume of the respective normal cones in relation to the volume of the whole space. By definition, $\mu_\pm, \mu^\ast \in [0; 1]$, since $\mathbb{R}^d$ is a cone containing all cones in $\mathbb{R}^d$ and $	ext{vol}(\mathbb{R}^d)$ equals the area of the surface of $\mathbb{S}^d$. Note that, in practice, permutation of the clusters/sites yield the same clustering and the respective normal cones have the same volume. For our theoretical purposes, it is more useful to consider each permutation as an individual clustering. However, note that, if all cluster bounds are symmetric, then the volumes defined above only take values between 0 and $\frac{1}{2}$, because they only take one possible permutation into account. We would like to stress that, depending on $k$, even small values already represent good clusterings. By Lemma [1], $\mu_\pm(C) \leq \mu^\ast(C)$ for all clusterings $C$.

It does not matter whether we consider the whole space or just the affine hull of the polytope in Definition 6, because all normal cones have the same lineality space equal to $\text{lin}(\mathcal{P}^\ast - v)^\perp$ with $v \in \mathcal{P}^\ast$ for $\ast \in \{\pm, \ast\}$, respectively. Since we consider the volume of a cone relative to the whole space, we can restrict ourselves to the affine hull of the polytopes. For further details on the restriction to the affine hull we refer to [23].

In fact, one can show that $\mathcal{P}^\pm$ and $\mathcal{P}^\ast$ are contained in a $(d \cdot (k - 1))$-dimensional affine subspace. In order to see that $\mathcal{P}^\ast$ for $\ast \in \{\pm, \ast\}$ is not full-dimensional, consider a vector $a = (\vec{a}^T, \ldots, \vec{a}^T)^T \in \mathbb{R}^d$ with $\vec{a} \in \mathbb{R}^d$ and an arbitrary clustering vector $w(C)$. Then

$$\vec{a}^T w(C) = \sum_{i=1}^k \vec{a}^T \sigma_i = \sum_{i=1}^k \sum_{c_i \in C_i} \vec{a}^T x = \sum_{j=1}^n \vec{a}^T x_j,$$

so $\mathcal{P}^\ast \subseteq \{x \in \mathbb{R}^d \mid (\vec{a}^T, \ldots, \vec{a}^T)x = \sum_{j=1}^n \vec{a}^T x_j\}$. The dimension of the set of all vectors consisting of $k$ copies of $d$-dimensional vectors is equal to $d$, so the dimension of $\mathcal{P}^\ast$ is at most $d \cdot k - d = d \cdot (k - 1)$.

**Theorem 7** The measures $\mu_\pm$ and $\mu^\ast$ are well-defined. For $\ast \in \{\pm, \ast\}$, $C$ is a vertex clustering of $\mathcal{P}^\ast$ if and only if $\mu_\pm(C) > 0$ and then

$$\mu_\ast(C) = \frac{\Gamma\left(\frac{d+1}{2}\right)}{2\pi^{\frac{d}{2}} \text{bd}[B(N^\ast(w(C)))]} \int_{\text{int}(N^\ast(w(C))) \cap \mathbb{S}^d} d(z; a), \quad (2)$$

with $z \in \text{int}(N^\ast(w(C))) \cap \mathbb{S}^d$ and $\Gamma$ being the well-known Gamma function.

Note that the coefficient of the integral in (2) is just the inverse of the area of the surface of $\mathbb{S}^d$. We postpone the proof of this theorem to Section 4.1.
3.2. Site Vectors in the Normal Cone

Overview: In this section we derive an explicit description of the normal cone of a vertex clustering of the bounded-shape partition polytope via a characterization of the edges of the polytope. This characterization generalizes previously known results on several special cases [11, 17, 18]. As a corollary, we also obtain a description for the normal cones of the single-shape partition polytopes. Informally, two clusterings of the polytope belong to neighboring vertices if they differ by only a single movement or cyclical movement. Further, there has to exist a set of sites for which both clusterings allow construction of a separating power diagram. The explicit information on the normal cone allows to represent the set of sites that define a separating power diagram for a vertex clustering as a convex area in the original space of the data set. These areas contain a representative of each equivalence class of site vectors that induce the clustering. A running example of two clusterings with different volumes is used for an illustration.

It is well-known that the edges incident to a vertex of a polytope are normal vectors to the facets of the normal cone. It is well-known that the edges incident to a vertex of a polytope are normal vectors to the facets of the normal cone. It is well-known that the edges incident to a vertex of a polytope are normal vectors to the facets of the normal cone. It is well-known that the edges incident to a vertex of a polytope are normal vectors to the facets of the normal cone.

The most closely related characterization was done by Fukuda et al. [17] who explicitly stated the neighborhood of a vertex \( w(C) \) of the all-shape partition polytope \( \mathcal{P} \). They showed that edges incident to \( w(C) \) correspond to movements of the form \( C_i \xrightarrow{\lambda} C_j \) or \( C_i \xleftarrow{\lambda} C_j \) with \( \lambda > 0 \) and \( i, j \in [k], i \neq j \). Moreover, they proved that, if there are \( x, y \in X \) and \( \lambda > 0 \) such that \( x = \lambda y \), then \( x \) and \( y \) are in the same cluster in a vertex clustering, so one can assume that no such pair exists. In particular, if there are no multiplies in the point set \( X \), i.e. \( \text{lin}(\{x\}) \cap X = \{x\} \) for all \( x \in X \) then all edges of \( \mathcal{P} \) correspond to movements which move a single element from one cluster to another. The following theorem extends these results to general lower and upper bounds \( s^- \) and \( s^+ \).

**Theorem 8** Let \( C := (C_1, \ldots, C_k), C' := (C'_1, \ldots, C'_k) \) be two clusterings such that \( w(C) = C_{i_1} \) and \( w(C') = C_{i'_1} \) are adjacent vertices of \( \mathcal{P} \). Let further no three points in \( X \) lie on a single line. Then \( C \) and \( C' \) differ by a single (cyclical) movement or by two movements and there are distinct \( i, j \in [k] \) such that both are of the form \( C_i \rightarrow C_j \).

The different cases that can occur for the clustering difference graph \( CDG(C, C') \) according to Theorem 8 are depicted in Figure 4. Note that the case of two movements is degenerated and that this can only occur, if all sites lie on a line. For the single-shape partition polytope, we obtain the following corollary [11].

**Corollary 9** Let \( C := (C_1, \ldots, C_k), C' := (C'_1, \ldots, C'_k) \) be two clusterings such that \( w(C) = C_{i_1} \) and \( w(C') = C_{i'_1} \) are adjacent vertices of \( \mathcal{P} \). If no four points in \( X \) lie on a single line, then \( C \) and \( C' \) only differ by a single cyclical movement.

We dedicate Section 4.2 to the proof of this theorem and corollary. Figure 5 shows two vertex clusterings of \( \mathcal{P} \) which are connected by an edge. The site vector of the plotted power diagram lies in the (boundary of the) normal cones of both vertices. As one can see, the data points on the boundary of the cells – which exist due to Corollary 9 – move to the other cell.

Figure 6 illustrates two possible clusterings of 27 data points (colored dots) in \( \mathbb{R}^2 \). These clusterings were computed by running the \( k \)-means algorithm twenty times with three random sites in the beginning. In every iteration the \( k \)-means
algorithm computes a LSA to the current sites and updates the sites to be the arithmetic mean of their corresponding cluster. This is repeated until the clustering does not change anymore. Note that the $k$-means algorithm is deterministic, but its result strongly depends on the choice of the initial sites. Whereas the clustering on the left hand side (except for permutation of the colors) was the result in about half of the runs, the one on the right hand side was computed only once. Intuitively, one sees that the left clustering captures the structure of the data better than the one on the right hand side. This fits with our quantitative measure when looking at the normal cones of the respective vertices of the all-shape partition polytope and single-shape partition polytope:

Both volumes of the clustering on the left hand side are higher ($\mu_+ \approx 0.0076$, $\mu_- \approx 0.041$) than the ones of the clustering on the right hand side ($\mu_+ < 0.0001$, $\mu_- < 0.01$). The area of the surface of the 6-dimensional sphere is \( \frac{2\pi^3}{\Gamma(3)} \approx 31 \). The volumes of the respective normal cones were computed with MATLAB using the function Volume Computation of Convex Bodies by Ben Cousins with an error tolerance of 0.001 \[14\].

The higher volumes of the left clustering can also be verified by computing the edges of the respective normal cones and projecting these (normalized) site vectors to the $d$-dimensional components corresponding to the $k$ sites. The sites of a site vector in the convex hull of the edges (in the normal cone) are now located in the $k$ convex hulls of the sites of the edges in the $d$-dimensional space.

Figures 7 and 8 illustrate the three areas of the 2-dimensional sites for the respective clusterings w.r.t. $\mathcal{P}$ and $\mathcal{P}^-$, respectively. In order to be comparable, the 6-dimensional site vectors (edges of the normal cones) were normalized to Euclidean norm equal to 4. Of course, this does not change the induced clustering by the invariance under scaling of sites.

\[\text{Link: } https://www.mathworks.com/matlabcentral/fileexchange/43596-volume-computation-of-convex-bodies?focused=3878144&tab=function\]
As one can see the areas on the left hand side are larger than the ones on the right hand side, especially for the all-shape partition polytope. Note that these areas do not mean that one can choose three sites arbitrarily within the three areas and obtain an optimal site vector for the respective clustering. Instead, for any arbitrary site within one of these areas, there are sites in the respective other areas such that the corresponding site vector is inside the normal cone of the clustering. Further, these areas contain sites of representatives for all equivalence classes of site vectors. For the single-shape partition polytope these areas depict representatives of all site vectors for which the respective clusterings are optimal constrained LSAs, see Figure 8.

3.3. Stability of Site Vectors

Overview: For any vertex clustering $C$ with clustering vector $w(C)$, its normal cone $N_{\varphi^+}(w(C))$ contains all site vectors for which $C$ is an optimal constrained LSA and the normal cone $N_{\varphi^+}(w(C))$ contains all sites with the properties stated in Corollary 3. So far, we observed that clusterings with large normal cones are good in two senses: they are likely to be computed with randomly chosen sites and they have high volume, i.e. each site vector strictly inducing this clustering can be perturbed with the same (high) amount without changing the clustering. Given a vertex clustering $C$ and its normal cone, we now characterize a most stable site vector in the normal cone. After introducing our notion of stability, which depends on the choice of a $p$-norm, we provide an optimization problem whose optimal solution gives us a site vector with the highest possible stability for this clustering. Moreover, we present how the optimal solutions for different $p$-norms are connected and how one can obtain an approximate solution for any $p$-norm via use of the Euclidean norm.

Let us begin by defining a notion for the stability of site vectors.
Definition 10 (Stability of Site Vectors) Let \( p \in [1; \infty] \). For \( a \in B(N_{\| \cdot \| \pm} (w(C))) \subseteq S^d \cdot k \) the BHR stability of the site vector \( a \) w.r.t. \( p \) is
\[
\tau_p^0 (a) := \max \{ \delta > 0 \mid \bar{a} \in N_{\| \cdot \| \pm} (w(C)) \text{ for all } \| a - \bar{a} \|_p \leq \delta \}.
\]
For \( a \in B(N_{\| \cdot \| -} (w(C))) \subseteq S^d \cdot k \) we call
\[
\tau_p^0 (a) := \max \{ \delta > 0 \mid \bar{a} \in N_{\| \cdot \| -} (w(C)) \text{ for all } \| a - \bar{a} \|_p \leq \delta \}
\]
the LSA stability of the site vector \( a \) w.r.t. \( p \).

The stability measures \( \tau_p^0, \tau_p^\pm \) indicate the amount by which we can perturb the site vector \( a \) within the respective normal cone without changing the induced clustering. Note that we can extend the above definition to the equivalence classes of site vectors (and therefore to all site vectors) by inserting the corresponding representative into \( \tau_p^\star, \star \in \{ \pm, = \} \).

Geometrically, the above characterization of a ”most stable” site vector can be described by informally dropping a \( p \)-norm unit ball into the normal cone with \( 0 \in \mathbb{R}^d \cdot k \) as gravity center and compute where it “gets stuck” due to being blocked by the facets of the cone. The center of this unit ball then gives us a vector which lies ”most centrally” within the normal cone. Figure 9 illustrates two 2D examples of this approach.

The optimization problem below yields a stable site vector in the sense of Definition 10. We postpone the proofs of the theorems to Section 4.3.

Theorem 11 Let \( p \in [1; \infty] \), \( \star \in \{ \pm, = \} \) and \( w(C) \in \mathcal{P}^\star \) be a vertex with incident edges \( v_1, \ldots, v_t \). Then the optimal solution of the following optimization problem yields a site vector inducing the clustering \( C \) with highest possible stability w.r.t. \( p \).

\[
\min \| z \|_2^2 \quad \text{s.t.} \quad \left( \begin{array}{c}
\gamma_j^\star \leq v_j^T z \\
\gamma_j^\star \leq v_j^T z \quad \forall j \in \mathcal{I},
\end{array} \right)
\]

with \( \gamma_1^\star, \ldots, \gamma_t^\star \) being the optimal objective values of the problems

\[
\min \quad v_j^T z \quad \text{s.t.} \quad z \in \mathbb{B}_p^\star (0),
\]

for each \( j \in \mathcal{I} \). Here \( \mathbb{B}_p^\star (b) := \{ y \in \mathbb{R}^d \cdot k \mid \| y - b \|_p \leq \lambda \} \) denotes the closed \( p \)-norm ball with center \( b \in \mathbb{R}^d \cdot k \) and radius \( \lambda > 0 \).
The stability of an optimal solution \( z^{(p)} \) of (3) is \( \tau_p(z^{(p)}) = \frac{1}{\|z^{(p)}\|_p} \). Note that the edges \( v_j \) encode single (cyclical) movements for all \( j \in [r] \) (Theorem 8 and Corollary 9). Problem (3) is a quadratic optimization problem with linear constraints and the auxiliary problems (4) are linear optimization problems over a convex set. For \( p = 1, \infty \) the feasible regions are polytopes. Note, however, that there might be exponentially many edges and, thus, we might have to solve exponentially many auxiliary problems.

Problem (3) models the facets of the normal cone blocking the unit ball. We justify our approach that the ball is directly correspond to the stability of the depicted optimal sites. The BHR stability measures of the left clustering are \( \tau^\pm(\tilde{z}^{(2)}) \approx 0.253 \) and \( \tau^\mp(\tilde{z}^{(\infty)}) \approx 0.139 \), whereas for the right clustering we have \( \tau^\pm(\tilde{z}^{(2)}) \approx 0.018 \) and \( \tau^\mp(\tilde{z}^{(\infty)}) \approx 0.009 \).

For the LSA stability measures the differences between left and right clustering are not as large, since the single-shape partition polytope fixes the cluster sizes. We obtain \( \tau^\pm_2(\tilde{z}^{(2)}) \approx 0.365 \), \( \tau^\mp_\infty(\tilde{z}^{(\infty)}) \approx 0.163 \) (left) and \( \tau^\pm_2(\tilde{z}^{(2)}) \approx 0.199 \).
Figure 11: The optimal constrained LSA sites for $p = 2$ with perturbation of the first site.

Figure 12: The optimal sites for $p = 2$ w.r.t. $P$ and the areas of possible perturbations.

Figure 13: The optimal sites for $p = \infty$ w.r.t. $P$ and the areas of possible perturbations.
Figure 14: The optimal constrained LSA sites for $p = 2$ and the areas of possible perturbations.

Figure 15: The optimal constrained LSA sites for $p = \infty$ and the areas of possible perturbations.

$\tau_{\infty}(\mathbf{z}^{(\infty)}) \approx 0.100$ (right). One sees that the sites on the left hand side can be perturbed much more than the ones on the right hand side, which again justifies the intuition that the left clustering is better than the right one.

Next, we show how to obtain a feasible approximate solution of (3). It is well-known that for all $p, q \in [1; \infty]$ there is a positive constant $c_{p,q} > 0$ only depending on $p, q$ and the dimension of the space such that $\|x\|_p \leq c_{p,q} \|x\|_q$ for all $x \in \mathbb{R}^d$.

**Theorem 12** Let $p \in [1; \infty]$, $* \in \{\pm, =\}$ and $\mathbf{z}^{(p)} \in \mathbb{R}^{d \cdot k}$ be an optimal solution of (3). Then for all $q \in [1; \infty]$ the vector $\mathbf{z}' = c_{p,q} \mathbf{z}^{(p)}$ satisfies $\mathbb{B}_q(\mathbf{z}') \subseteq N_{p\ast}(w(C))$. In particular, $d(\mathbf{z}^{(p)}, \mathbf{z}') = 0$, i.e. both site vectors are in the same equivalence class. Moreover, the objective value of $\mathbf{z}'$ satisfies $\|\mathbf{z}'\|_2^2 \leq \max\{c_{p,q}, c_{q,p}\}^2 \|\mathbf{z}^{(q)}\|_2^2$ where $\mathbf{z}^{(q)} \in \mathbb{R}^{d \cdot k}$ is the optimal solution of (3) when considering the $q$-norm ball.

Note that an upper bound on the objective value of the approximate solution $\mathbf{z}'$ implies a lower bound on its stability $\tau_{\infty}(\mathbf{z}')$. If we choose $p = 2$, then the auxiliary problems (4) of Theorem 11 have optimal objective values $\gamma_j = -\|v_j\|_2$ for all $j \in [t]$. Problem (3) then reduces to

$$\begin{align*}
\min & \quad \|\mathbf{z}\|_2^2 \\
\text{s.t.} & \quad \mathbf{v}_j^T \mathbf{z} \leq -\|v_j\|_2 \quad \forall j \in [t], \\
& \quad \mathbf{z} \in \mathbb{R}^{d \cdot k}.
\end{align*}$$

Page 14
After having computed an optimal solution \( z^{(2)} \) of (5), we get a site vector \( z' = c_2, \rho^\star z^{(2)} \) for any arbitrary \( p \in [1; \infty] \) whose norm can be bounded from above by the norm of the most stable (w.r.t. \( p \)) site vector \( z(p) \) and a constant factor only depending on \( p \) (Theorem 12). Hence, we obtain a provable approximation.

4. Proofs

In this section, we provide the necessary proofs for Sections 3.1, 3.2 and 3.3. We will prove the theorems of Section 3.1 and 3.3 regarding equivalent statements for the bounded-shape and single-shape polytopes just for the bounded-shape case, since they work analogously: Essentially, just replace “±” by “=”.

4.1. Proofs for Section 3.1

Theorem 7 The measures \( \mu_\pm \) are well-defined. For \( * \in \{\pm, =\} \), \( C \) is a vertex clustering of \( \mathcal{D}^\star \) if and only if \( \mu_\star (C) > 0 \) and then

\[
\mu_\star (C) = \frac{\Gamma(d_k)}{2\pi^{d_k/2}} \int_{\text{bd}(B(N_{\rho^\star}(\mathcal{W}(C))))} d(z, a) \, da,
\]

with \( z \in \text{int}(N_{\rho^\star}(\mathcal{W}(C))) \cap S^{d_k} \) and \( \Gamma \) being the well-known Gamma function.

Proof. Let \( C \) be a clustering with normal cone \( N_{\rho^\star}(\mathcal{W}(C)) \). \( C \) is not a vertex clustering if and only if its normal cone has empty interior yielding \( \text{vol}(N_{\rho^\star}(\mathcal{W}(C))) = 0 \) and \( \mu_\pm = 0 \). So assume \( C \) is a vertex clustering.

For every \( a \in N_{\rho^\star}(\mathcal{W}(C)) \), its representative \( \frac{1}{\|a\|_2} a \in S^{d_k} \) is contained in the normal cone. Suppose we are given a representative of \( a \in \text{int}(N_{\rho^\star}(\mathcal{W}(C))) \cap S^{d_k} \). Then it is contained in the relative interior of the base of the normal cone. Therefore, starting at \( z \), we can walk into any direction along the unit sphere for at least a small distance while staying inside the interior of the normal cone. More precisely, for every geodesic \( \gamma \subseteq S^{d_k} \) starting at \( z \), there exists \( a \in \text{int}(N_{\rho^\star}(\mathcal{W}(C))) \cap S^{d_k} \cap \gamma \) with \( d(z, a) > 0 \).

Since the base of a cone is spherically convex, for all \( a \in \text{bd}(B(N_{\rho^\star}(\mathcal{W}(C)))) \) the geodesic \( \gamma \) “connecting” \( z \) and \( \frac{1}{\|a\|_2} a \), i.e. the \( \gamma \) satisfying \( L(\gamma) = d(z, a) \), is contained in the base of the normal cone. The area of \( B(N_{\rho^\star}(\mathcal{W}(C))) \), which is equal to the volume of \( N_{\rho^\star}(\mathcal{W}(C)) \), can be computed by the integral over all directions (geodesics through \( z \)) of \( d(z, a) \) with \( a \in \text{bd}(B(N_{\rho^\star}(\mathcal{W}(C)))) \) maximizing \( d(z, \cdot) \) along the direction of \( \gamma \). Note that this does not depend on the choice of \( z \) in the relative interior of the base and that this is well-defined due to the spherical convexity of the base. In particular, for every \( a \in \text{bd}(B(N_{\rho^\star}(\mathcal{W}(C)))) \subseteq \text{bd}(N_{\rho^\star}(\mathcal{W}(C))) \), there is a geodesic \( \gamma \) satisfying \( L(\gamma) = d(z, a) > 0 \), and for every geodesic \( \gamma \) starting at \( z \) we have

\[
\max\{d(z, a) \mid a \in \gamma, a \in \text{bd}(N_{\rho^\star}(\mathcal{W}(C))))\} = d(z, \bar{a}) > 0
\]

with \( \{\bar{a}\} = \text{bd}(B(N_{\rho^\star}(\mathcal{W}(C)))) \cap \gamma \). We obtain

\[
\text{vol}(N_{\rho^\star}(\mathcal{W}(C))) = \int_{\text{bd}(B(N_{\rho^\star}(\mathcal{W}(C))))} d(z, a) \, da,
\]

which, together with the fact that \( \text{vol}(R^{d_k}) \) is equal to the area of the surface of \( S^{d_k} \), yields representation (6).

Let \( K \subseteq R^{d_k} \) be a cone with equal volume, i.e. \( \text{vol}(K) = \text{vol}(N_{\rho^\star}(\mathcal{W}(C))) \), and consider a diffeomorphism \( f : B(N_{\rho^\star}(\mathcal{W}(C))) \to B(K) \). Then \( f(z) \) satisfies \( d(f(z), a) > 0 \) for all \( a \in \text{bd}(B(K)) \), because open sets (in \( B(N_{\rho^\star}(\mathcal{W}(C))) \)) are mapped to open sets (in \( B(K) \)) and vice versa. Furthermore, as \( \text{vol}(K) = \text{vol}(N_{\rho^\star}(\mathcal{W}(C))) \) and the bases \( B(K) \) and \( B(N_{\rho^\star}(\mathcal{W}(C))) \) are spherically convex, we can perturb \( f(z) \) and \( z \) in equal measure without leaving the bases of the respective cones. This shows that \( \mu_\pm \) as a quality measure is well-defined.

\[\Box\]
The statements and proofs are based on the dissertation [11] and master’s thesis [23] of the authors. First, we need a few auxiliary lemmas.

Fix $C, C'$ to be the clusterings of Theorem 8. Let $CDG(C, C')$ be their clustering difference graph and let $\mathcal{M}$ be the set of movements corresponding to a decomposition of $CDG(C, C')$ into cycles and paths. Denote by $\mathcal{M}_{\geq 2} \subseteq \mathcal{M}$ the set of movements corresponding to cycles and non-trivial paths, i.e. paths (and cycles) containing at least two edges.

**Lemma 13** Let $a \in \mathbb{R}^{d \times k}$ such that $a^T w(C) = a^T w(C') > a^T w$ for all other vertices $w \in \mathcal{P} \setminus \{w(C), w(C')\}$. Then $a^T w(M) = 0$ for all $M \in \mathcal{M}$.

**Proof.** Let $M \in \mathcal{M}$ and suppose $a^T w(M) > 0$. Applying $M$ to $C$ yields a (feasible) clustering $\tilde{C}$ with $a^T w(\tilde{C}) = a^T (w(C) + w(M)) > a^T w(C)$ which is a contradiction. For $a^T w(M) < 0$ we obtain the same contradiction by applying $M^{-1}$ to $C'$, as we then would derive a (feasible) clustering $\tilde{C}$ with

$$a^T w(\tilde{C}) = a^T (w(C') - w(M)) = a^T w(C') - a^T w(M) > a^T w(C').$$

So $a^T w(M) = 0$, yielding the claim. \hfill \Box

This implies that all movements corresponding to $CDG(C, C')$ operate on the same clusters.

**Lemma 14** There exists an index subset $I := \{i_1, \ldots, i_j\} \subseteq [k]$ such that all (cyclical) movements are of the form $i_1 \rightarrow i_2 \rightarrow \cdots \rightarrow i_j \rightarrow i_1$.

**Proof.** By Lemma 13 the vectors of all movements in $\mathcal{M}$ are collinear. $X$ contains non-zero distinct data points by assumption. So, if two movements $M_1$ and $M_2$ would move items between different subsets of clusters, their corresponding vectors would leave different components of the clustering vectors unchanged and therefore cannot be collinear.

Let $M_1 : C_{i_1} \rightarrow C_{i_2} \rightarrow \cdots \rightarrow C_{i_j}$ be a movement in $\mathcal{M}$ for some $2 \leq t \leq k$. Note that for $t \geq 3$, $i_1 = i_j$ might be possible in which case $M_1$ is a cyclical movement. Suppose there exists a movement $M_2$ whose corresponding path in the CDG contains an edge $(i_l, i_j)$ with $j \neq l + 1$ (cyclical indexing). Then the corresponding paths in $CDG(C, C')$ would be as illustrated in Figure 16. Figure 16a shows the path corresponding to $M_1$.

For $l + 1 < j$, $M_2$ “skips” all clusters between $i_l$ and $i_j$, see Figure 16b. For $l > j$ we get a cyclical movement “skipping” all clusters $C_{i_l}$ with $r < j$ or $r > l$, c.f. Figure 16c. In both cases, we find a (cyclical) movement that leaves different components of the clustering vectors unchanged: $w(M_2)$ does not change cluster $i_{l+1}$, whereas $w(M_1)$ does. Therefore, the respective vectors cannot be collinear to the one of $M_1$. This is a contradiction to Lemma 13. \hfill \Box

**Lemma 15** If no three points in $X$ lie on a single line, then $|\mathcal{M}_{\geq 2}| \leq 1$.  

\hfill Page 16
contain at least one cycle. So suppose which are collinear. But this means that  

We obtain for  

and  

By Lemma 14, all movements in  which contains both paths, c.f. Figure 17a. We can construct two movements  and  of edge-length greater than 1. We are now ready to prove Theorem 8 and Corollary 9.

In Figure 17b, we see that  equals  for  between the nodes  and  and between the  and the  clusters  equals  and  equals  After having applied the movements  and  to  we obtain the same clustering as after applying  and  so we could replace the paths corresponding to  and  by the ones corresponding to  and in the decomposition of  and . By Lemma 13 the vectors  and  and  are collinear. Let us now consider the components of  corresponding to the changes of  . We obtain for  and  (in this order) 

which are collinear. But this means that  lie on a single line, which is a contradiction to our assumption on  . Hence,  

In particular, the CDG of vertex clusterings whose clustering vectors are adjacent contains at most one cycle or path of edge-length greater than 1. We are now ready to prove Theorem 8 and Corollary 9.

Theorem 8 Let  be  be two clusterings such that  and  are adjacent vertices of  . Let further no three points in  lie on a single line. Then  and  differ by a single (cyclical) movement or by two movements and there are distinct  such that both are of the form  

Proof. If  then the statement follows from Lemma 15 and the fact that  implies that  must contain at least one cycle. So suppose  . If  contains a path  then  with  then, by Lemma 14 and Lemma 15 there cannot be any other paths or cycles. Now let  and suppose there are three movements  and  which all have to be of the form  By collinearity of the vectors  and  the respective components corresponding to the cluster  are collinear. By definition, these entries equal the labels of the three edges – which have to be collinear. This, however, is a contradiction to the assumption that no three points in  lie on a single line. 

Figure 17: Construction of  and  in CDG(C, C').
Corollary 9 Let $C := (C_1, \ldots, C_k)$, $C' := (C'_1, \ldots, C'_k)$ be two clusterings such that $w(C)$ and $w(C')$ are adjacent vertices of $\mathcal{P}$. If no four points in $X$ lie on a single line, then $C$ and $C'$ only differ by a single cyclical movement.

PROOF. Note that $|C| = |C'|$ implies that $\mathcal{M} = \mathcal{M}_{\geq 2}$, since the CDG decomposes into cycles. Suppose there exist two cycles in $CDG(C, C')$ with corresponding cyclical movements $M_1$ and $M_2$ as follows

$$
M_1 : \quad C_{i_1} \rightarrow C_{i_2} \rightarrow \ldots \rightarrow C_{i_t} \rightarrow C_{i_{t+1}},
$$

$$
M_2 : \quad C_{i'_1} \rightarrow C_{i'_2} \rightarrow \ldots \rightarrow C_{i'_t} \rightarrow C_{i'_{t+1}},
$$

for some $t \geq 2$. Again, we have $x_{i_l} \neq x_{i'_l}$ for all $l, r \in [t]$. Similarly to the proof of Lemma 15, we can construct two different cycles $M'_1$ and $M'_2$ via $M'_1 = M_j$ between $i_1$ and $i_t$ and $M'_2 = M_2$ and $M'_2 = M_1$ between clusters $C_{i_1}$ and $C_{i_t}$. The corresponding vectors $w(M'_1), w(M'_2), w(M_1')$ and $w(M_2')$ are collinear and, looking at the $d$ entries corresponding to $x_{i_1}$, we obtain (in this order)

$$
x_{i_l} - x_{i_1}, \quad x'_{i_l} - x'_{i_1}, \quad x_{i_t} - x_{i_1}, \quad x'_{i_t} - x'_{i_1},
$$

which are collinear. However, this implies that $x_{i_1}, x'_{i_1}, x_{i_t}$ and $x'_{i_t}$ lie on a line, contradicting our assumption. \qed

4.3. Proofs for Section 3.3

Theorem 11 Let $p \in [1; \infty], \ast \in \{\pm, =\}$ and $w(C) \in \mathcal{P}^*$ be a vertex with incident edges $v_1, \ldots, v_t$. Then the optimal solution of the following optimization problem yields a site vector inducing the clustering $C$ with highest possible stability w.r.t. $p$.

$$
\begin{align*}
\min_{z} & \quad \|z\|^2_p \\
\text{s.t.} & \quad v_j^T z \leq \gamma_j^p \quad \forall j \in [t], \\
& \quad z \in \mathbb{R}^{d_k}
\end{align*}
$$

with $\gamma_1^p, \ldots, \gamma_t^p$ being the optimal objective values of the problems

$$
\begin{align*}
\min_{z} & \quad v_j^T z \\
\text{s.t.} & \quad z \in \mathbb{P}_p^\ast(0), \\
& \quad z \in \mathbb{R}^{d_k}
\end{align*}
$$

for each $j \in [t]$. Here $\mathbb{P}_p^\ast(b) := \{y \in \mathbb{R}^{d_k} \mid \|y - b\|^p_p \leq \lambda\}$ denotes the closed $p$-norm ball with center $b \in \mathbb{R}^{d_k}$ and radius $\lambda > 0$.

PROOF. Let $p \in [1; \infty], w(C)$ be a vertex clustering with normal cone $N_{p, \ast}(w(C))$ and let $F_1, \ldots, F_t$ be the facets of the normal cone. The “gravity center” at $0 \in \mathbb{R}^{d_k}$ which attracts the $p$-norm ball can be modeled by minimizing the squared Euclidean norm $\|z\|^2_p$. Together with the constraints that the $p$-norm unit ball with center $z$ has to be inside the normal cone $N_{p, \ast}(w(C))$, this yields the optimization problem

$$
\begin{align*}
\min_{z} & \quad \|z\|^2_p \\
\text{s.t.} & \quad z \in N_{p, \ast}(w(C)), \\
& \quad \text{dist}(z, F_j)_p \geq 1 \quad \forall j \in [t], \\
& \quad z \in \mathbb{R}^{d_k}
\end{align*}
$$

Here, $\text{dist}(z, F_j)_p$ denotes the distance w.r.t. the $p$-norm of $z$ and the facet $F_j$, i.e. $\text{dist}(z, F_j)_p = \inf\{\|z - f\|_p \mid f \in F_j\}$. First we show that an optimal solution $z^{(p)}$ of (9) has the highest stability among all site vectors in the normal cone.
The stability of \( z^{(p)} \) is \( \tau^p_z(z^{(p)}) = \frac{1}{\|z^{(p)}\|_2} > 0 \) by construction. Note that any site vector on the boundary of the normal cone has stability 0. So let \( z \in \text{int}(N_{\gamma^p} (w(C))) \) and let \( \delta_i := \min\{\text{dist}(z,F_j) | j \in [t]\} > 0 \) be the minimal \( p \)-norm distance to the boundary of the normal cone. Then \( z \) has stability \( \tau^p_z(z) = \frac{\delta_i}{\|z\|_2} \). Moreover, the vector \( \frac{1}{\delta_i} z \) is feasible for \( \mathbf{9} \) and, by optimality of \( z \), we obtain

\[
\tau^p_z(z) = \frac{\delta_i}{\|z\|_2} = \frac{1}{\|\frac{1}{\delta_i} z\|_2} \leq \frac{1}{\|z^{(p)}\|_2} = \tau^p_z(z^{(p)}).
\]

We will now prove that \( \mathbf{9} \) is equivalent to \( \mathbf{7} \). Let us start by motivating the choices of \( \gamma^{\ell}_i, \ldots, \gamma^{\ell}_F \). The “distance constraints” \( \mathbf{9} \) define two half spaces for each \( j \in [t] \). Technically speaking, we have to replace \( F_j \) by \( \text{lin}(F_j) \), however, this makes no difference, since we only consider vectors inside the normal cone. The optimal objective value of \( \mathbf{9} \), i.e. \( \gamma^{\ell}_j := v_j^T z_j^* < 0 \) with \( z_j^* \) being an optimal solution of \( \mathbf{8} \), then satisfies

\[
dist(z, \text{lin}(F_j))_p \geq 1 \iff z \in H_{\gamma^{\ell}_j, \gamma^{\ell}_j} \cup H_{\gamma^{\ell}_j, -\gamma^{\ell}_j}.
\]

Furthermore, for all \( j \in [t] \) we have \( F_j = \text{lin}(F_j) \cap N_{\gamma^p} (w(C)) \) and \( \text{lin}(F_j) = \{v_j\}_F \) is a \((d \cdot k - 1)\)-dimensional linear subspace, i.e. a hyperplane through the origin with normal vector \( v_j \). In particular, we can write any \( z \in \mathbb{R}^{d \cdot k} \) as linear combination of an orthogonal basis \( \{v_j, f_1, \ldots, f_{d \cdot k - 1}\} \) of \( \mathbb{R}^{d \cdot k} \) with \( f_j \in \text{lin}(F_j) \) for all \( l \in [d \cdot k - 1] \), i.e.

\[
z = v_0^j v_j + \sum_{l=1}^{d \cdot k - 1} v_l^j f_l \quad \text{with } v_0^j, \ldots, v_{d \cdot k - 1}^j \in \mathbb{R}.
\]

For \( z \in N_{\gamma^p} (w(C)) \), we observe that \( v_0^j \leq 0 \), since \( v_j \) points along the edge direction and, thus, away from the normal cone. If \( z \) is in the interior of \( N_{\gamma^p} (w(C)) \), we even obtain \( v_0^j < 0 \), because otherwise, if \( v_0^j = 0 \), then \( z \in \text{lin}(F_j) \subseteq \text{bd}(N_{\gamma^p} (w(C))) \). Together with \( \mathbf{10} \), we obtain

\[
N_{\gamma^p} (w(C)) \cap \{z \in \mathbb{R}^{d \cdot k} \mid \text{dist}(z,F_j)_p \geq 1\} = N_{\gamma^p} \cap H_{\gamma^{\ell}_j, \gamma^{\ell}_j}^\leq\}
\]

for all \( j \in [t] \). Rewriting the normal cone in its representation with the facet-defining supporting half spaces

\[
N_{\gamma^p} (w(C)) = \bigcap_{j=1}^t H_{\gamma^{\ell}_j, \gamma^{\ell}_j}^\leq
\]

and recalling that \( \gamma^{\ell}_j \leq 0 \), we see that the half space \( H_{\gamma^{\ell}_j, \gamma^{\ell}_j}^\leq \) is strictly contained in \( H_{\gamma^{\ell}_j, \gamma^{\ell}_j}^\leq \) for all \( j \in [t] \). Thus,

\[
\bigcap_{j=1}^t H_{\gamma^{\ell}_j, \gamma^{\ell}_j}^\leq \subseteq \bigcap_{j=1}^t H_{\gamma^{\ell}_j, \gamma^{\ell}_j}^\leq = N_{\gamma^p} (w(C)).
\]

Therefore, the constraint \( z \in N_{\gamma^p} (w(C)) \) is redundant and \( \mathbf{9} \) is equivalent to \( \mathbf{7} \). \( \square \)

In order to see why lower dimensional faces blocking the \( p \)-norm ball in Section 3.3 do not guarantee a stable site vector, consider the following 3-dimensional example for \( p = 2 \).

Let \( z = (0,0,2)^T, a_1 = (0, \sqrt{3}, 3)^T, a_2 = (0, -\sqrt{3}, 3)^T, a_3 = (\sqrt{3}, 0, 3)^T \in \mathbb{R}^3 \) and consider the cone spanned by \( \{a_1, a_2, a_3\} \). It is easy to see that \( a_1, a_2 \) and \( a_3 \) are linearly independent. The vector \( z \) can be written as

\[
z = \frac{1}{3} a_1 + \frac{1}{3} a_2 + 0 \cdot a_3 \quad \Rightarrow \quad z \in \text{bd}(\text{pos}(\{a_1, a_2, a_3\}))
\]

One can easily verify that \( \text{dist}(z, \text{pos}(\{a_1\}))_2 = \|z - \frac{1}{2} a_1\|_2 = 1 \) for all \( j \in [3] \), see Figure 18 for a sketch. The given \( z = (0,0,2)^T \) is indeed optimal, since we cannot decrease its norm without getting infeasible, i.e. being outside of the cone.
Thus, suffices to show that $z$ is an optimal solution of (7). For all $q \in [1;\infty]$ the vector $z = c_{p.q}z^{(p)}$ satisfies $\mathbb{B}^p_q(z) \subseteq N_{\gamma_p}(w(C))$. In particular, $d(z^{(p)}, z') = 0$, i.e. both site vectors are in the same equivalence class. Moreover, the objective value of $z$ satisfies $\|z\|_2^2 \leq \max\{c_{p.q}, c_{q.p}\}^2 \|z^{(q)}\|_2^2$ where $z^{(q)} \in \mathbb{R}^{d.k}$ is the optimal solution of (7) when considering the $q$-norm ball.

**Theorem 12** Let $p \in [1;\infty]$, $\star \in \{\pm, =\}$ and $z^{(p)} \in \mathbb{R}^{d.k}$ be an optimal solution of (7). Then for all $q \in [1;\infty]$ the vector $z' = c_{p.q}z^{(p)}$ satisfies $\mathbb{B}^p_q(z') \subseteq N_{\gamma_p}(w(C))$. For the objective value of $z'$ satisfies $\|z\|_2^2 \leq \max\{c_{p.q}, c_{q.p}\}^2 \|z^{(q)}\|_2^2$ where $z^{(q)} \in \mathbb{R}^{d.k}$ is the optimal solution of (7) when considering the $q$-norm ball.

**Proof.** Let $q \in [1;\infty]$ with $q \neq p$, else the claim is trivial, because $c_{p.q} = 1$. By assumption $\mathbb{B}^p_q(z^{(p)}) \subseteq N_{\gamma_p}(w(C))$, thus, $\mathbb{B}^p_q(\lambda z^{(p)}) \subseteq N_{\gamma_p}(w(C))$ for all $\lambda > 0$. The vector $z' = c_{p.q}z^{(p)}$ already satisfies $d(z^{(p)}, z') = 0$. It suffices to show that $\lambda = c_{p.q}$ satisfies $\mathbb{B}^p_q(0) \subseteq \mathbb{B}^p_q(0)$, whence $z' = \lambda z^{(p)} = c_{p.q}z^{(p)}$ has the desired properties. Let $y \in \mathbb{B}^p_q(0)$, then we obtain

$$1 \geq \|y\|_q \geq \frac{1}{c_{p.q}} \|y\|_p \quad \Rightarrow \quad \|y\|_p \leq c_{p.q}.$$ 

Hence, choosing $\lambda = c_{p.q} > 0$ yields $\mathbb{B}^p_q(c_{p.q}z^{(p)}) \subseteq \mathbb{B}^p_q(c_{p.q}z^{(p)}) \subseteq N_{\gamma_p}(w(C))$.

Now let $z^{(q)}$ be the optimal solution of (7) for the $q$-norm distance, i.e. with constraints $\mathbb{B}^p_q(z^{(q)})_q \geq 1$ for all $q \in [1;\infty]$. In the following, whenever we refer to (7), we mean the optimization problem for the $p$-norm distance. Denote by $\gamma_1^q, \ldots, \gamma_d^q$ the optimal objective values of (7) with $\mathbb{B}^p_q(0)$ as feasible region and by $\gamma_1^p, \ldots, \gamma_d^p$ for $\mathbb{B}^p_q(0)$, respectively. We distinguish between two cases: $p < q$ and $p > q$.

For $p < q$, we have $\mathbb{B}^p_q(0) \subseteq \mathbb{B}^p_q(0)$ and, therefore, $z^{(q)}$ satisfies

$$\gamma_j^p(z^{(q)}) \leq \gamma_j^q \leq \gamma_j^p$$

for all $j \in [1;\infty]$. Hence, $z^{(q)}$ is feasible for (7) and $\mathbb{B}^p_q(z^{(q)}) \subseteq N_{\gamma_p}(w(C))$. For the objective value of $z' = c_{p.q}z^{(p)}$, we obtain

$$\|z\|_2^2 = c_{p.q}^2 \|z^{(p)}\|_2^2 \leq c_{p.q}^2 \|z^{(q)}\|_2^2,$$

because $z^{(p)}$ is an optimal solution of (7).

For the other case $p > q$, consider $\widetilde{z} = c_{p.q}z^{(q)}$ which satisfies $\mathbb{B}^p_q(\tilde{z}) \subseteq N_{\gamma_p}(w(C))$ due to the first part of this theorem. Note that $z^{(q)}$ is not feasible for (7). Hence, by optimality of $z^{(p)}$,

$$\|z\|_2^2 = \|z^{(p)}\|_2^2 \leq \|\tilde{z}\|_2^2 = c_{q.p}^2 \|z^{(q)}\|_2^2.$$
In total, since either $c_{p,q}$ or $c_{p,q}$ equals 1, (11) and (12) yield
\[ \|z'(q)\|_2^2 \leq \max\{c_{p,q}, c_{q,p}\}^2 \|z'(q)\|_2^2. \]

\[ \square \]

5. Outlook

We would like to stress that all computations of the volume of the normal cone, the computations of edges of the normal cone, as well as the optimal solutions of (3) are challenging due to the possibly exponential number of edges incident to a given vertex of the polytope. However, this significant computational effort is rewarded with deep insights into the structure and behavior of site vectors. Our contribution can be seen as a step towards a strong understanding of the relation of site vectors and the clusterings they induce. The next natural steps include the design of approximation algorithms that trade small errors in the construction of normal cones for a significant reduction in computation times.

References

[1] D. Aloise, A. Deshpande, P. Hansen, and P. Popat. NP-hardness of Euclidean sum-of-squares clustering. Machine Learning, 75:245–249, 2009.

[2] E. Anderes, S. Borgwardt, and J. Miller. Discrete Wasserstein Barycenters: Optimal Transport for Discrete Data. Mathematical Methods of Operations Research, 84(2):389–409, 2016.

[3] F. Aurenhammer. Power Diagrams: Properties, Algorithms and Applications. SIAM Journal on Computing, 16(1):78–96, 1987. doi: 10.1137/0216006. URL http://dx.doi.org/10.1137/0216006.

[4] F. Aurenhammer, F. Hoffmann, and B. Aronov. Minkowski-Type Theorems and Least-Squares Clustering. Algorithmica, 20(1):61–76, 1998. ISSN 1432-0541. doi: 10.1007/PL00009187. URL http://dx.doi.org/10.1007/PL00009187.

[5] Sharon Aviran, Nissan Lev-Tov, Shmuel Onn, and Uriel G. Rothblum. Vertex Characterization of Partition Polytopes of Bipartitions and of Planar Point Sets. Discrete Applied Mathematics, 124(1):1 – 15, 2002. ISSN 0166-218X. doi: http://dx.doi.org/10.1016/S0166-218X(01)00326-2. URL http://www.sciencedirect.com/science/article/pii/S0166218X01003262.

[6] M. Balinski and A. Russakoff. On the Assignment Polytope. SIAM Review, 16(4), 1974.

[7] E. R. Barnes, A. J. Hoffman, and U. G. Rothblum. Optimal Partitions Having Disjoint Convex and Conic Hulls. Mathematical Programming, 54(1):69–86, 1992. ISSN 1436-4646. doi: 10.1007/BF01586042. URL http://dx.doi.org/10.1007/BF01586042.

[8] K. P. Bennett and O. L. Mangasarian. Multicategory discrimination via linear programming. Optimization Methods and Software, 3:27–39, 1992.

[9] P. Berkhin. Survey of clustering data mining techniques. Technical report, Accrue Software, 2002.

[10] N. Bonifas, M. Di Summa, E. Eisenbrand, N. Hähle, and M. Niemeier. On Sub-determinants and the Diameter of Polyhedra. In Computational Geometry, pages 1–15, 2009. ISSN 1432-0444. doi: 10.1007/s00454-014-9601-x. URL http://dx.doi.org/10.1007/s00454-014-9601-x.

[11] S. Borgwardt. A Combinatorial Optimization Approach to Constrained Clustering. Dissertation, Technische Universität München, München, 2010.

[12] S. Borgwardt. On Soft Power Diagrams. Journal of Mathematical Modelling and Algorithms in Operations Research, 14(2):173–196, 2015. doi: 10.1007/s10852-014-9263-y. URL http://dx.doi.org/10.1007/s10852-014-9263-y.

[13] A. Brieden and P. Gritzmann. On Optimal Weighted Balanced Clusterings: Gravity Bodies and Power Diagrams. SIAM Journal on Discrete Mathematics, 26(2):415–434, 2012. doi: 10.1137/110832707. URL http://dx.doi.org/10.1137/110832707.

[14] B. Cousins. Volume Computation of Convex Bodies, June 2015. MATLAB Function.

[15] J. A. De Loera and E. D. Kim. Combinatorics and Geometry of Transportation Polytopes: An Update. Discrete & Computational Geometry, 52(1):102–115, 2014. ISSN 1432-0444. doi: 10.1007/s00454-014-9601-x. URL http://dx.doi.org/10.1007/s00454-014-9601-x.

[16] J.-P. Doignon and M. Regenwetter. An Approval-voting Polytope for Linear Orders. Journal of Mathematical Psychology, 41(2):171–188, July 1997. ISSN 0022-2496. doi: 10.1006/jmps.1997.1155. URL http://dx.doi.org/10.1006/jmps.1997.1155.

[17] K. Fukuda, S. Onn, and V. Rosta. An Adaptive Algorithm for Vector Partitioning. Journal of Global Optimization, 25(3):305–319, 2003. ISSN 1573-2916. doi: 10.1023/A:1022417803474. URL http://dx.doi.org/10.1023/A:1022417803474.

[18] B. Gao, F. K. Hwang, W. C. W. Li, and U. G. Rothblum. Partition Polytopes Over 1-dimensional Points. Mathematical Programming, 85(2):335–362, 1999. ISSN 1436-4646. doi: 10.1007/s101070050080. URL http://dx.doi.org/10.1007/s101070050080.

[19] J. Gill and S. Linusson. The k-assignment Polytope. In Discrete Optimization, volume 6(2), pages 148–161. Elsevier Science, 2009.

[20] E. S. Gottlieb and M. R. Rao. The Generalized Assignment Problem: Valid Inequalities and Facets. Mathematical Programming, 46(1):31–52, 1990. ISSN 1436-4646. doi: 10.1007/BF01585725. URL http://dx.doi.org/10.1007/BF01585725.

[21] M. Grötschel and Y. Wakabayashi. Facets of the Clique Partitioning Polytope. Mathematical Programming, 47(1):367–387, 1990. ISSN 1436-4646. doi: 10.1007/BF01580870. URL http://dx.doi.org/10.1007/BF01580870.

[22] R. Guralnick and D. Perkinson. Permutation Polytopes and Indecomposable Elements in Permutation Groups. Journal of Combinatorial Theory, Ser. A, 113:1243–1256, 2006.

[23] F. Happach. Stable Clusterings and the Cones of Outer Normals. Master’s thesis, Technische Universität München, München, 2016.
[24] F. K. Hwang, S. Onn, and U. G. Rothblum. Representations and Characterizations of Vertices of Bounded-shape Partition Polytopes. *Linear Algebra and its Applications*, 278(1):263 – 284, 1998. ISSN 0024-3795. doi: http://dx.doi.org/10.1016/S0024-3795(97)10092-1. URL [http://www.sciencedirect.com/science/article/pii/S0024379597100921](http://www.sciencedirect.com/science/article/pii/S0024379597100921).

[25] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering - a review. In *ACM Computing Surveys*, volume 31-3, pages 264–323, 1999.

[26] A. Kalman. *Newton Polytopes of Cluster Variables*. Dissertation, University of California, Berkeley, Berkeley, 2014.

[27] S. P. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

[28] J. B. MacQueen. Some methods of classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

[29] M. Mahajan, P. Nimbhorkar, and K. Varadarajan. The planar k-means problem is NP-hard. *Theoretical Computer Science*, 442:13–21, 2012.

[30] G. L. Nemhauser and L. A. Wolsey. *Integer and Combinatorial Optimization*. Wiley, New York, NY, 1999.

[31] S. Onn. Geometry, Complexity, and Combinatorics of Permutation Polytopes. *Journal on Combinatorial Theory, Ser. A*, 64(1):31–49, 1993.

[32] A. Schrijver. *Theory of Linear and Integer Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1986. ISBN 0-471-90854-1.

[33] R. Suck. Geometric and Combinatorial Properties of the Polytope of Binary Choice Probabilities. *Mathematical Social Sciences*, 23(1):81 – 102, 1992. ISSN 0165-4896. doi: http://dx.doi.org/10.1016/0165-4896(92)90039-8. URL [http://www.sciencedirect.com/science/article/pii/0165489692900398](http://www.sciencedirect.com/science/article/pii/0165489692900398).

[34] R. Suck. Polytopes in Measurement and Data Analysis. Review of Lectures on Polytopes, by Günther M. Ziegler. *Journal of Mathematical Psychology*, 41(3):299–303, 1997. doi: 10.1006/jmps.1997.1173.

[35] U. v. Luxburg. Clustering Stability: An Overview. *Foundations and Trends in Machine Learning*, 2(3):235–274, 2010. ISSN 1935-8237. doi: 10.1561/2200000008.

[36] A. Vattani. k-Means requires exponentially many iterations even in the plane. *Discrete Computational Geometry*, 45:596–616, 2011.

[37] R. Xu and D. Wunsch. Survey of clustering algorithms. In *IEEE Transactions on Neural Networks*, volume 16:3, pages 645–678, 2005.

[38] G. M. Ziegler. *Lectures on Polytopes*, volume 152 of *Graduate Texts in Mathematics*. Springer Verlag, 1995. URL [http://www.springer.com/math/geometry/book/978-0-387-94365-7](http://www.springer.com/math/geometry/book/978-0-387-94365-7).