Chapter 4

Multilevel Representative Clustering

In this chapter we will extend the basic reduction algorithm (see section 2.3) to a multilevel approach. The main idea is to iterate the decomposition based representative clustering method until the decomposition is fine enough so that the optimal solution of the reduced cluster problem determines an optimal clustering of the original cluster problem.

We will present a general approach that can be always used if the number of clusters k is known a priori. For special homogeneity functions we will additionally describe a powerful extension based on *Perron Cluster* analysis that can be used for cluster problems, where the number of clusters is unknown.

4.1 General approach

Let $V = \{v_1, \ldots, v_n\} \subset \Omega$ be any data set in Ω with frequency function f and homogeneity function h. A point very critical within the application of the basic reduction algorithm (see section 2.3) is the fulfillment of the condition that the decomposition of V has to be a covering of an optimal k-cluster set of (V, f, h).

Suppose now that we have any decomposition Θ of V with codebook W and any optimal k-cluster set $\mathcal{C}$ of $(W, \check{f}, \check{h}_f)$. We know that the extension $\hat{\mathcal{C}}$ of $\mathcal{C}$ on V is a k-cluster set of (V, f, h). Since Θ is a covering of $\mathcal{C}$ by construction, it is also a covering of $\hat{\mathcal{C}}$. Therefore we have $\Gamma_{\check{f},\check{h}_f}(\mathcal{C}) = \Gamma_{f,h}(\hat{\mathcal{C}})$. At the moment we cannot be sure that Θ is also a covering of an optimal k-cluster set of (V, f, h), what would imply that $\hat{\mathcal{C}}$ is optimal. Therefore we try to refine Θ.

Let $\Theta`$ with codebook vector $W`$ be the result of a suitable refinement process, e.g., as it will be described in the next section. If we now compute an optimal k-cluster set $\mathcal{C}`$ of $(W`, \check{f}, \check{h}_f)$, we can extend it to $\hat{\mathcal{C}}`$ and compute the weighted intra-cluster homogeneity $\Gamma_{f,h}(\hat{\mathcal{C}}`) = \Gamma_{\check{f},\check{h}_f}(\mathcal{C}`)$. If $\Gamma_{f,h}(\hat{\mathcal{C}}`) > \Gamma_{f,h}(\hat{\mathcal{C}})$, the new clustering is better, i.e. Θ was definitely not a covering of an optimal k-cluster set of (V, f, h). With Θ‘ we have a new candidate that is a finer decomposition as Θ and that is a covering of a k-cluster set with improved quality.

From the above reflections one easily derives the main idea of the new multilevel representative clustering approach: Refine iteratively the decomposition Θ, until no further improvement of the corresponding representative clustering is observable.

Multilevel Reduction Algorithm

The following algorithm embeds the basic reduction algorithm in a multilevel refinement process. Note that k has to be known a priori.
(1) Compute a decomposition Θ (based on a codebook W) with adaptive choice of n_k, $k \leq n_k \leq \Bbbk \ll n$.
(2) Compute an optimal k-cluster set $\mathcal{C}$ of $(W, \check{f}, \check{h}_f)$.
(3) Extend $\mathcal{C}$ on V: $\hat{\mathcal{C}}$ is a k-cluster set of (V, f, h). Since Θ is a covering of $\hat{\mathcal{C}}$, we have $\Gamma_{\check{f},\check{h}_f}(\mathcal{C}) = \Gamma_{f,h}(\hat{\mathcal{C}})$.
(4) Refine Θ so that the new decomposition Θ‘ of V with codebook W' is also a covering of $\hat{\mathcal{C}}$.
(5) Compute an optimal k-cluster set $\mathcal{C}$‘ of $(W$‘$, \check{f}, \check{h}_f)$.
(6) Extend $\mathcal{C}$‘ on V: $\hat{\mathcal{C}}$‘ is k-cluster set of (V, f, h). Since Θ‘ is a covering of $\hat{\mathcal{C}}$‘, we have $\Gamma_{\check{f},\check{h}_f}(\mathcal{C}$‘$) = \Gamma_{f,h}(\hat{\mathcal{C}}$‘$)$.
(7) If $\Gamma_{\check{f},\check{h}_f}(\mathcal{C}$‘$) > \Gamma_{\check{f},\check{h}_f}(\mathcal{C})$ then set $\Theta := \Theta$‘ and go to step (4), else stop.

For the computation of Θ with adaptive choice of n_k and the codebook W we can use the algorithms described in chapter 3. In the following section we will describe techniques for a refinement of an existing decomposition so that the quality of the corresponding codebook clustering increases.

4.2 Adaptive decomposition refinement

Let $\widetilde{\mathcal{C}}$ be any optimal k-cluster set of (V, f, h) and $\hat{\mathcal{C}}$ be any nearly optimal k-cluster set of (V, f, h). Further let $\Theta := \{\Theta_1, \ldots, \Theta_{n_k}\}$ any decomposition of V with codebook $W = \{w_1, \ldots, w_{n_k}\}$ so that Θ is a covering of $\hat{\mathcal{C}}$, but not of $\widetilde{\mathcal{C}}$. Then there exist clusters $\widetilde{C}_1, \widetilde{C}_2 \in \widetilde{\mathcal{C}}$, $\hat{C}_1, \hat{C}_2 \in \hat{\mathcal{C}}$ and partitions $\Theta_s, \Theta_p \in \Theta$ so that $\widetilde{C}_1 \cap \Theta_s \neq \emptyset$, $\widetilde{C}_2 \cap \Theta_s \neq \emptyset$, $\widetilde{C}_1 \cap \Theta_p \neq \emptyset$ and $\Theta_s \subset \hat{C}_1, \Theta_p \subset \hat{C}_2$.

Suppose now that $\bar{\Theta}$ is a decomposition of $\bar{V} := \Theta_s \cup \Theta_p$. Then the refined decomposition Θ‘ $:= \Theta \setminus \{\Theta_s, \Theta_p\} \cup \{\bar{\Theta}_i \cap \Theta_s \,|\, \bar{\Theta}_i \in \bar{\Theta}\} \cup \{\bar{\Theta}_i \cap \Theta_p \,|\, \bar{\Theta}_i \in \bar{\Theta}\}$ would be better fitting to $\widetilde{\mathcal{C}}$, while still being a covering of $\hat{\mathcal{C}}$. The problem is how to identify the partitions Θ_s and Θ_p, without knowing $\widetilde{\mathcal{C}}$.

The following qualitative observation offers a heuristic solution: Since $\widetilde{\mathcal{C}}$ is optimal, we have $\hat{h}_f(\widetilde{C}_1, \widetilde{C}_1) \gg 0$ and therefore $\hat{h}_f(\Theta_s \cap \widetilde{C}_1, \Theta_p \cap \widetilde{C}_1) \gg 0$. But this gives $\hat{h}_f(\Theta_s, \Theta_p) \gg 0$, which is equivalent to $\check{h}_f(w_s, w_p) \gg 0$. So if we refine all partitions with $\check{h}_f(w_s, w_p) \gg 0$, we can be sure to refine also all partitions which destroy the covering property of Θ for $\widetilde{\mathcal{C}}$. Note that partitions that are already fitting to $\widetilde{\mathcal{C}}$, are also fitting after a refinement.

Decomposition refinement algorithm

Let Θ be any decomposition of V with codebook $W := \{w_1, \ldots, w_{n_k}\}$.
(1) Identify all indices $s, p \in \{1, \ldots, n_k\}$ so that $\check{h}_f(w_s, w_p) > \sigma$ with $\sigma \gg 0$. Let I be the resulting index subset.
(2) Set $\bar{V} := \bigcup_{s \in I} \Theta_s$.
(3) Compute a decomposition $\bar{\Theta}$ of $\bar{V}$ with $\bar{n}_k$ partitions $\bar{\Theta}_i$.
(4) Set $\Theta` := \Theta \setminus \{\Theta_s \,|\, s \in I\} \cup \{\bar{\Theta}_i \cap \Theta_s \,|\, \bar{\Theta}_i \in \bar{\Theta}\,,\ s \in I\}$.

Obviously the above algorithm increases the number of partitions from n_k to maximally $n_k + (\bar{n}_k - 1)|I|$ partitions. Often several of the new partitions $\bar{\Theta}_i \cap \Theta_s$ are nearly empty. Therefore step (4) is improved by the following condition: Θ_s is replaced only by those $\bar{\Theta}_i \cap \Theta_s$, with $f(\bar{\Theta}_i \cap \Theta_s) \gg 0$. Note that in this case the refined Θ has to be adapted slightly to guarantee that it is still a decomposition. This can be easily done, if we use the SOM algorithm for the computation of the decomposition of $\bar{V}$:

Let $\bar{W}$ be the codebook of $\bar{\Theta}$ generated by the SOM algorithm. For each $s \in I$ we set $I_s := \{i \,|\, f(\bar{\Theta}_i \cap \Theta_s) \gg 0\}$. Then the reduced codebook W_{I_s} defines a decomposition $\bar{\Theta}_{W_{I_s}}$ of $\bar{V}$. If we replace Θ_s by $\{\bar{\Theta}_{s,i} \cap \Theta_s \,|\, \bar{\Theta}_{s,i} \in \bar{\Theta}_{W_{I_s}}\}$ for all $s \in I$, the refined Θ is still a decomposition of V.

Instead of the suggested refinement algorithm, one could also think about using methods that tries to grow the SOM adaptively [26, 17]. In this case one has to assure that the growing process is driven by the homogeneity function h. If the cluster problem is geometrically based, this should be no problem.

4.3 Approach based on Perron Cluster analysis

In this section, we will extend our general multilevel cluster approach by using results and methods from the theory of *Perron Cluster* analysis that has been recently developed by DEUFLHARD ET AL.. We will show that for cluster problems with a stochastic homogeneity functions, this extended approach can be used for a fast identification and efficient description of clusters, even if a correct number of clusters k is not known a priori.

4.3.1 Theoretical background

In the following, we will give a short description of the theory of Perron Cluster analysis. For details and proofs see [16, 13].

Suppose we have a primitive stochastic $n_{\mathbb{k}} \times n_{\mathbb{k}}$ matrix $\mathcal{S}$, i.e. there exist an $m \in N$ so that $\mathcal{S}^m > 0$, the entries $\mathcal{S}_{i,j}$ are non-negative and the sum of each row equals one. As a consequence, the constant vector $e = (1, \dots, 1)^T$ is an eigenvector corresponding to the simple eigenvalue $\lambda_1 = 1$ of $\mathcal{S}$. For all other eigenvalues λ_i of $\mathcal{S}$ we have $|\lambda_i| < 1$.

Let $\pi = (\pi_1, \dots, \pi_{n_{\mathbb{k}}})^T$ any strictly positive distribution so that $\pi^T e = 1$ and $\pi^T \mathcal{S} = \pi^T$. We suppose that $\mathcal{S}$ is reversible with respect to π, i.e. $\mathcal{D}^2 \mathcal{S} = \mathcal{S}^T \mathcal{D}^2$, where $\mathcal{D} := diag(\sqrt{\pi_i})$ is called a *weighting matrix* of $\mathcal{S}$. If $\mathcal{S}$ is reversible, it is self-adjoint with respect to the weighted scalar product $< x, y >_\pi := x^T \mathcal{D}^2 y$ and consequently, all eigenvalues are real. Additionally there exist a basis of π-orthogonal right eigenvectors, which diagonalizes $\mathcal{S}$ and for every right eigenvector Y there is an associated left eigenvector $\bar{Y} = \mathcal{D}^2 Y$, which corresponds to the same eigenvalue.

In the following let $I_1, \dots, I_k$ any disjoint index subsets with $I_p \subset \{1, \dots, n_{\mathbb{k}}\}$, $p \in \{1, \dots, k\}$, and $\bigcup_{p=1}^{k} I_p = \{1, \dots, n_{\mathbb{k}}\}$. Based on these index subsets we define a so called *coupling matrix* $\hat{\mathcal{S}} := (\mathcal{S}_{I_s,I_p})_{1 \le s,p \le k}$ via

$$\mathcal{S}_{I_s,I_p} := \sum_{i \in I_s} \sum_{j \in I_p} \frac{\pi_i \mathcal{S}(i,j)}{\sum_{i \in I_s} \pi_i}. \tag{4.1}$$

Lemma 4.3.1 *The matrix $\hat{\mathcal{S}}$ is stochastic and reversible with respect to the distribution $\hat{\pi} := (\hat{\pi}_1, \dots, \hat{\pi}_k)^T$ where $\hat{\pi}_p := \sum_{i \in I_p} \pi_i$.*

Proof: Since $\mathcal{S}$ is stochastic, i.e. $\sum_{j=1}^{n_{\mathbb{k}}} \mathcal{S}(i,j) = 1$ for $1 \le i \le n_{\mathbb{k}}$, we have

$$\begin{aligned} \sum_{p=1}^{k} \mathcal{S}_{I_s,I_p} &= \sum_{i \in I_s} \frac{\pi_i}{\hat{\pi}_s} \sum_{p=1}^{k} \sum_{j \in I_p} \mathcal{S}(i,j) \\ &= \sum_{i \in I_s} \frac{\pi_i}{\hat{\pi}_s} \sum_{j=1}^{n_{\mathbb{k}}} \mathcal{S}(i,j) \;=\; \sum_{i \in I_s} \frac{\pi_i}{\hat{\pi}_s} \;=\; 1 \end{aligned}$$

and therefore $\hat{\mathcal{S}}$ is stochastic. We further have

$$\hat{\pi}_s \mathcal{S}_{I_s,I_p} = \sum_{i \in I_s} \sum_{j \in I_p} \pi_i \mathcal{S}(i,j) \text{ for } 1 \le s, p \le k.$$

Since $\mathcal{S}$ is reversible, i.e. $\pi_i \mathcal{S}(i,j) = \pi_j \mathcal{S}(j,i)$ the reversibility of $\hat{\mathcal{S}}$ follows immediately. □

We are interested in index subsets $I_1, \ldots, I_k$ that lead to a nearly diagonal coupling matrix:

Definition 4.3.2 *Choose $p \in \{1, \ldots, k\}$. Then we call I_p an almost invariant aggregate of $\mathcal{S}$, if $\mathcal{S}_{I_p,I_p} \approx 1$. If I_p is an almost invariant aggregate of $\mathcal{S}$ for all $p \in \{1, \ldots, k\}$, we call $I_1, \ldots, I_k$ a covering set of almost invariant aggregates of $\mathcal{S}$. In this case we call k an optimal number of almost invariant aggregates of $\mathcal{S}$.*

One easily checks that almost invariant aggregates correspond to a permutation of $\mathcal{S}$ so that the matrix is nearly block-diagonal:

Lemma 4.3.3 *Let $I_1, \ldots, I_k \subset \{1, \ldots, n_{\Bbbk}\}$ any covering set of almost invariant aggregates of $\mathcal{S}$. Then the indices $\{1, \ldots, n_{\Bbbk}\}$ can be ordered so that the matrix $\mathcal{S}$ is of block-diagonally dominant form:*

$$\mathcal{S} = D + E = \begin{pmatrix} D_{1,1} & E_{1,2} & \ldots & E_{1,k} \\ E_{2,1} & D_{2,2} & \ldots & E_{2,k} \\ \ldots & \ldots & \ldots & \ldots \\ E_{k,1} & E_{k,2} & \ldots & D_{k,k} \end{pmatrix}.$$

Herein the perturbation matrix E satisfies $E = O(\epsilon)$ where ϵ is some perturbation parameter.

Supposing that the conditions of Lemma 4.3.3 hold, we set:

$$\mathcal{S}(\epsilon) := \mathcal{S}(0) + \epsilon\mathcal{S}^{(1)} + \epsilon^2\mathcal{S}^{(2)} + \ldots,$$

where $\mathcal{S}(0) = D$ is the unperturbed part of $\mathcal{S}$.

It follows from perturbation theory [45] that the spectrum of $\mathcal{S}(\epsilon)$ can be divided into two parts:

1. The *Perron Cluster* including the *Perron Root* $\lambda_1 = 1$ and the $k-1$ eigenvalues $\lambda_2(\epsilon), \ldots, \lambda_k(\epsilon)$ approaching 1 for $\epsilon \to\ 0$.
2. The remaining part of the spectrum, bounded away from 1 for $\epsilon \to 0$.

The eigenvectors corresponding to eigenvalues of the Perron Cluster have a useful property:

Lemma 4.3.4 *Let $\lambda_1(\epsilon), \ldots, \lambda_k(\epsilon)$ be the Perron Cluster of $\mathcal{S}$. Then there exits a covering set of almost invariant aggregate $I_1, \ldots, I_k$ of $\mathcal{S}$ so that the eigenvectors $Y_1, \ldots, Y_k \in \mathbf{R}^{n_{\Bbbk}}$, corresponding to $\lambda_1(\epsilon), \ldots, \lambda_k(\epsilon)$, are almost constant on each I_s, i.e. we have for all $s \in \{1, \ldots, k\}$:*

$$i, j \in I_s \implies (\forall p \in \{1, \ldots, k\})\, Y_p(i) \approx Y_p(j).$$

The above theoretical results lead to a powerful method for the determination of an optimal number k of almost invariant aggregates of $\mathcal{S}$:

Suppose there exist — a priori unknown — index subsets $I_1, \ldots, I_k$ so that the conditions of Lemma 4.3.3 hold. Then there exists an ϵ_* so that $\mathcal{S}(\epsilon_*) = \mathcal{S}$. If ϵ_* is sufficiently small, we can find a large gap within the spectrum of $\mathcal{S}$ between the eigenvalues λ_k and λ_{k+1} of $\mathcal{S}$. In this case k is an optimal number of almost invariant aggregates of $\mathcal{S}$.

But we cannot only determine an optimal number of almost invariant aggregates, also the index subsets themselves can be computed based on Lemma 4.3.4:

Let $Y_1, \ldots, Y_k \in \mathbf{R}^{n_{\Bbbk}}$ be the eigenvectors corresponding to the eigenvalues $\lambda_1(\epsilon), \ldots, \lambda_k(\epsilon)$ of $\mathcal{S}$. Then the identification of k groups of nearly identical k-tuple $Y(i) := (Y_1(i), \ldots, Y_k(i))^T$ of eigenvector components associated with each $i \in \{1, \ldots, n_{\Bbbk}\}$, is sufficient to identify the covering set of almost invariant aggregates $I_1, \ldots, I_k$ of $\mathcal{S}$. Obviously such a grouping can be done via the computation of a k-cluster set of the set $V_Y := \{Y(1), \ldots, Y(n_{\Bbbk})\}$ with frequency function $f_Y(v) := 1$ for $v \in V_Y$ and a suitable homogeneity function h_Y, e.g., $h_Y = h_d$, where d is a distance function in $\mathbf{R}^k$.

4.3.2 Stochastic homogeneity functions

In the following we suppose that the homogeneity function h is stochastic in V with respect to f:

Definition 4.3.5 *We call any homogeneity function $h : \Omega \times \Omega \longrightarrow [0, 1]$ stochastic in V with respect to f if we have*

$$\sum_{w \in V} h(v, w) f(w) = 1 \ \text{forall}\, v \in V. \tag{4.2}$$

Set $P(v, w) := h(v, w) f(w)$ for any $v, w \in V$. We can directly extend P on subsets of V, if we define for any non-void subsets $V_1, V_2 \subset V$:

$$\hat{P}(V_1, V_2) := \sum_{v \in V_1} \sum_{w \in V_2} \frac{f(v) P(v, w)}{f(V_1)}. \tag{4.3}$$

Using earlier definitions (see section 2.3) we get:

Lemma 4.3.6 $\hat{P}(V_1, V_2) = \hat{h}_f(V_1, V_2) \hat{f}(V_2)$ *for any non-void $V_1, V_2 \subset V$.*

Proof:

$$\begin{aligned}
\hat{h}_f(V_1, V_2) &= \frac{1}{f(V_1)f(V_2)} \sum_{v \in V_1} \sum_{w \in V_2} h(v, w) f(v) f(w) \\
&= \frac{1}{f(V_1)f(V_2)} \sum_{v \in V_1} \sum_{w \in V_2} P(v, w) f(v) \\
&= \frac{1}{\hat{f}(V_1)\hat{f}(V_2)} \hat{P}(V_1, V_2) f(V_1) \; = \; \frac{\hat{P}(V_1, V_2)}{\hat{f}(V_2)}
\end{aligned}$$

□

We have a sort of reversibility of P with respect to f:

Lemma 4.3.7 $f(v)P(v,w) = f(w)P(w,v)$ *for all* $v, w \in V$.

Proof: Since h is a homogeneity function, we have $h(v,w) = h(w,v)$ and therefore also

$$f(v)P(v,w) \; = \; f(v)h(v,w)f(w) \; = \; f(v)h(w,v)f(w) \; = \; P(w,v)f(w).$$

for all $v, w \in V$. □

From Lemma 4.3.7 directly follows:

$$f(V_1)\hat{P}(V_1, V_2) = f(V_2)\hat{P}(V_2, V_1)$$

for all non-void subsets $V_1, V_2 \subset V$.

Based on $\hat{P}$ and a decomposition of V we can define a stochastic and reversible matrix $\mathcal{S}$:

Lemma 4.3.8 *Let* $\Theta := \{\Theta_1, \dots, \Theta_{n_{\mathbb{k}}}\}$ *be any decomposition of* V. *Define the* $n_{\mathbb{k}} \times n_{\mathbb{k}}$ *matrix* $\mathcal{S}$ *via* $\mathcal{S}(i,j) := \hat{P}(\Theta_i, \Theta_j)$. *Further set* $\pi := (\pi_1, \dots, \pi_{n_{\mathbb{k}}})^T$ *with* $\pi_s := \frac{f(\Theta_s)}{f(V)}$. *Then we have:*
(i) *If for any* $i, j \in \{1, \dots, n_k\}$ *there exist* $p_1, \dots, p_m$, $m \in N$, *so that* $p_1 = i$, $p_m = j$ *and* $\mathcal{S}(p_t, p_{t+1}) > 0$ *for* $1 \le t \le m-1$, *then the matrix* $\mathcal{S}$ *is primitive.*
(ii) *The matrix* $\mathcal{S}$ *is stochastic.*
(iii) π *is a strictly positive distribution with* $\pi^T e = 1$ *and* $\pi^T \mathcal{S} = \pi^T$.
(iv) *The matrix* $\mathcal{S}$ *is reversible with respect to* π.

Proof:
(i) is obvious and (ii) follows directly from the fact that h is stochastic and Θ is a decomposition of the data set V.

(iii) Obviously we have $\pi^T e = 1$. Further let $\mathcal{S}_{*j} := (\mathcal{S}(1,j),\ldots,\mathcal{S}(n_{\mathbb{k}},j))^T$ be the j-th column of the matrix $\mathcal{S}$. Using Lemma 4.3.7 we have:

$$
\begin{aligned}
\pi^T\mathcal{S}_{*j} &= \frac{1}{f(V)}\sum_{i=1}^{n_{\mathbb{k}}} f(\Theta_i)\hat{P}(\Theta_i,\Theta_j)\\
&= \frac{1}{f(V)}\sum_{i=1}^{n_{\mathbb{k}}} f(\Theta_i)\sum_{v\in\Theta_i}\sum_{w\in\Theta_j}\frac{f(v)P(v,w)}{f(\Theta_i)}\\
&= \frac{1}{f(V)}\sum_{i=1}^{n_{\mathbb{k}}}\sum_{v\in\Theta_i}\sum_{w\in\Theta_j} f(v)P(v,w)\\
&= \frac{1}{f(V)}\sum_{i=1}^{n_{\mathbb{k}}}\sum_{v\in\Theta_i}\sum_{w\in\Theta_j} f(w)P(w,v)\\
&= \frac{1}{f(V)}\sum_{w\in\Theta_j} f(w)\sum_{i=1}^{n_{\mathbb{k}}}\sum_{v\in\Theta_i} P(w,v)\\
&= \frac{1}{f(V)}\sum_{w\in\Theta_j} f(w) \;=\; \pi_j.
\end{aligned}
$$

(iv) For any $i,j\in\{1,\ldots,n_{\mathbb{k}}\}$ we have:

$$
\begin{aligned}
\pi_i\mathcal{S}(i,j) &= \frac{f(\Theta_i)}{f(V)}\hat{P}(\Theta_i,\Theta_j)\\
&= \frac{1}{f(V)}\sum_{v\in\Theta_i}\sum_{w\in\Theta_j} f(v)P(v,w)\\
&= \frac{1}{f(V)}\sum_{v\in\Theta_i}\sum_{w\in\Theta_j} f(w)P(w,v)\\
&= \frac{f(\Theta_j)}{f(V)}\sum_{w\in\Theta_j}\sum_{v\in\Theta_i}\frac{f(w)P(w,v)}{f(\Theta_j)} = \pi_j\mathcal{S}(j,i).
\end{aligned}
$$

□

Based on $\mathcal{S}$ we can use Perron Cluster analysis to determine an optimal number k and to identify the almost invariant aggregates of $\mathcal{S}$. The following Theorem shows that a covering set of k almost invariant aggregates of $\mathcal{S}$ corresponds to a nearly optimal k-cluster set of $(\Theta,\hat{f},\hat{h})$ what we know is equivalent to a nearly optimal representative clustering for any codebook W of Θ (see Theorem 2.3.9).

Theorem 4.3.9 *Let k be an optimal number of almost invariant aggregates of the matrix $\mathcal{S}$ and let $I_1, \ldots, I_k \subset \{1, \ldots, n_{\mathbb{k}}\}$ be the corresponding covering set of almost invariant aggregates. Then we have:*
(i) $\frac{1}{k}\sum_{s=1}^{k} \mathcal{S}_{I_s,I_s} \geq 1 - \epsilon_*$*, with small* $\epsilon_* := 1 - \min_s \mathcal{S}_{I_s,I_s}$
(ii) *If we set* $\bar{\mathcal{C}} := \{\bar{C}_1, \ldots, \bar{C}_k\}$ *with* $\bar{C}_s = \{\Theta_p \,|\, p \in I_s\}$*, then* $\bar{\mathcal{C}}$ *is an nearly optimal* k*-cluster set of* $(\Theta, \hat{f}, \hat{h})$*, with* $\Gamma_{\hat{f},\hat{h}_f}(\bar{\mathcal{C}}) = \frac{1}{k}\sum_{s=1}^{k} \mathcal{S}_{I_s,I_s}$

Proof:
(i) Since each I_s is almost invariant, we have $\mathcal{S}_{I_s,I_s} \approx 1$ for $s = 1, \ldots, k$.
(ii) Obviously $\mathcal{C}$ is a k-cluster set of Θ. We have:

$$\begin{aligned}
\Gamma_{\hat{f},\hat{h}_f}(\bar{\mathcal{C}}) &= \frac{1}{k}\sum_{s=1}^{k} \frac{1}{\hat{f}(\bar{C}_s)} \sum_{V_1 \in C_s} \sum_{V_2 \in C_s} \hat{h}_f(V_1, V_2)\hat{f}(V_1)\hat{f}(V_2) \\
&= \frac{1}{k}\sum_{s=1}^{k} \frac{1}{\hat{f}(\bar{C}_s)} \sum_{p_1 \in I_s} \sum_{p_2 \in I_s} \hat{h}_f(\Theta_{p_1}, \Theta_{p_2})\hat{f}(\Theta_{p_1})\hat{f}(\Theta_{p_2}) \\
&= \frac{1}{k}\sum_{s=1}^{k} \frac{1}{\hat{f}(\bar{C}_s)} \sum_{p_1 \in I_s} \sum_{p_2 \in I_s} \hat{P}(\Theta_{p_1}, \Theta_{p_2})\hat{f}(\Theta_{p_1}) \\
&= \frac{1}{k}\sum_{s=1}^{k} \frac{1}{\hat{f}(\bar{C}_s)} \sum_{p_1 \in I_s} \sum_{p_2 \in I_s} \mathcal{S}(p_1, p_2) f(\Theta_{p_1}) \\
&= \frac{1}{k}\sum_{s=1}^{k} \frac{1}{\hat{f}(\bar{C}_s)} \sum_{p_1 \in I_s} \sum_{p_2 \in I_s} \frac{(\sum_{p \in I_s} f(\Theta_p))\mathcal{S}(p_1, p_2) f(\Theta_{p_1})}{\sum_{p \in I_s} f(\Theta_p)} \\
&= \frac{1}{k}\sum_{s=1}^{k} \frac{\sum_{p \in I_s} f(\Theta_p)}{\hat{f}(\bar{C}_s)} \sum_{p_1 \in I_s} \sum_{p_2 \in I_s} \frac{\mathcal{S}(p_1, p_2)\pi_{p_1}}{\sum_{p \in I_s} \pi_p} \\
&= \frac{1}{k}\sum_{s=1}^{k} \mathcal{S}_{I_s,I_s}.
\end{aligned}$$

Since $\frac{1}{k}\sum_{s=1}^{k} \mathcal{S}_{I_s,I_s} \leq 1$, we have $\Gamma_{\hat{f},\hat{h}_f}(\bar{\mathcal{C}}) \leq 1$ and therefore (i) guarantees that $\bar{\mathcal{C}}$ is nearly optimal. □

If we set $\mathcal{C} := \{C_1, \ldots, C_k\}$ with $C_s := \bigcup_{p \in I_s} \Theta_p$, then using Theorem 2.3.9 and Lemma 2.3.5 we have $\Gamma_{f,h}(\mathcal{C}) = \Gamma_{\hat{f},\hat{h}_f}(\bar{\mathcal{C}})$ and therefore $\mathcal{C}$ is a nearly optimal k-cluster set of (V, f, h).

Note that it is possible that there exist different k so that k is an optimal number of invariant aggregates. But this is not surprising, because cluster problems might also have different correct numbers of clusters.

Multilevel Reduction Algorithm for stochastic homogeneity functions

We can use our results for a special version of the multilevel reduction algorithm that can be used even if the number of clusters k is not known a priori:
(1) Compute a decomposition Θ (based on a codebook W) with adaptive choice of $n_{\Bbbk}$, $n_{\Bbbk} \leq \Bbbk \ll n$.
(2a) Compute the matrix $\mathcal{S}$.
(2b) Compute an optimal number k of almost invariant aggregates of $\mathcal{S}$ via Perron Cluster analysis.
(2c) Compute an optimal k-cluster set of (V_Y, f_Y, h_Y), leading to a covering set of k almost invariant aggregates $I_1, \ldots, I_k \subset \{1, \ldots, n_{\Bbbk}\}$ of $\mathcal{S}$.
(3) Set $\mathcal{C} := \{C_1, \ldots, C_k\}$ with $C_s = \bigcup_{p \in I_s} \Theta_p$. Then $\mathcal{C}$ is a k-cluster set of (V, f, h) with $\Gamma_{f,h}(\mathcal{C}) = \frac{1}{k} \sum_{s=1}^{k} \mathcal{S}_{I_s, I_s}$.
(4) Refine Θ so that the new decomposition Θ‘ of V with codebook W' is also a covering of $\mathcal{C}$.
(5) Repeat the steps (2a)-(2c) and (3) with Θ‘ instead of Θ, leading to a k‘-clustering $\mathcal{C}$‘ of (V, f, h).
(6) If $k‘ \neq k$ then set $\Theta := \Theta‘$ and $k = k‘$ and go to step (4)
(7) If $\Gamma_{f,h}(\mathcal{C}‘) > \Gamma_{f,h}(\mathcal{C})$, then set $\Theta := \Theta‘$ and go to step (4), else stop.

Identification of discriminating attributes based on Perron Cluster analysis.

In section 2.4.2 we have presented an algorithm for the identification of discriminating attributes. We now give a simple heuristic criterion to decide if an attribute set $\mathcal{A}(J^C)$ is redundant or not:

Let $\mathcal{C} := \{C_1, \ldots, C_k\}$ be any optimal k-cluster set of a data set V with a covering Θ_W that is defined based on a codebook W according to Eq. (3.2). Further let $W(J)$ be the projection of W on $\Omega(J)$ and $\Theta_{W(J)} := \{\Theta_1, \ldots, \Theta_{n_k}\}$ be the corresponding decomposition of $V(J)$. If the eigenvalue spectrum of the matrix $\mathcal{S}$ corresponding to Θ_W, is nearly the same as the spectrum of matrix $\mathcal{S}$‘ corresponding to $\Theta_{W(J)}$, then the attribute set $\mathcal{A}(J^C)$ is redundant.

The above criterion uses the obvious fact that an attribute set $\mathcal{A}(J^C)$ is redundant, if the cluster structure of the cluster problem is independent of the attributes in $\mathcal{A}(J^C)$.

Natural and artificial stochastic homogeneity functions

For a reversible dynamic system, the homogeneity function h_s, as defined in Lemma 1.1.4, is stochastic:

Lemma 4.3.10 *Let $(X(t))_{t=1,\ldots,T}$ any representative trajectory of length $T \in \mathbf{N}$ of a reversible dynamic system in Ω, i.e. $|\{t \,|\, X(t) = v, X(t+1) = w\}| = |\{t \,|\, X(t) = w, X(t+1) = v\}|$ for all $v, w \in \Omega$. Then the homogeneity function h_s is stochastic with respect to the frequency function f that is given by $f(v) := |\{t \,|\, X(t) = v\}|$.*

Proof: We have $f(v)S(v,w) = f(w)S(w,v)$ and so $h_S(v,w) = \frac{S(v,w)}{f(w)}$ for any $v, w \in V$. Since $\sum_{w \in V} S(v,w) = 1$ for all $v \in V$, h_S is stochastic. □

Note that the condition for $\mathcal{S}$ primitive (see Lemma 4.3.8) is true for any decomposition of $V := \{X(t) \,|\, t = 1, \ldots, T\}$ because one easily checks that for all $v, w \in V$, there exist $v_1, \ldots, v_m \in V$, $m \leq T$, so that $v = v_1$, $w = v_m$ and $S(v_i, v_{i+1}) > 0$ for $1 \leq i \leq m-1$.

In addition to natural given stochastic homogeneity functions, we can also construct them artificially: For each homogeneity function h there exists a transformation into a stochastic homogeneity function $\widetilde{h}$ with respect to a suitable frequency function:

Lemma 4.3.11 *Let V be any data set in Ω with homogeneity function h and frequency function f. Set $\widetilde{f}(v) := \sum_{w \in V} h(v,w)f(v)f(w)$ for all $v \in \Omega$. Define $\widetilde{h} : \Omega \times \Omega \longrightarrow [0,1]$ via*

$$\widetilde{h}(v,w) := \frac{h(v,w)f(v)f(w)}{\widetilde{f}(v)\widetilde{f}(w)} \quad v, w \in \Omega.$$

Then $\widetilde{h}$ is a stochastic homogeneity function with respect to $\widetilde{f}$.

For well structured simple cluster problems, i.e. cluster problems with clusters of nearly identical size and a nearly identical homogeneity and a nearly constant frequency function, we have $\widetilde{f}(v) \approx \mathrm{const.}$ and therefore $\widetilde{h}(v,w) \approx \mathrm{c} \cdot h(v,w)$, where c is a constant value. This guarantees that an optimal k-cluster set of $(V, \widetilde{f}, \widetilde{h})$ is nearly an optimal k-cluster set of (V, f, h). Note that in the case of geometric cluster problems with a distance function d, we usually have $h_d(v,w) > 0$ for nearly all $v, w \in V$, because h_d vanishes only for objects with maximal distance. Therefore the constructed matrix $\mathcal{S}$ will be primitive. We will use this observation to compute an optimal number of clusters for our simple example from section 1.4:

Example: Determination of a correct number of clusters

Obviously the cluster problem for the data set V as given by Figure 1.3 is well structured and the frequency function f is constant with $f(v) = 1$ for all $v \in V$. For $h = h_d$ with $d = d_{euclid}$ we get $\widetilde{f}(v) \in [4.90, 7.24]$ for $v \in V$. To reduce the variance we slightly modify our homogeneity function. We set

$$h(v,w) := 1 - \frac{d(v,w)}{(\max_{v,w \in V} d(v,w))} \ , v, w \in \Omega.$$

Obviously this homogeneity function is still suitable for the computation of geometrically based clusters. Now we get $\widetilde{f}(v) \in [3.81, 5.78]$ for $v \in V$, i.e. the variance has decreased. The modification of h has no influence on the ranking of optimal k-cluster sets for different k. We still cannot use the values $\Gamma_{f,h}(\mathcal{C}(k))$ to determine the optimal number of clusters (see Table 4.1).

optimal k-cluster set $\mathcal{C}(k)$	$\Gamma_{f,h}(\mathcal{C}(k))$
$\mathcal{C}(1) := V$	4.85
$\mathcal{C}(2) := \{\{a,b,c,d,e,f\},\{g,h,i\}\}$	3.67
$\mathcal{C}(3) := \{\{a,b,c\},\{d,e,f\},\{g,h,i\}$	2.72
$\mathcal{C}(4) := \{\{a\},\{b,c\},\{d,e,f\},\{g,h,i\}\}$	2.10
$\mathcal{C}(9) := \{\{a\},\{b\},\{c\},\{d\},\{e\},\{f\},\{g\},\{h\},\{i\}\}$	1.00

Table 4.1: **Example: Optimal k-cluster sets of (V, f, h) for different k with modified homogeneity function.**

Based on $\widetilde{h}$ and the trivial decomposition $\Theta_V := \{\{v\} \,|\, v \in V\}$, we can compute the matrix $\mathcal{S}$. The spectrum of $\mathcal{S}$ is given in Table 4.2:

λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9
1.000	0.577	0.165	0.046	0.041	0.025	0.018	0.015	0.010

Table 4.2: **Example: Spectrum of matrix $\mathcal{S}$.**

Obviously the large gap between the Perron Cluster and the remaining part of the spectrum is at $k = 2$, indicating that $\mathcal{S}$ has two almost invariant aggregates. Therefore the optimal number of clusters for our cluster problem is also 2. The fact that the distance between the Perron Root and λ_2 is also very large, is a result of the artificial construction of the stochastic homogeneity function $\widetilde{h}$. We will see in chapter 5 that for natural stochastic homogeneity functions, as e.g., the dynamically based function h_S, the Perron Cluster is always approaching 1, if at least there exist two clusters within the data.

Since the eigenvector associated with the Perron Root is the constant vector $e = (1, \ldots, 1)^T$, we only need the eigenvector Y_2 associated with λ_2, to compute the almost invariant aggregates.

We have $Y_2 = (-0.35, -0.21, -0.20, -0.13, -0.13, -0.13, 0.54, 0.53, 0.42)^T$. Comparing the components of Y_2 we can directly identify the almost invariant aggregates $I_1 := \{1, \ldots, 6\}$ and $I_2 := \{7, \ldots, 9\}$. One easily checks that this solution corresponds to the optimal k-cluster set $\mathcal{C}(2)$ of (V, f, h).