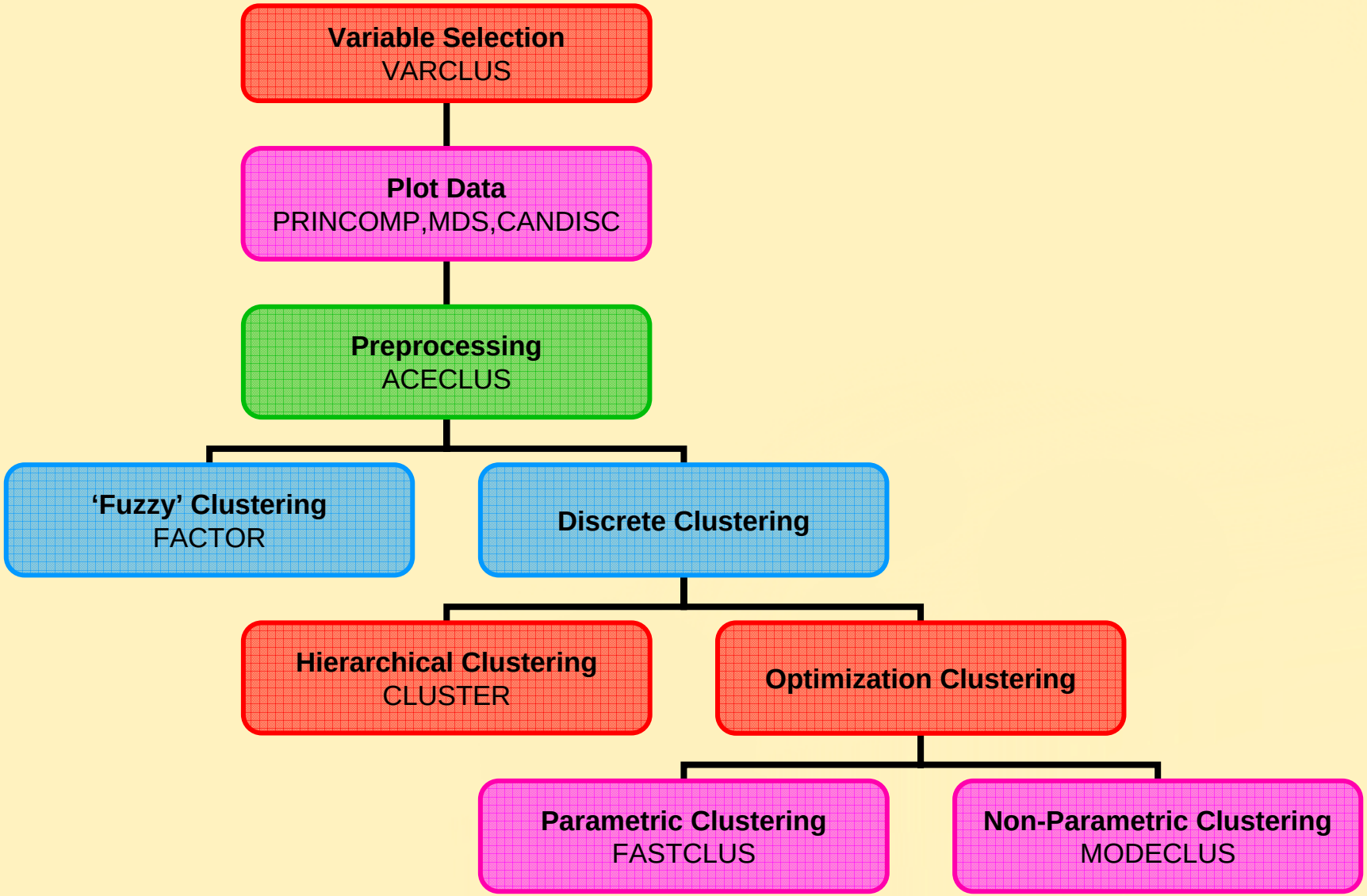
Clusterverfahren – bewährte statistische Technik und Basis für Data Mining Analysen

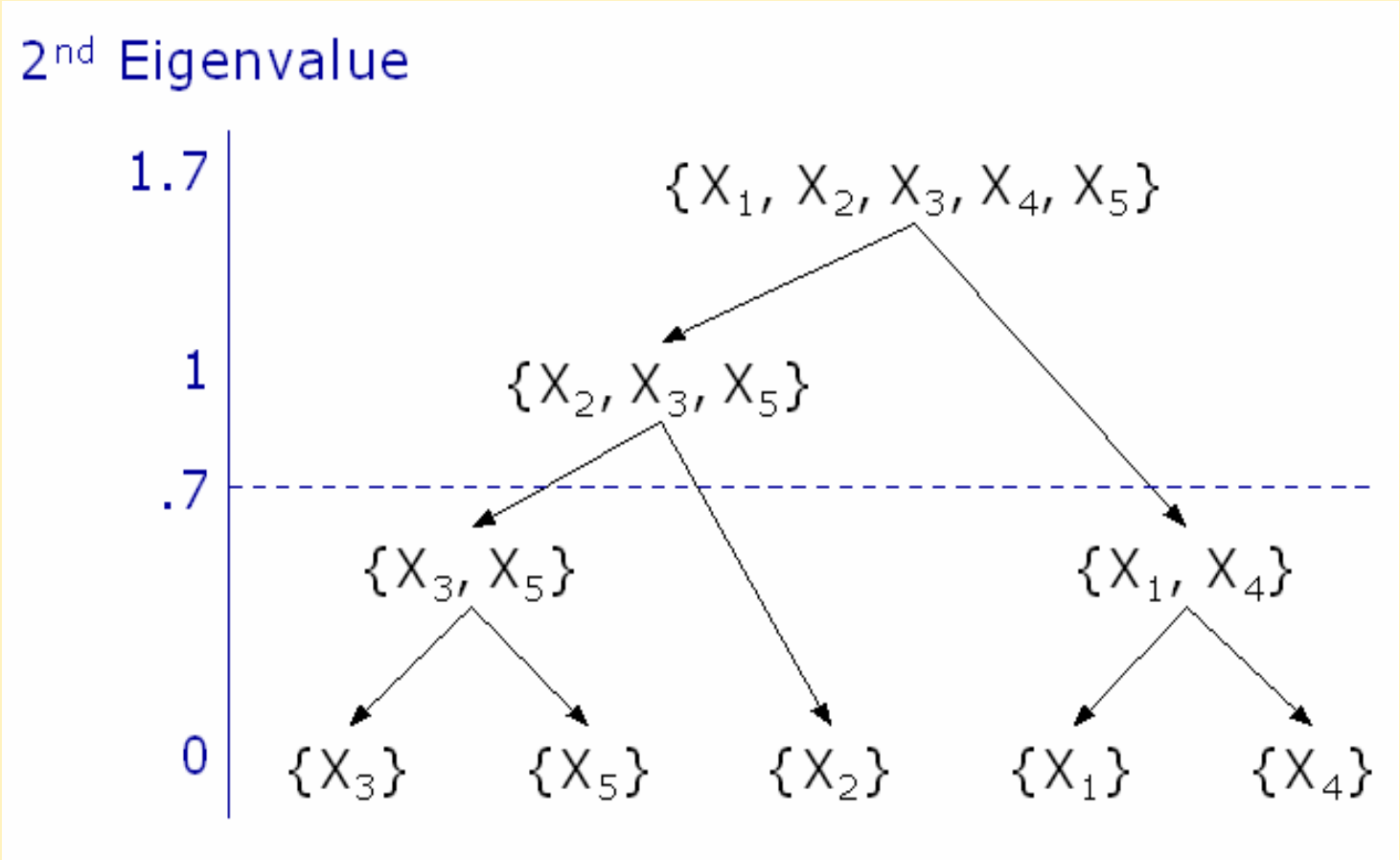
Dr. Reinhard Strüby
SAS Deutschland
Business Competence Center Analytical Solutions

Cluster-Techniken mit SAS



Divisive Clustering

PROC VARCLUS nutzt Divisive Clustering um Untergruppen von Variablen zu bilden, die sich möglichst stark unterscheiden.

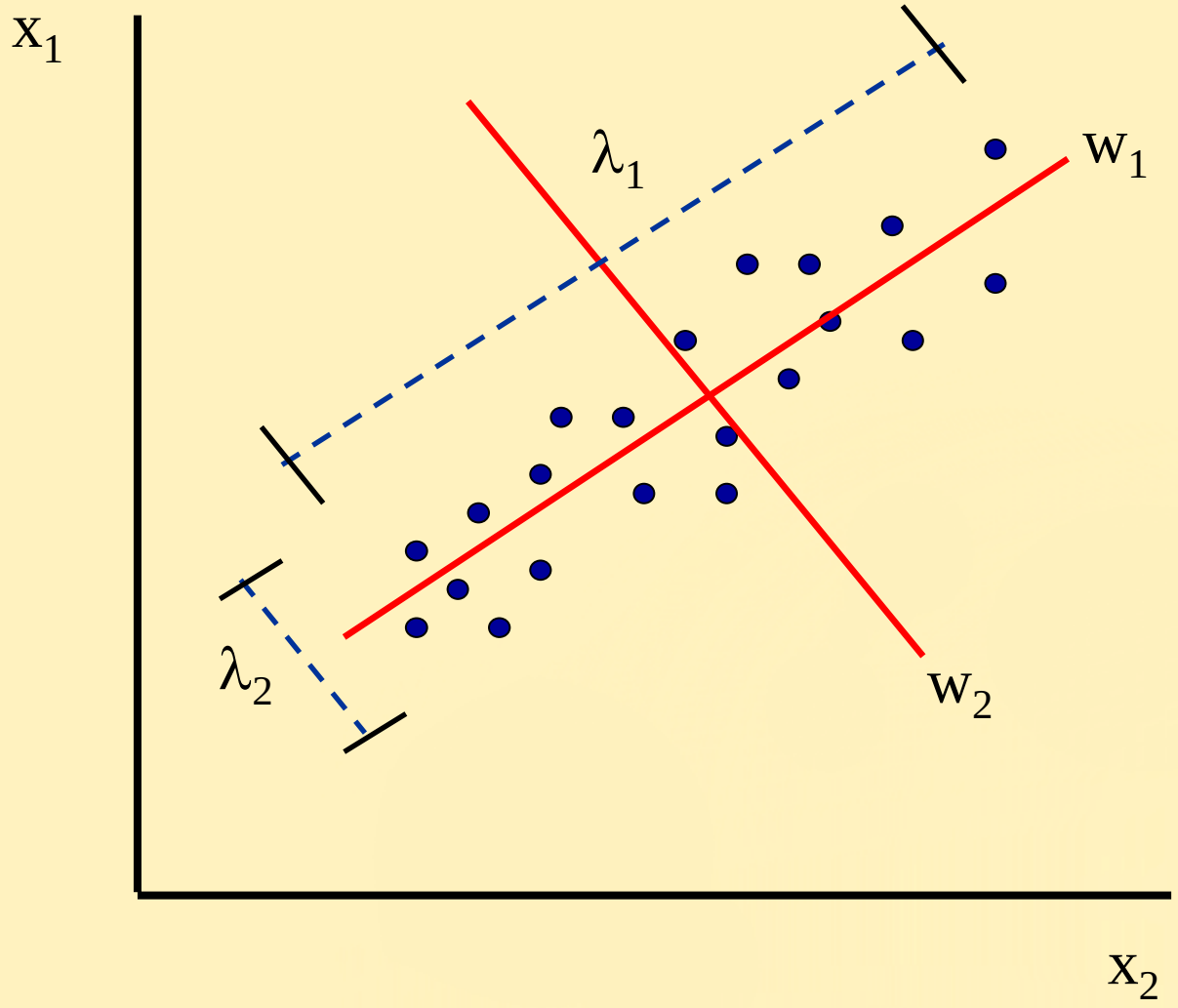


Variablen-Reduktion

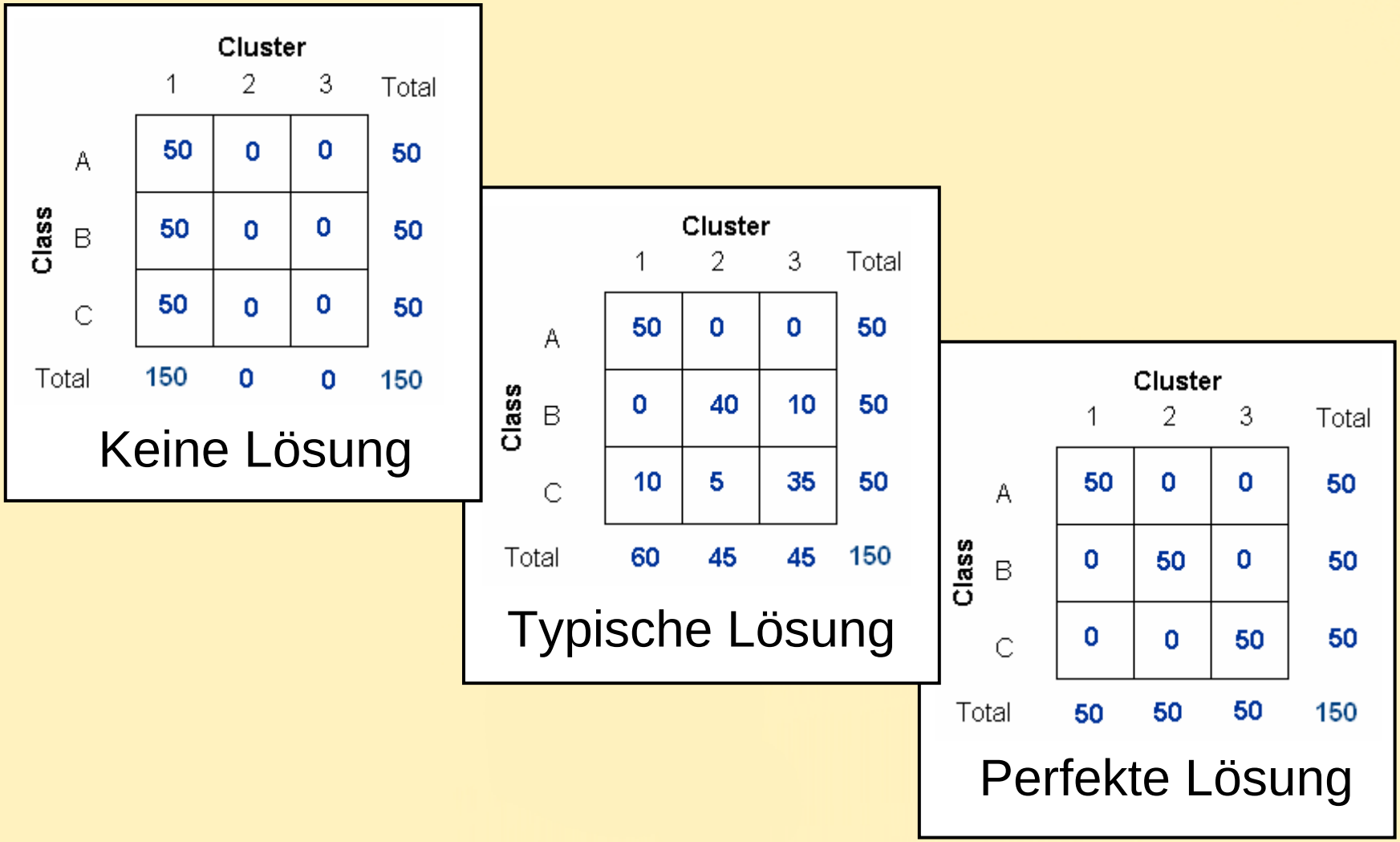
Cluster	Variable	R-squared with Own Cluster	R-squared with Next Closest	1-R**2 Ratio
Cluster 1	RedMeat	0.5350	0.2185	0.5950
	WhiteMeat	0.4544	0.3331	0.8181
	Eggs	0.7926	0.4902	0.4067
	Milk	0.5529	0.2721	0.6142
Cluster 2	Cereal	0.8255	0.4630	0.3250
	Nuts	0.8255	0.4549	0.3201
Cluster 3	Fish	0.7019	0.1365	0.3452
	Starch	0.7019	0.3075	0.4304
Cluster 4	FruitVeg	1.0000	0.0538	0.0000

PROC VARCLUS reduziert Variable, nicht Dimensionen.

Eigenwerte und Eigenvektoren



Qualität der Cluster-Lösung

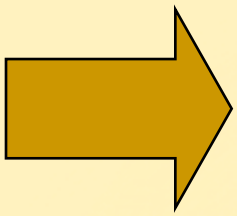


Wahrscheinlichkeit der Clusterzugehörigkeit

Gegeben durch die relativen Cluster-Häufigkeiten je Klasse:

Class	Cluster 1	2	3	Total
A	50	0	0	50
B	0	40	10	50
C	10	5	35	50
Total	60	45	45	150

Frequency



Class	Cluster 1	2	3	Total
A	1	0	0	1
B	0	0.8	0.2	1
C	0.2	0.1	0.7	1

Probability

Die Chi-Quadrat Statistik

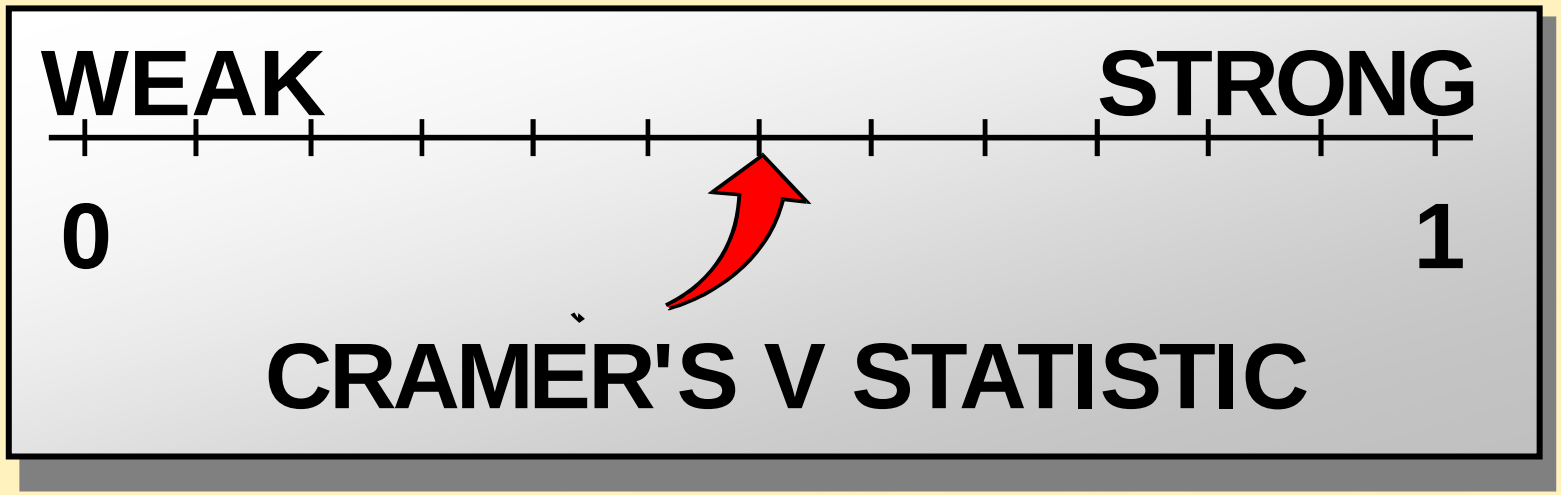
$$\lambda^2 = \sum_i \sum_j \frac{(\text{observed}_{ij} - \text{expected}_{ij})^2}{\text{expected}_{ij}}$$

The chi-square statistic (and associated probability)

- Prüft auf Assoziation
- Abhängig vom Stichprobenumfang
- Mißt **nicht** die Stärke der Assoziation

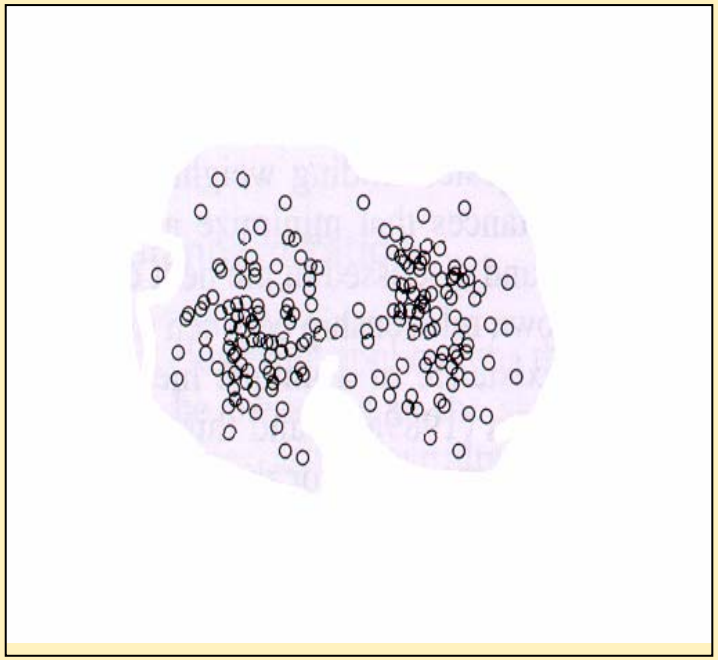
Beurteilung der Stärke der Assoziation

$$\text{Cramer's V} = \sqrt{\frac{\chi^2 / n}{\min(r-1, c-1)}}$$

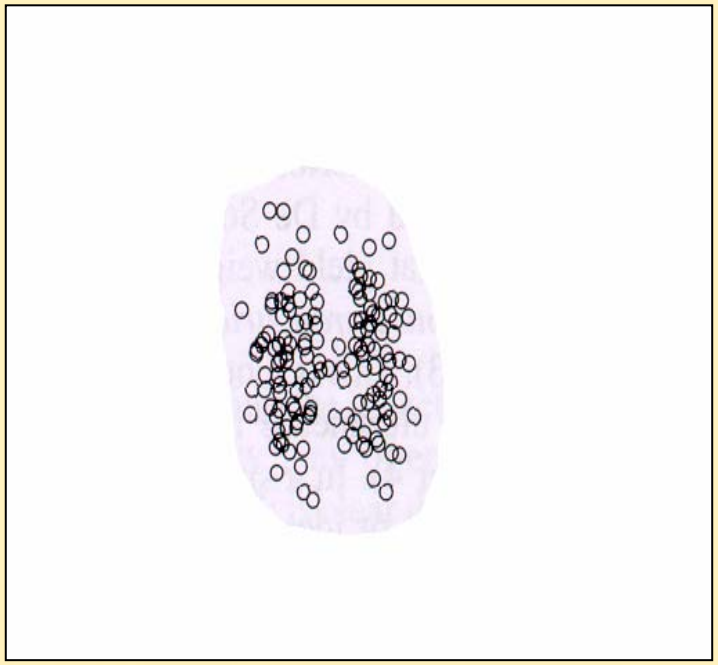


Cramer's V nimmt Werte von -1 bis 1 für 2x2-Tabellen an, Werte von 0 bis 1 sonst.

Das Standardisierungsproblem



vorher



nachher

Standardisierung kann Gruppendifferenzen mildern.

Die Prozedur STDIZE

Allgemeine Form:

```
PROC STDIZE METHOD=method <options>;
      VAR variables;
RUN;
```

Approximate Covariance Estimation for Clustering (ACECLUS)

Viele Clustermethoden arbeiten gut mit näherungsweise sphärischen Clustern. Sie sind nicht optimal für irregulär geformte Cluster.

Wenn man die Inner-Cluster-Covarianzmatrix kennen würde, könnten Daten so transformiert werden, dass mehr sphärisch geformte Cluster entstehen !

Da die Cluster nicht bekannt sind, kann die Kovarianzmatrix je Cluster nicht direkt berechnet werden.

Aber es gibt Schätzmethoden: Prozedur ACECLUS (Approximate Covariance Estimation for Clustering).

Die Prozedur ACECLUS

Starte mit einer Matrix **A** als Anfangsschätzung.

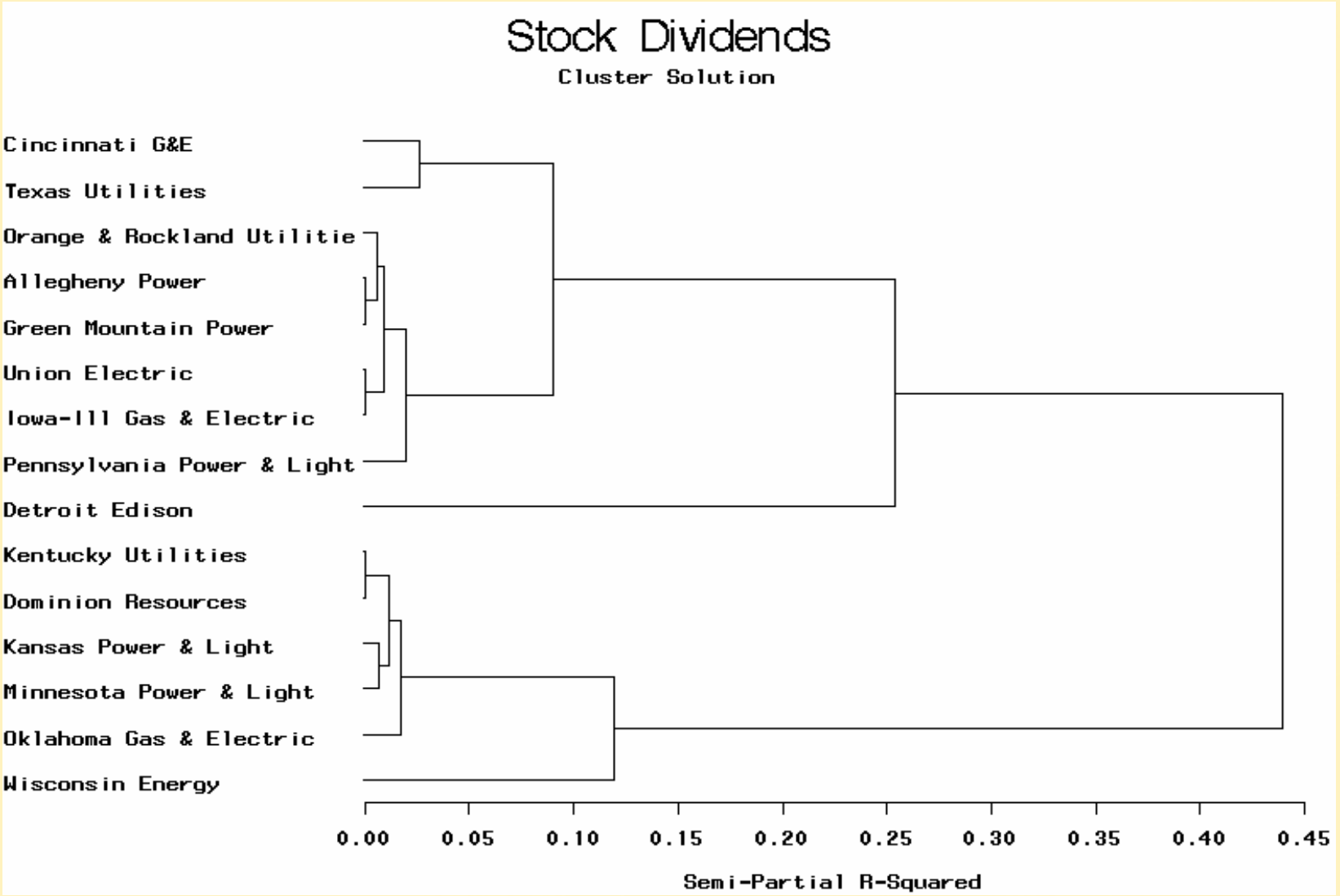
Berechne die inverse Matrix **M**=**A**$^{-1}$.

Berechne **A** neu durch:

$$a_{jk} = \frac{\sum_{i=2}^{n}\sum_{h=1}^{i-1} d_{ih}\left(x_{ij} - x_{hj}\right)\left(x_{ik} - x_{hk}\right)}{2\sum_{i=2}^{n}\sum_{h=1}^{i-1} d_{ih}}$$

4. Wiederhole 2 und 3 bis sich die Schätzung stabilisiert.

Hierarchical Clustering



The CLUSTER Procedure

General form of the CLUSTER procedure:

```
PROC CLUSTER METHOD=name <options>;
     COPY variables;
     VAR variables;
 RUN;
```

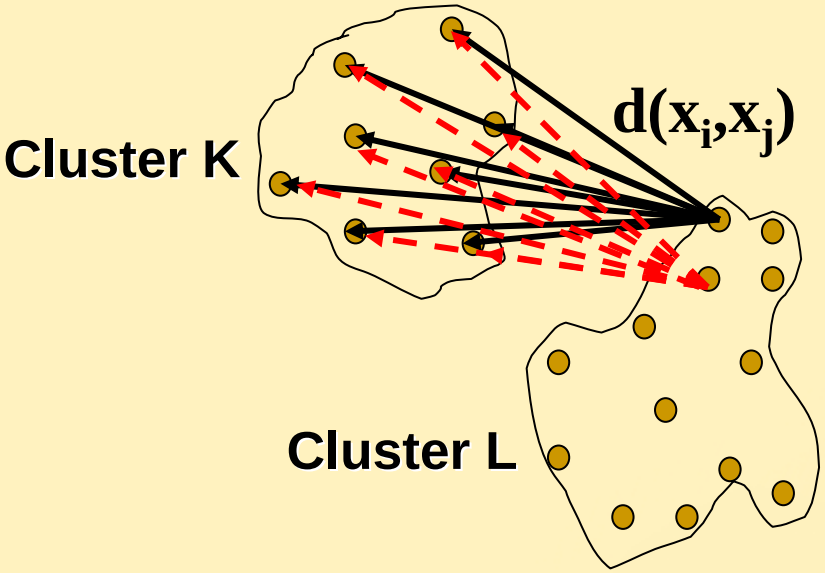
The TREE Procedure

General form of the TREE procedure:

```
PROC TREE OUT=SAS-data-set <options>;
     COPY variables;
 RUN;
```

Average Linkage

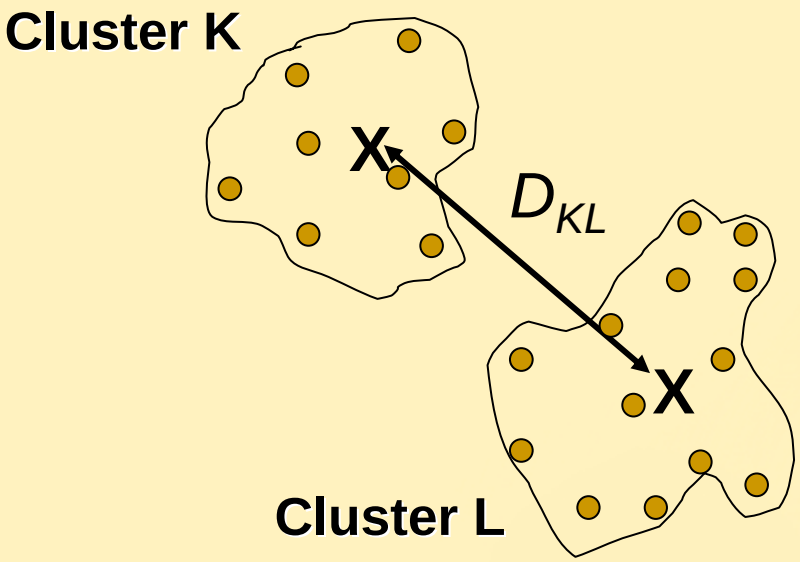
Der Abstand zweier Cluster ist der mittlere Abstand von Beobachtungspaaren.



$$D_{KL} = \frac{1}{n_K n_L} \sum_{i \in C_K} \sum_{j \in C_L} d\left(x_i, x_j\right)$$

Centroid Linkage

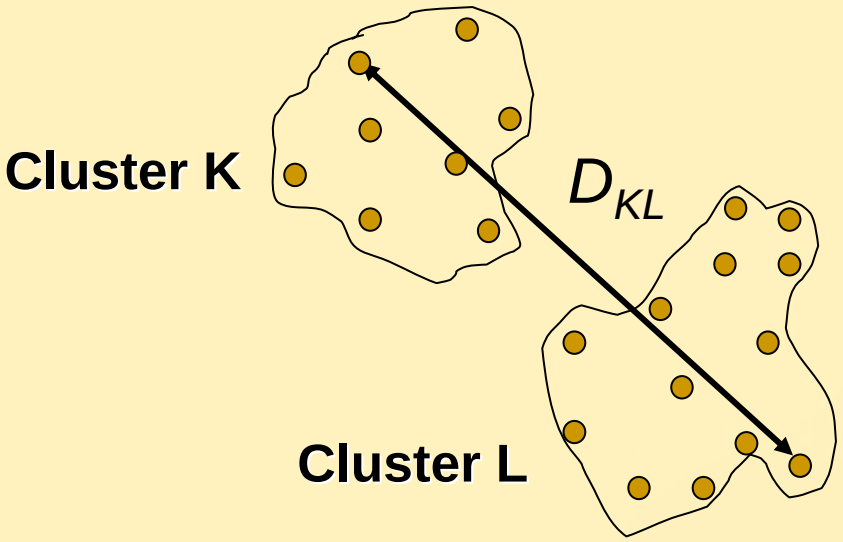
Der Clusterabstand ergibt sich aus dem quadrierten Euclidischen Abstand der Clustercentroids.



$$D_{KL} = \left\| \bar{x}_K - \bar{x}_L \right\|^2$$

Complete Linkage

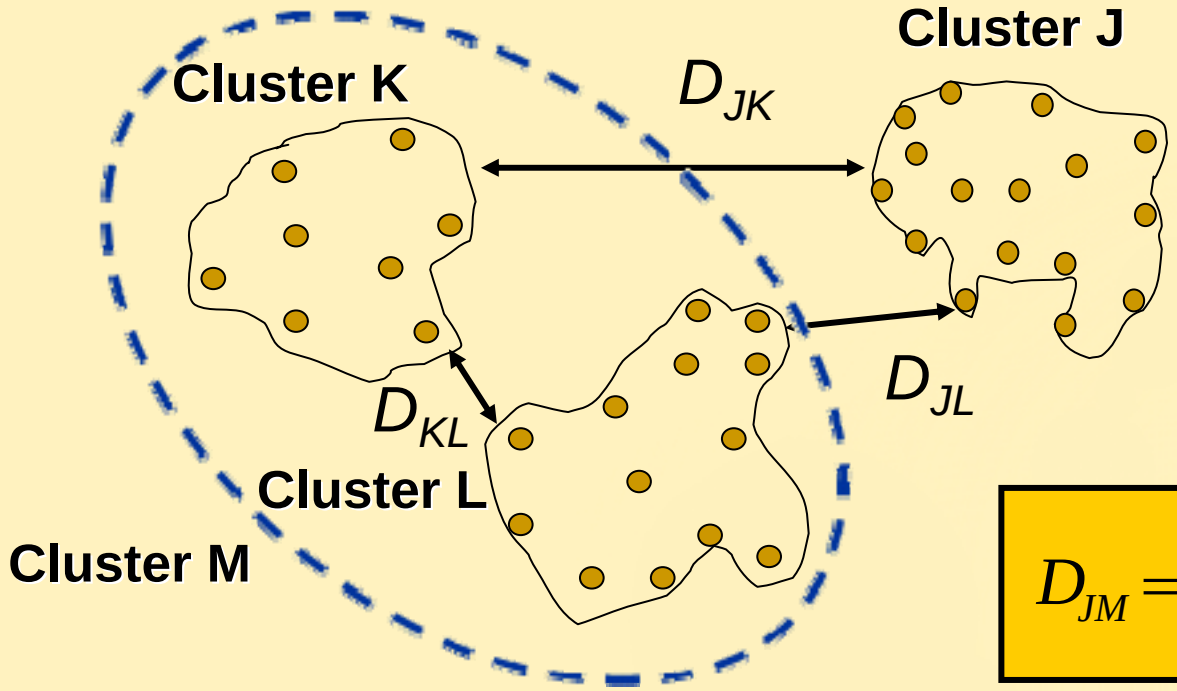
Maximaler Abstand von Beobachtungspaaren



$$D_{KL} = \max i \in C_K \max j \in C_L \; d(x_i, x_j)$$

Median Linkage

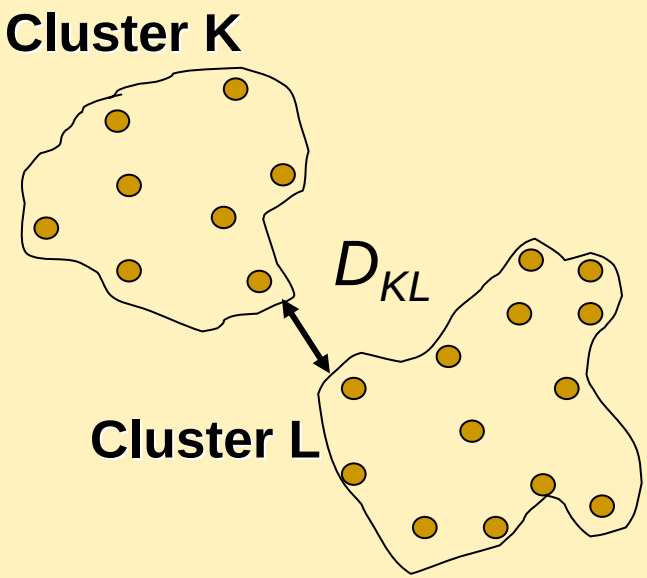
Mittlerer Abstand zwischen externem Cluster J und jedem der Komponentencluster (C_K and C_L), minus ¼ des Abstandes zwischen den Komponentenclustern.



$$D_{JM} = \frac{D_{JK} + D_{JL}}{2} - \frac{D_{KL}}{4}$$

Single Linkage

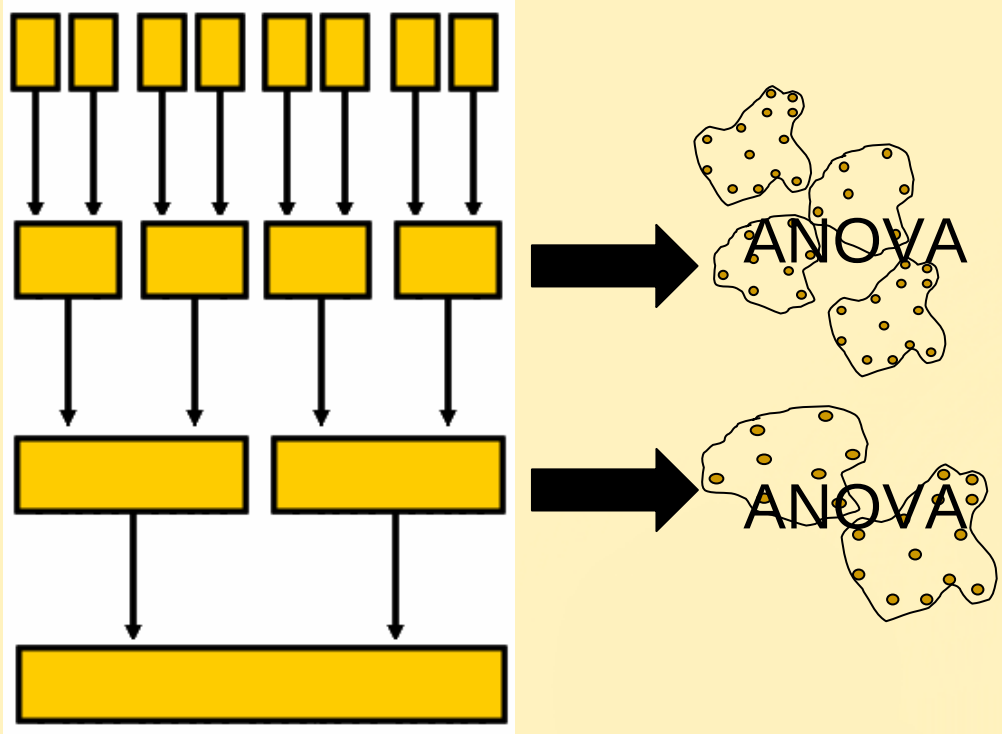
Minimaler Abstand zwischen Beobachtungspaaren.



$$D_{KL} = \min i \in C_K \min j \in C_L \, d(x_i, x_j)$$

Ward's Minimum-Variance

Ward's Methode nutzt ANOVA um in sich homogene Cluster mit maximaler Trennkraft zu finden.



$$D_{KL} = \frac{\left\| \bar{x}_K - \bar{x}_L \right\|^2}{\left(\frac{1}{n_K} + \frac{1}{n_L} \right)}$$

k-Means Clustering: 3-Schritt-Verfahren

Wahl der Anfangscluster (Centroids).

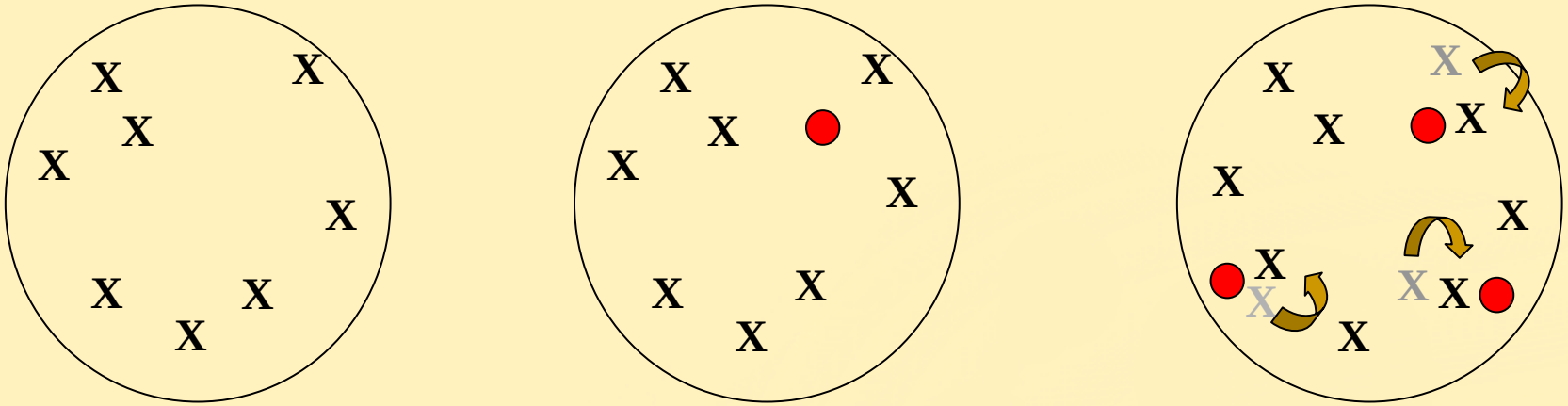
Einlesen der Beobachtungen und Update der Cluster-Zentren.

Schritt 2 wird wiederholt bis die Verschiebungen der Clusterzentren klein werden.

Jede Beobachtung wird dem nächstgelegenen Clusterzentrum zugeordnet. Dadurch entstehen die Endcluster.

MAXITER Option

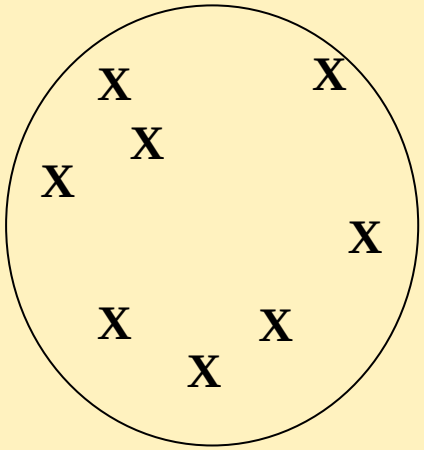
Die MAXITER Option adjustiert die Centroids simultan, nachdem alle Beobachtungen eingelesen wurden.



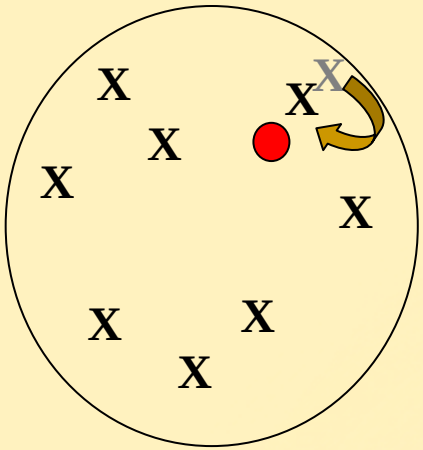
$Time_0$ $Time_1$ ... $Time_n$

DRIFT Option

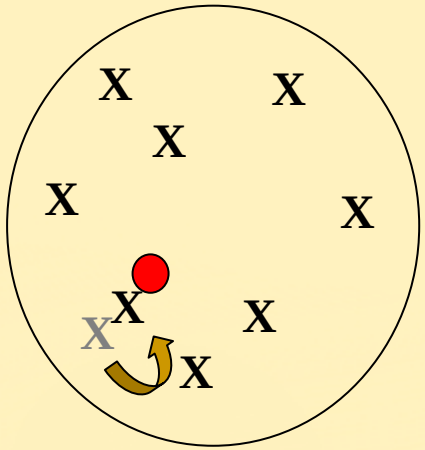
Die DRIFT Option adjustiert den nächstgelegenen Centroid, wenn Beobachtungen hinzukommen.



$Time_0$



$Time_1$



$Time_2$ …

The FASTCLUS Procedure

General form of the FASTCLUS procedure:

```
PROC FASTCLUS <MAXC= | RADIUS=><options>;
     VAR variables;
RUN;
```

Sarle's Cubic Clustering Criterion

$$\mathrm{CCC} = \ln\left[\frac{1-E(R^2)}{1-R^2}\right]\frac{\sqrt{\frac{np^*}{2}}}{\left(0.001+E(R^2)\right)^{1.2}}$$

Das *cubic clustering criterion* (CCC) testet die Hypothesen:

H_0 = Die Daten stammen aus einer Gleichverteilung in einem Unterraum

H_1 = Die Datenstammen aus einer Mischung von multivariaten Normalverteilungen mit gleichen Varianzen und gleichen Gewichten.

Positive Werte sprechen für mehr Struktur in den Daten als unter Normalverteilung zu erwarten wäre.

CCC soll maximiert werden:

Cluster History								
NCL	Clusters Joined		FREQ	SPRSQ	RSQ	ERSQ	CCC	Tie
10	CL12	CL44	25	0.0099	.911	.891	2.84	
9	CL11	CL15	45	0.0129	.898	.882	2.16	
8	CL18	CL27	18	0.0130	.885	.870	1.86	
7	CL17	CL22	29	0.0137	.872	.854	1.87	
6	CL8	CL20	26	0.0149	.857	.834	2.24	
5	CL10	CL16	30	0.0150	.842	.806	3.24	
4	CL19	CL7	49	0.0364	.805	.764	3.28	
3	CL9						4.00	
2	CL3	CL5	101	0.1331	.619	.552	2.86	
1	CL4	CL2	150	0.6188	.000	.000	0.00	

Empfohlene Lösung

Die Pseudo *F*-Statistik

$$\mathrm{PSF} = \frac{\left(\sum_{i=1}^{n} \left\|\mathbf{x}_i - \bar{\mathbf{x}}\right\|^2\right) / (g-1)}{\left(\sum_{j=1}^{g} \sum_{i \in c_k} \left\|\mathbf{x}_i - \bar{\mathbf{x}}_k\right\|^2\right) / (n-g)}$$

- Die Pseudo *F*-Statistik (PSF) mißt die Separation zwischen Clustern zu einem erreichten Hierarchielevel.
- Sie ist **nicht** *F*-verteilt.

Pseudo *F* soll ebenfalls maximiert werden:

Cluster History

NCL	Clusters Joined		FREQ	SPRSQ	RSQ	ERSQ	CCC	PSF	PST2	Tie
15	Oman	CL37	5	0.0039	.957	.933	6.03	132	12.1	
14	CL31	CL22	13	0.0040	.953	.928	5.81	131	9.7	
13	CL41	CL17	32	0.0041	.949	.922	5.70	131	13.1	
12	CL19	CL21	10	0.0045	.945	.916	5.65	132	6.4	
11	CL39	CL15	9	0.0052	.940	.909	5.60	134	6.3	
10	CL76	CL27	6	0.0075	.932	.900	5.25	133	18.1	
9	CL23	CL11	15	0.0130	.919	.890	4.20	125	12.4	
8	CL10	Afghanistan	7	0.0134	.906	.879	3.55	122	7.3	
7	CL9	CL25	17	0.0217	.884	.864	2.26	114	11.6	
6	CL8	CL20	14	0.0239	.860	.846	1.42	112	10.5	
5	CL14	CL13	45	0.0307	.829	.822	0.65	112	59.2	
4							0.57	122	14.8	
3								153	11.6	
2							-.82	135	48.9	
1	CL5	CL2	97	0.5866	.000	.000	0.00	.	135	

Empfohlene Lösung

Scoring nach FASTCLUS-Berechnungen

1. Rechne eine Clusternanalyse und speichere die Centroide.

```
PROC FASTCLUS OUTSTAT=centroids;
```

2. Lade die gespeicherten Centroide und score neue Daten.

```
PROC FASTCLUS INSTAT=centroids OUT=scored;
```

Berechnen einer k-Means Cluster-Lösung und Scoring neuer Daten

```
proc fastclus data=temp maxc=3 least=2
maxiter=5 out=clusout outstat=centroids;
   var &inputs;
run;
proc fastclus data=sbankte instat=centroids
least=2 out=scored;
   var &inputs;
run;
```

Scoring nach hierarchischer Clusterung

1. Hierarchische Clusteranalyse

```
PROC CLUSTER METHOD=method OUTTREE=tree;
RUN;
```

2. Cluster-Zuordnungen

```
PROC TREE DATA=tree OUT=out N=nclusters;
RUN;
```

(Continued)

Scoring nach hierarchischer Clusterung

3. Berechnung der Clustercentroide

```
PROC MEANS DATA=out;
        OUTPUT OUT=centroids;
RUN;
```

4. Einlesen der Centroide und Scoring

```
PROC FASTCLUS  DATA=new MAXC=nclusters
                        SEED=centroids MAXITER=0
                        OUT=scored;
RUN;
```

Scoring nach hierarchischer Clusterung

```
proc cluster data=distances method=ward pseudo
outtree=tree;
    var dist:; copy &inputs invest;
run;

proc means data=treeout noprint;
   class cluster;
   var &inputs;   output out=centroids mean=;
run;

proc fastclus data=sbankte seed=centroids maxc=3 maxiter=0
out=scored noprint;
   var &inputs;
run;
```