Unüberwachtes Lernen

Mustererkennung und Klassifikation, Vorlesung No. 12

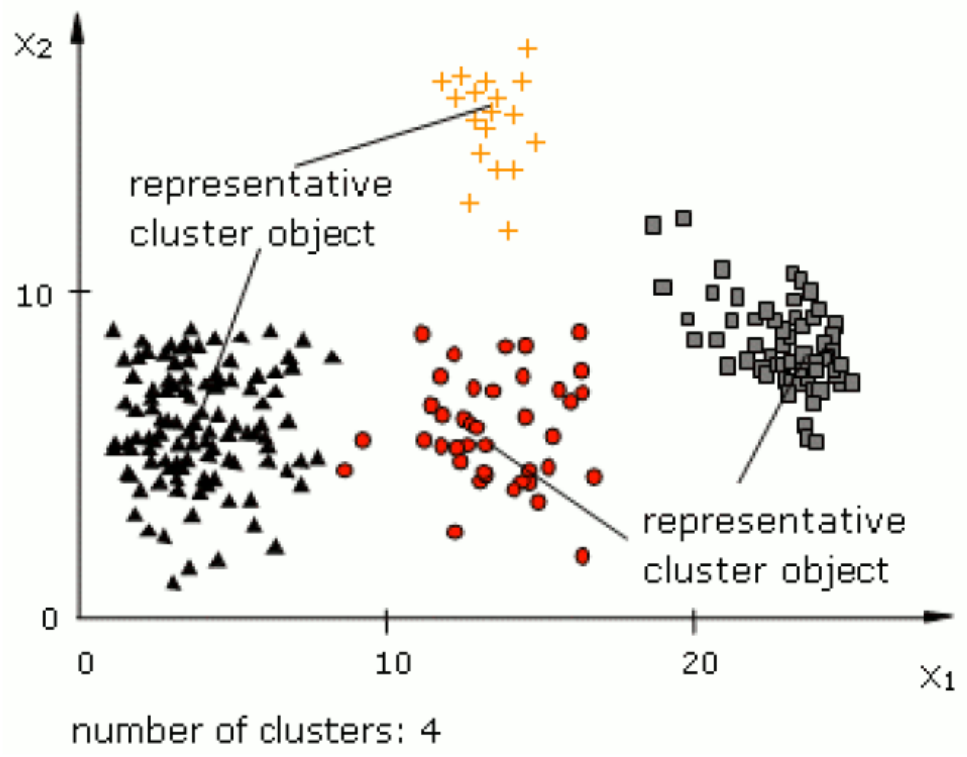
M. O. Franz

17.01.2008

Übersicht

Übersicht

Unüberwachtes Lernen

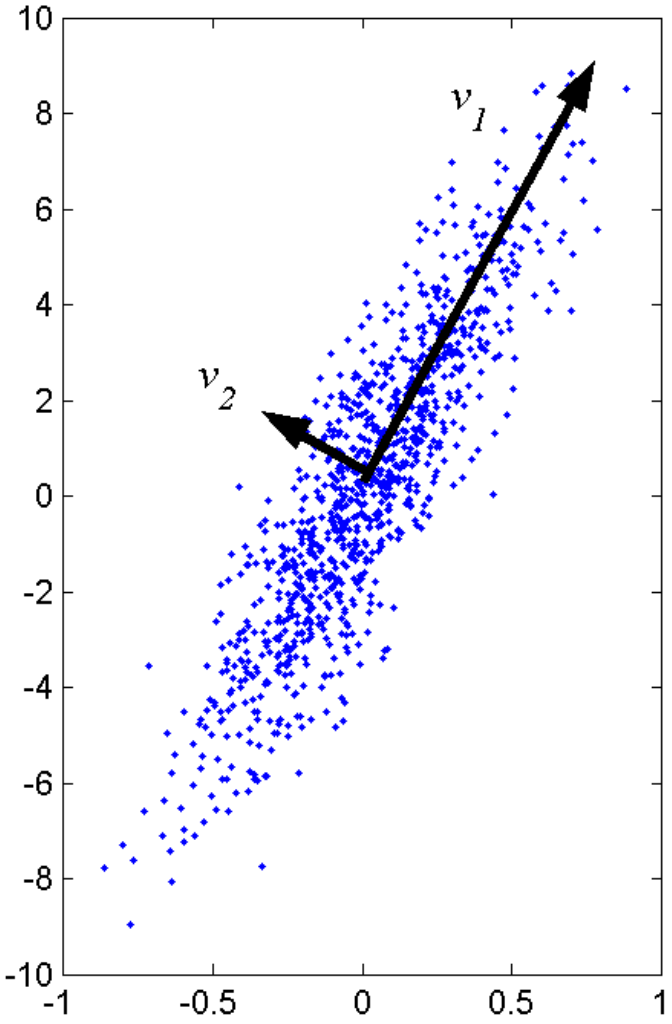


Bei unüberwachten Lernproblemen ist die Klassenzugehörigkeit der Trainingsbeispiele unbekannt.

Anwendungsfälle:

- Clusteranalyse: Die Daten werden aufgrund bestimmter Kriterien (z.B. räumliche Nähe) in Klassen zusammengefaßt.
- Dimensionsreduktion: Suche nach niedrigdimensionalen Repräsentationen der Daten zur Interpretation oder für Lernaufgaben.

Hauptkomponentenanalyse



In der Hauptkomponentenanalyse wird nach Richtungen in den Daten gesucht, entlang derer die Daten extremale (d.h. maximale oder minimale) Varianz haben.

Anwendung: Dimensionsreduktion durch Weglassen der Richtungen mit der kleinsten Varianz.

Annahme: die Daten seien zentriert, d.h. der Schwerpunkt

$$m = \frac{1}{n} \sum_i x_i$$

liegt am Ursprung.

Varianz entlang einer Raumrichtung

Gesucht sind Richtungen mit extremaler Varianz. Eine beliebige Richtung im $\mathbb{R}^d$ wird durch einen Einheitsvektor q dargestellt, auf die Datenpunkte x projiziert werden:

$$A = x^\top q = q^\top x \quad \text{mit} \quad \|q\| = \sqrt{q^\top q} = 1$$

Wir nehmen an, daß die Datenpunkte durch eine vektorwertige Zufallsvariable mit Erwartungswert $E[x] = 0$ erzeugt werden. Damit ist auch die Projektion A auf q eine Zufallsvariable mit

$$E[A] = q^\top E[x] = 0$$

Varianz in Richtung q:

$$\begin{aligned}
\sigma^2(q) &= E[A^2] - E[A]^2 \\
&= E[(q^\top x)(x^\top q)] - 0 \\
&= q^\top E[xx^\top] q \\
&= q^\top R q \quad \text{mit Korrelationsmatrix } R = E[xx^\top]
\end{aligned}$$

Raumrichtungen extremaler Varianz als Eigenwertproblem

An Richtungen extremaler Varianz ist die Ableitung von $\sigma^2(q) = 0$, d.h. bei einer infinitesimalen Richtungsänderung δq gilt

$$\sigma^2(q + \delta q) = \sigma^2(q) + O(\delta q^2)$$

Mit Hilfe von $\sigma^2 = q^\top R q$ ergibt sich

$$\sigma^2(q + \delta q) = (q + \delta q)^\top R(q + \delta q) = q^\top R q + 2(\delta q)^\top R q + O(\delta q^2)$$

$$\text{und} \quad \sigma^2(q + \delta q) = \sigma^2(q) + 2(\delta q)^\top R q$$

Damit beide Gleichungen gültig sind, muß $(\delta q)^\top R q = 0$ sein. Weiterhin gilt $\|q + \delta q\| = 1$, d.h. $(q + \delta q)^\top (q + \delta q) = 1$ damit muß auch hier $(\delta q)^\top q = 0$ sein. Zusammen ergibt sich

$$(\delta q)^\top R q + \lambda (\delta q)^\top q = 0 \quad \text{bzw.} \quad (\delta q)^\top (R q - \lambda q) = 0$$

und damit die Eigenwertgleichung

$$R q = \lambda q \quad \text{mit} \quad \lambda = \sigma^2(q)$$

Hauptkomponenten

Die Lösungen der Eigenwertgleichung der Korrelationsmatrix R

$$Rq = \lambda q$$

sind also die Richtungen extremaler Varianz, die Eigenwerte geben die Varianz der Daten entlang dieser Richtungen an.

Die Eigenwertgleichung hat i.A. d verschiedene, zueinander orthogonale Richtungen q_i als Lösung. Diese Richtungsvektoren werden als neue Basis im $\mathbb{R}^d$ gewählt. Die Projektionen eines Datenpunktes auf die q_i

$$a_i = q_i^\top x = x^\top q_i$$

heißen die Hauptkomponenten von x.

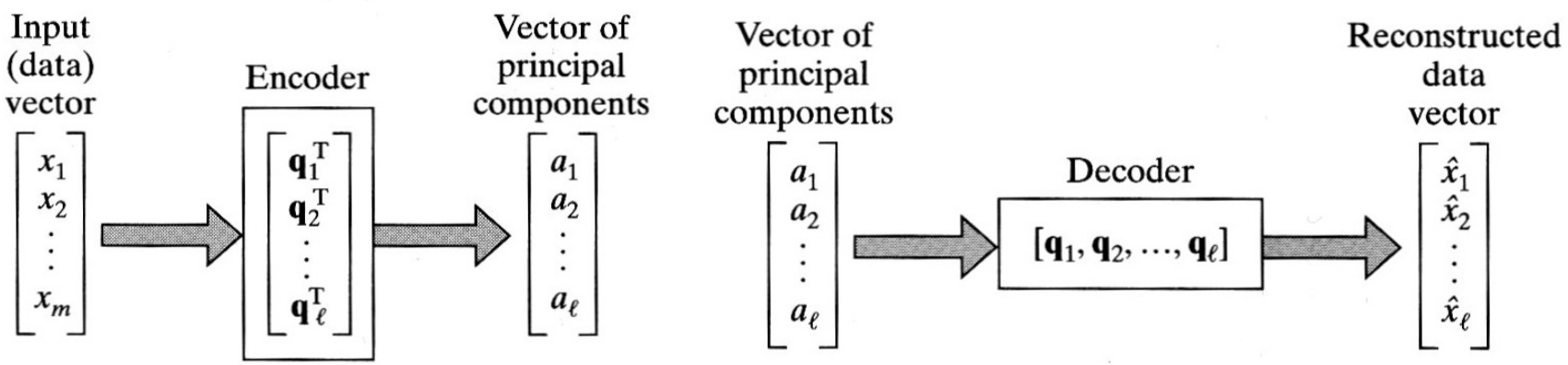
Die Darstellung eines Datenpunktes in diesem neuen Koordinatensystem entspricht also einer Rotation, nach der die Koordinatenachsen entlang der Richtungen extremaler Varianz zeigen.

Dimensionsreduktion

Prinzip: es werden nur die l Basisvektoren mit der größten Varianz beibehalten, Richtungen kleinerer Varianz werden verworfen. Der resultierende Vektor in $\mathbb{R}^l$ ist

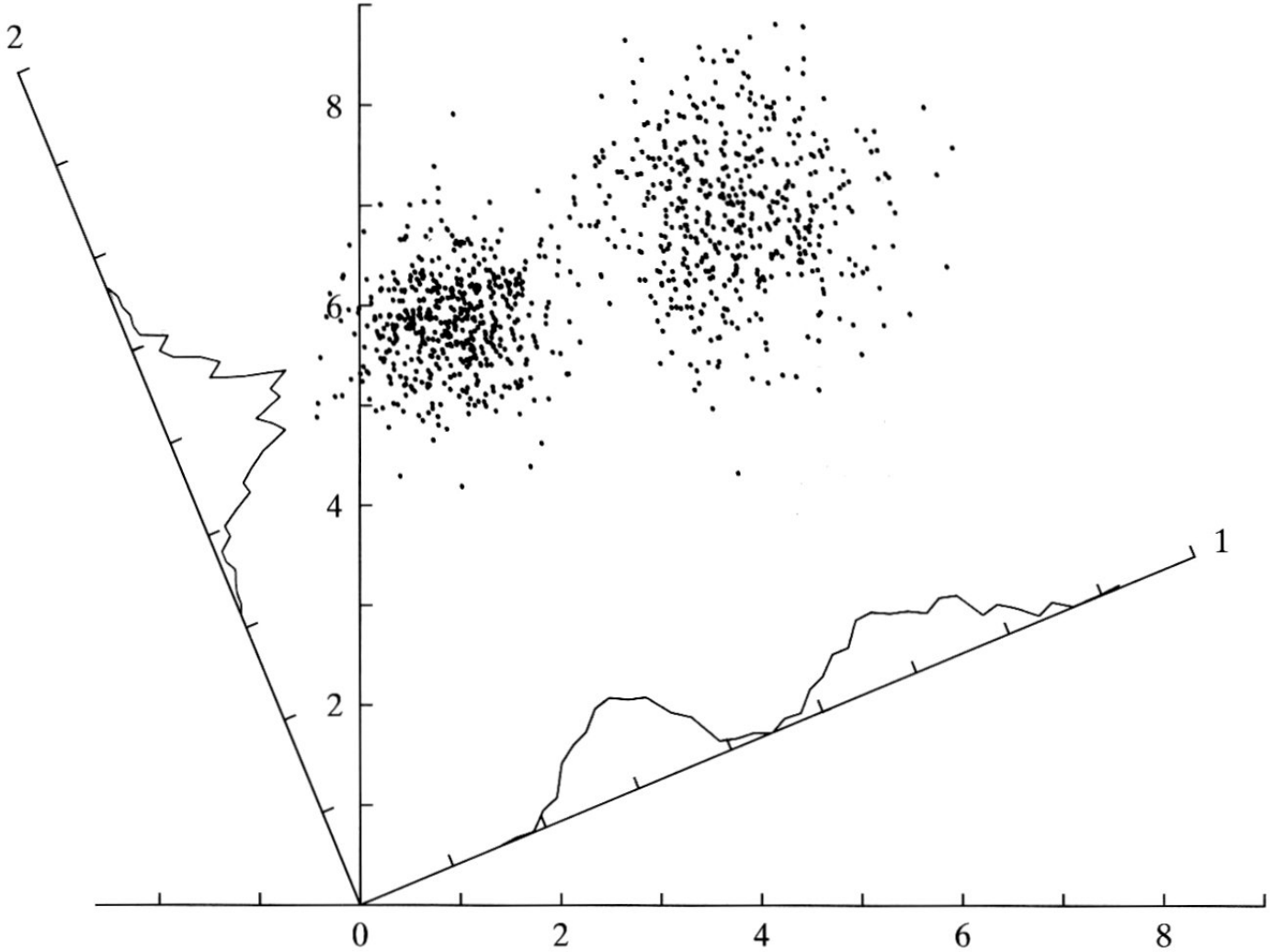
$$\hat{x} = \sum_{i=1}^{l} a_i q_i \quad \text{mit} \quad a_i = q_i^\top x.$$

Man kann zeigen, daß die so gewählte Basis den kleinsten Rekonstruktionsfehler unter allen Basen von $\mathbb{R}^l$ hat (s. Haykin, 99).



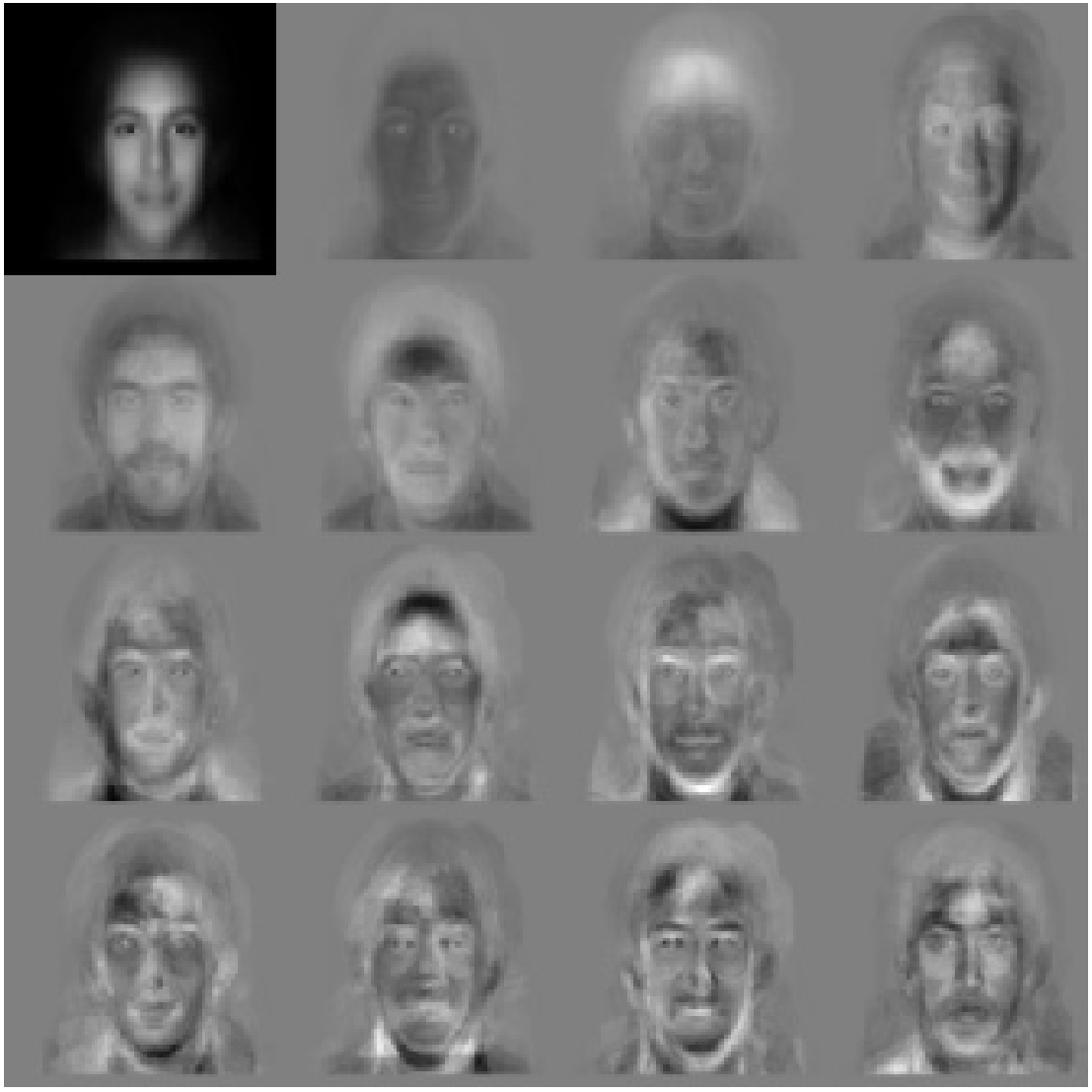
[Haykin 99]

2D-Beispiel



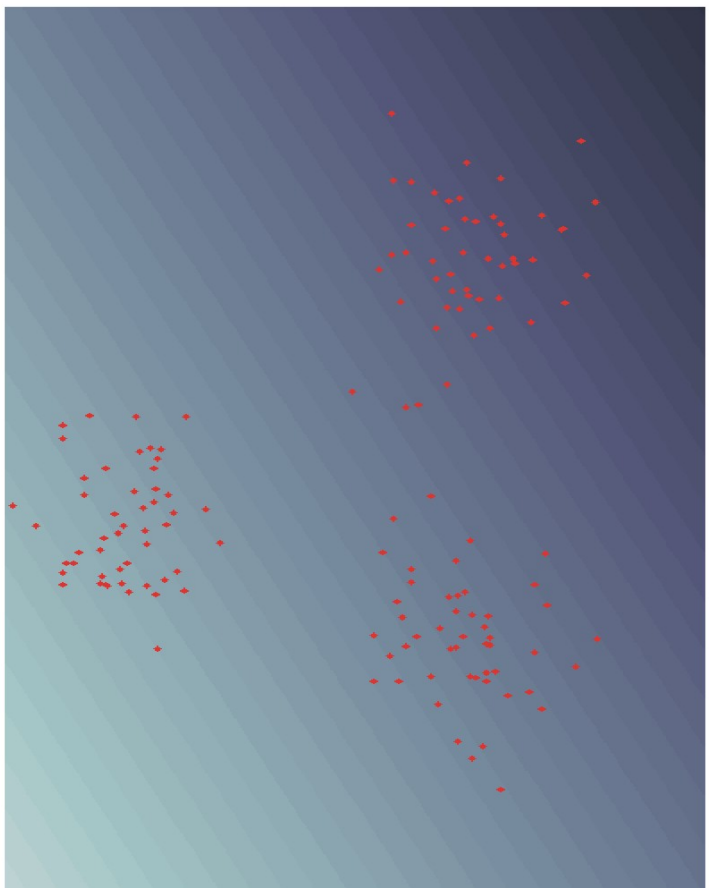
[Haykin 99]

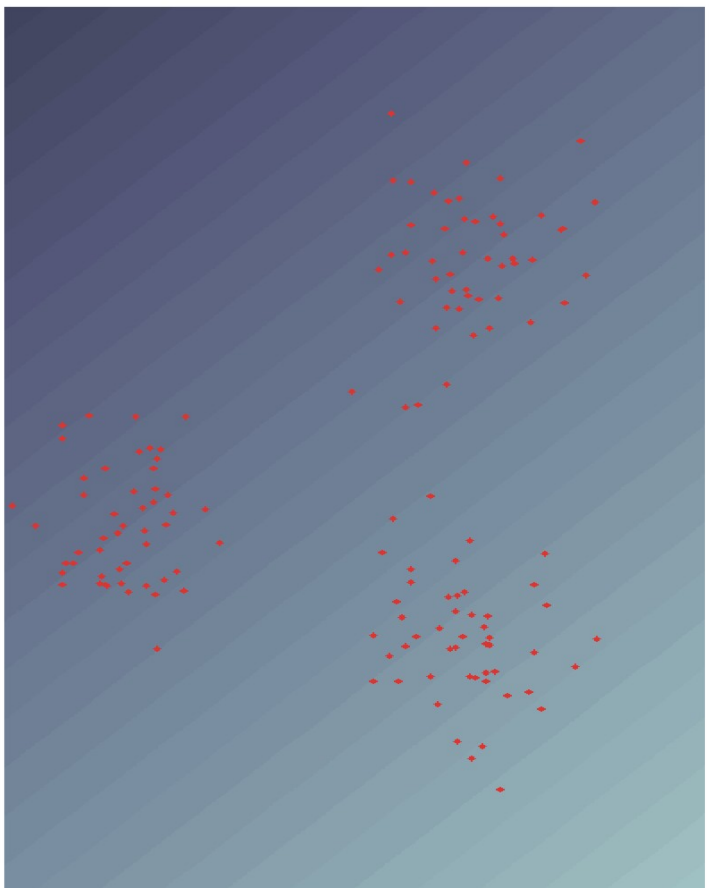
Beispiel: Eigengesichter



Übersicht

Hauptkomponentenanalyse auf 3 Clustern





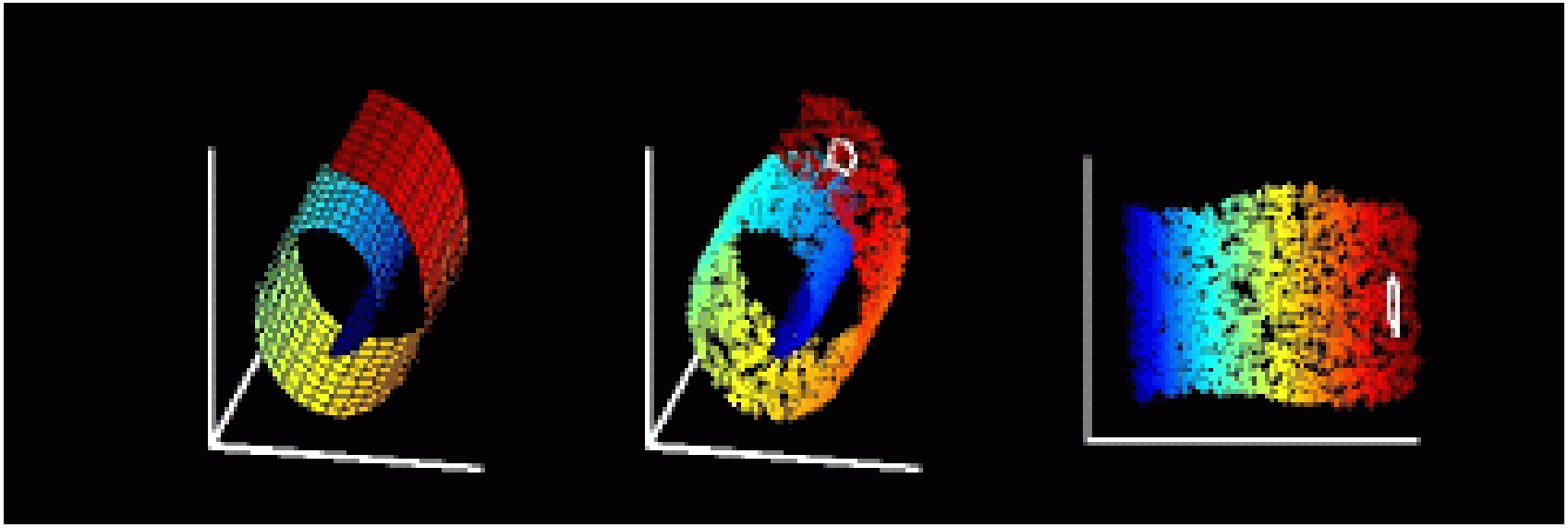
Dimensionsreduktion mit linearen Methoden

Die Hauptkomponentenanalyse ist eine lineare Methode. Sie beschreibt Unterräume mit besonders hoher bzw. niedriger Varianz.

- Die Basisvektoren q_i dieser Unterräume sind Linearkombinationen der Trainingsdatenpunkte, d.h. lineare Superpositionen dieser Daten (s. z.B. Eigengesichter). Für Bilder führen solche Repräsentationen oft zu unerwünschten Artefakten (Geisterbilder).
- Oft befinden sich die Daten auf niedrigdimensionalen Untermannigfaltigkeiten, die keinem Unterraum entsprechen, sondern gekrümmten Hyperflächen.
- Die Hauptkomponentenanalyse fängt nur paarweise Korrelationen ein, keine komplexeren statistischen Abhängigkeiten.

Für solche Probleme müssen nichtlineare Methoden zur Dimensionsreduktion eingesetzt werden.

Nichtlineare Hauptkomponentenanalyse (1)



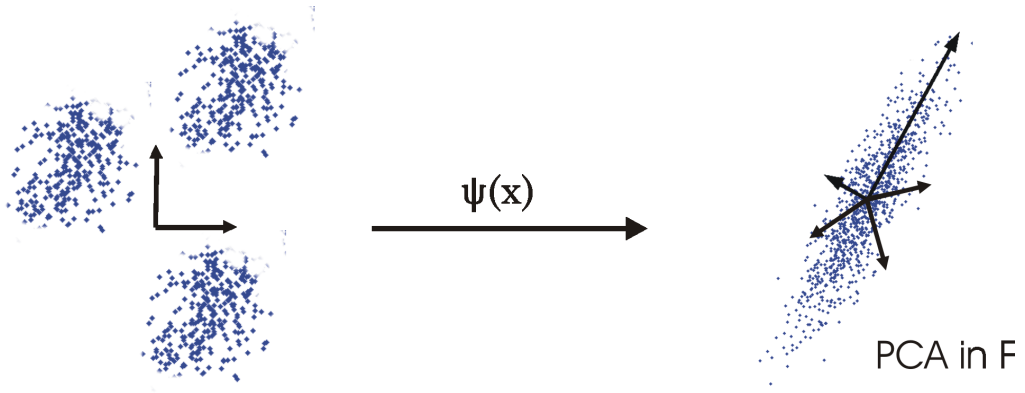
Grundidee: Nichtlineare Transformation der Daten in einen (oft höherdimensionalen) Merkmalsraum.

Hier: gleicher Ansatz wie bei nichtlinearer Klassifikation, d.h. wir (1) definieren eine Basis aus m nichtlinearen Funktionen $\varphi(x)$, (2) bilden darüber die Daten nichtlinear in einen erweiterten Merkmalsraum $\mathbb{R}^m$ ab und machen (3) dort eine Hauptkomponentenanalyse.

Nichtlineare Hauptkomponentenanalyse (2)

Analog zur Klassifikation werden die Basisfunktionen in einen erweiterten Merkmalsvektor geschrieben, z.B für eine Basis aus RBFs $\varphi_i(x) = e^{-\frac{\|x-x_i\|^2}{2}}$

$$y = \Phi(x) = \begin{bmatrix} e^{-\frac{\|x-x_1\|^2}{2}} \\ e^{-\frac{\|x-x_2\|^2}{2}} \\ \vdots \\ e^{-\frac{\|x-x_n\|^2}{2}} \end{bmatrix} \in \mathbb{R}^n$$

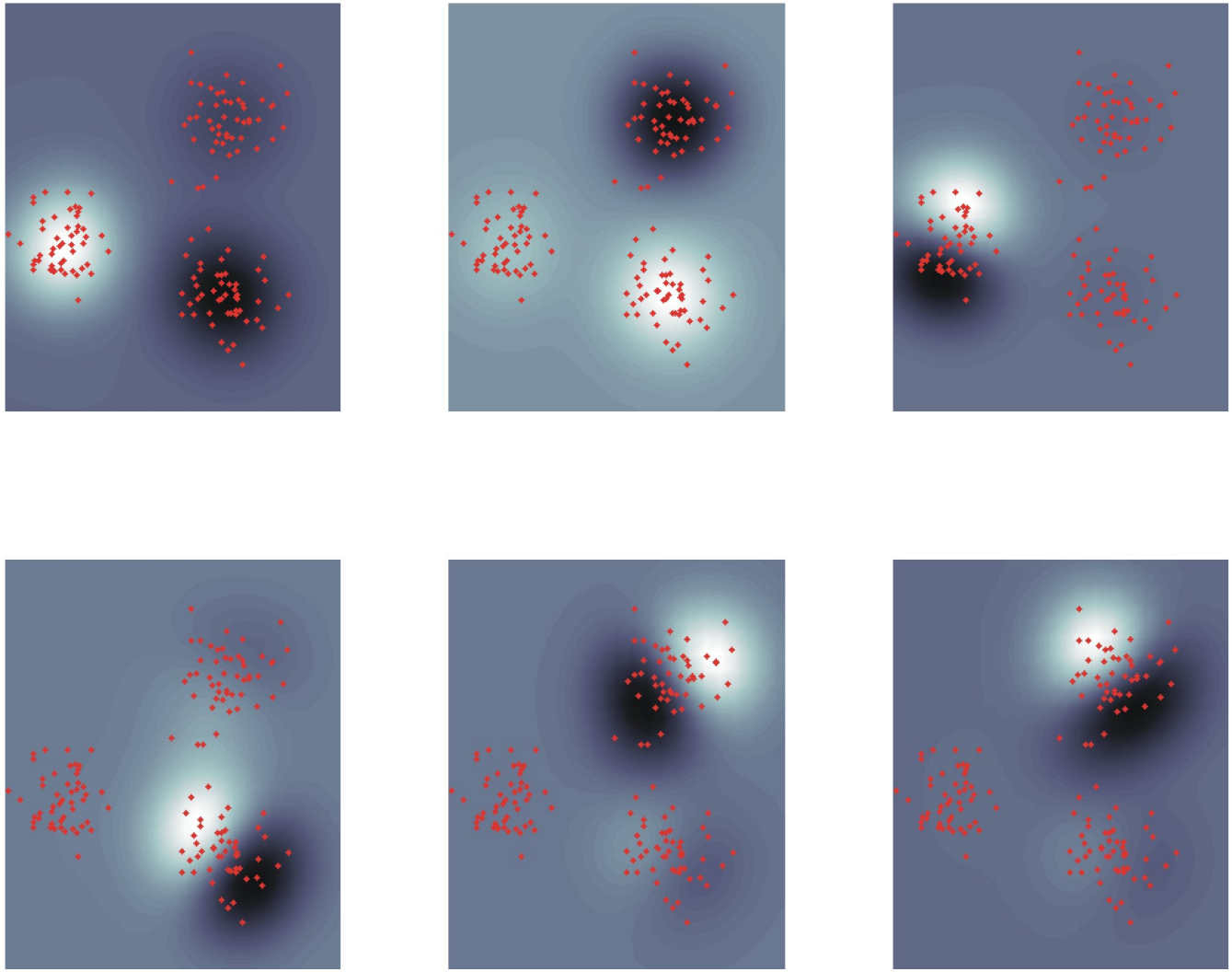


ergibt sich dann eine Korrelationsmatrix

$$R = E[\Phi(x)\Phi(x)^\top] = \frac{1}{n}\sum_{i=1}^{n} \Phi(x_i)\Phi(x_i)^\top,$$

deren Eigenwertzerlegung eine nichtlineare Hauptkomponenten-analyse im $\mathbb{R}^n$ darstellt.

Nichtlineare Hauptkomponentenanalyse auf 3 Clustern

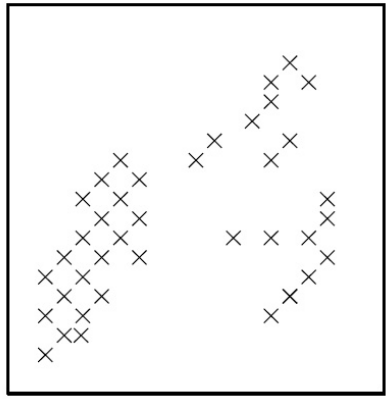


Übersicht

Clustering

Situation:

Eine Menge von Objekten und ein Ähnlichkeitsmaß zwischen diesen Objekten. Die Objekte seien als Punkte in einem Raum darstellbar, ihre Ähnlichkeiten als Distanzen.



Clustern:

geg. eine Punktmenge, partitioniere sie so in *Cluster*, daß sich Punkte im selben Cluster nahe sind, Punkte in verschiedenen Clustern weit auseinanderliegen.

- Hierarchisches Clustern
- Zentroidbasiertes Clustern
- Single-Linkage Clustering
- Probabilistische generative Modelle
- ...

K-Means-Clustering

1. Initialisierung: geg. Datenpunkte $x_n, n = 1 \ldots N$, setze K Prototypen auf Zufallswerte $m_k^{(0)}$
2. Zuweisungsschritt: weise jedem Datenpunkt x_n den nächsten Prototyp $\hat{k}_n$ zu, d.h.

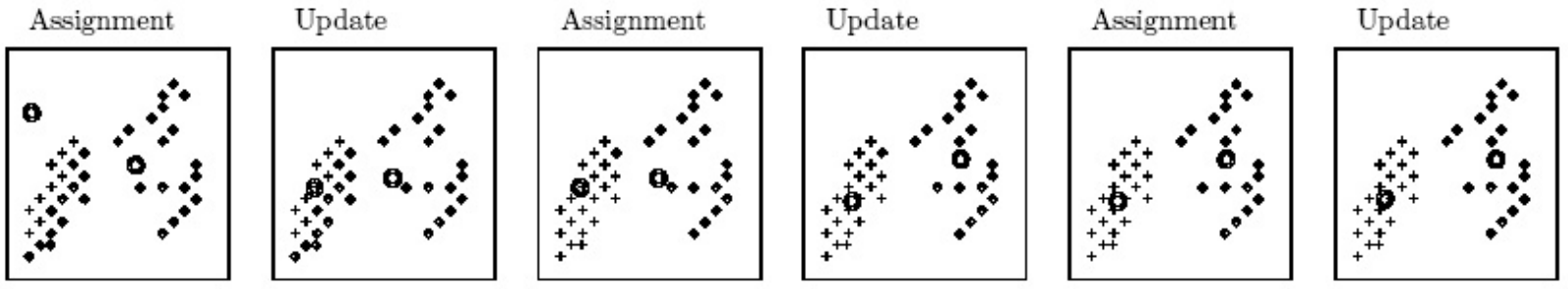
$$\hat{k}_n = \text{argmin}_k \|x_n - m_k\|^2$$

3. Anpassungsschritt: Setze m_k auf den Mittelwert der ihm zugewiesenen Datenpunkte

$$m_k = \frac{\sum_n r_{kn} x_n}{\sum_n r_{kn}}$$

   mit dem Indikator $r_{kn} = 1$ wenn $k = \hat{k}_n$, sonst $r_{kn} = 0$.
4. Wiederhole Zuweisungs- und Anpassungsschritt, bis sich die Zuordnung nicht mehr ändert.

Beispiel: K-means auf 2D-Daten



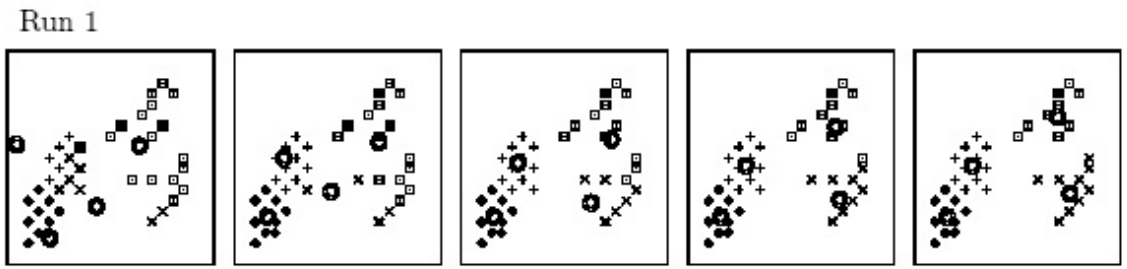
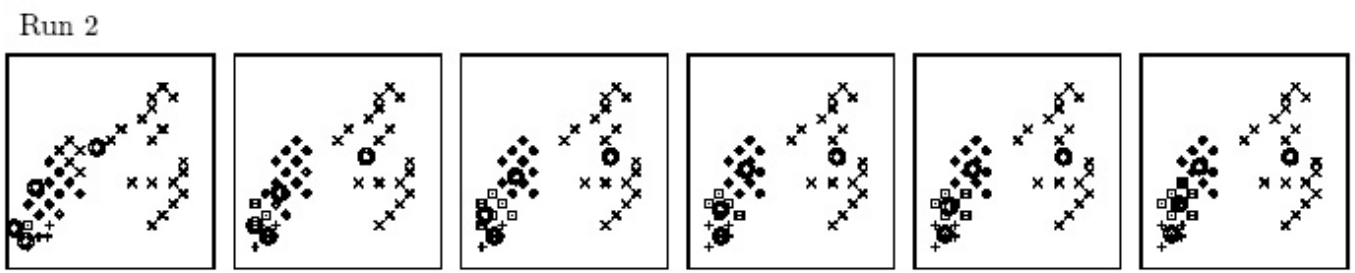
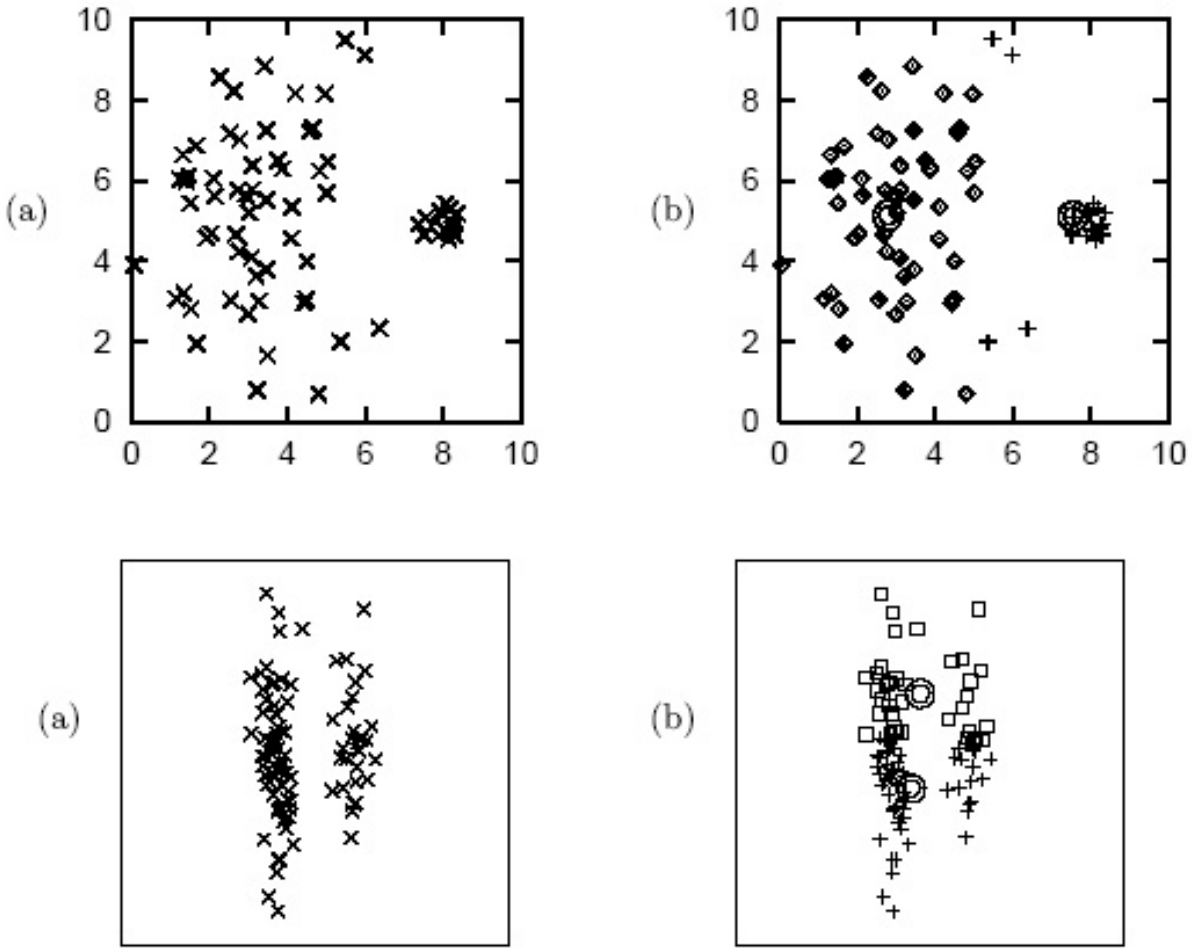


Figure 20.4. K-means algorithm applied to a data set of 40 points. Two separate runs, both with $K = 4$ means, reach different solutions. Each frame shows a successive assignment step.



[McKay 2003]

Problematische Fälle



[McKay 2003]

Probleme bei K-Means

Ad-hoc-Merkmale

- Anzahl der Prototypen steht von vornherein fest.
- Wie soll im Falle von mehreren möglichen Clusterzuteilungen entschieden werden?
- Warum ist gerade der Mittelwert ein guter Repräsentant für den Cluster?
- Gibt es bessere Distanzfunktionen?

Algorithmische Probleme

- Nur die Distanz zählt, nicht die Clustergröße.
- Keine Repräsentation der Clusterform
- Sog. harte Zuweisung: jeder Punkt trägt nur zu einem Cluster bei und hat dort dasselbe Gewicht.

Soft K-Means

1. Initialisierung: geg. Datenpunkte $x_n, n = 1 \ldots N$, setze K Prototypen auf Zufallswerte $m_k^{(0)}$
2. Zuweisungsschritt: Jeder Datenpunkt x_n erhält eine "weiche" Zuweisung zu allen Prototypen über die Indikatorfunktion

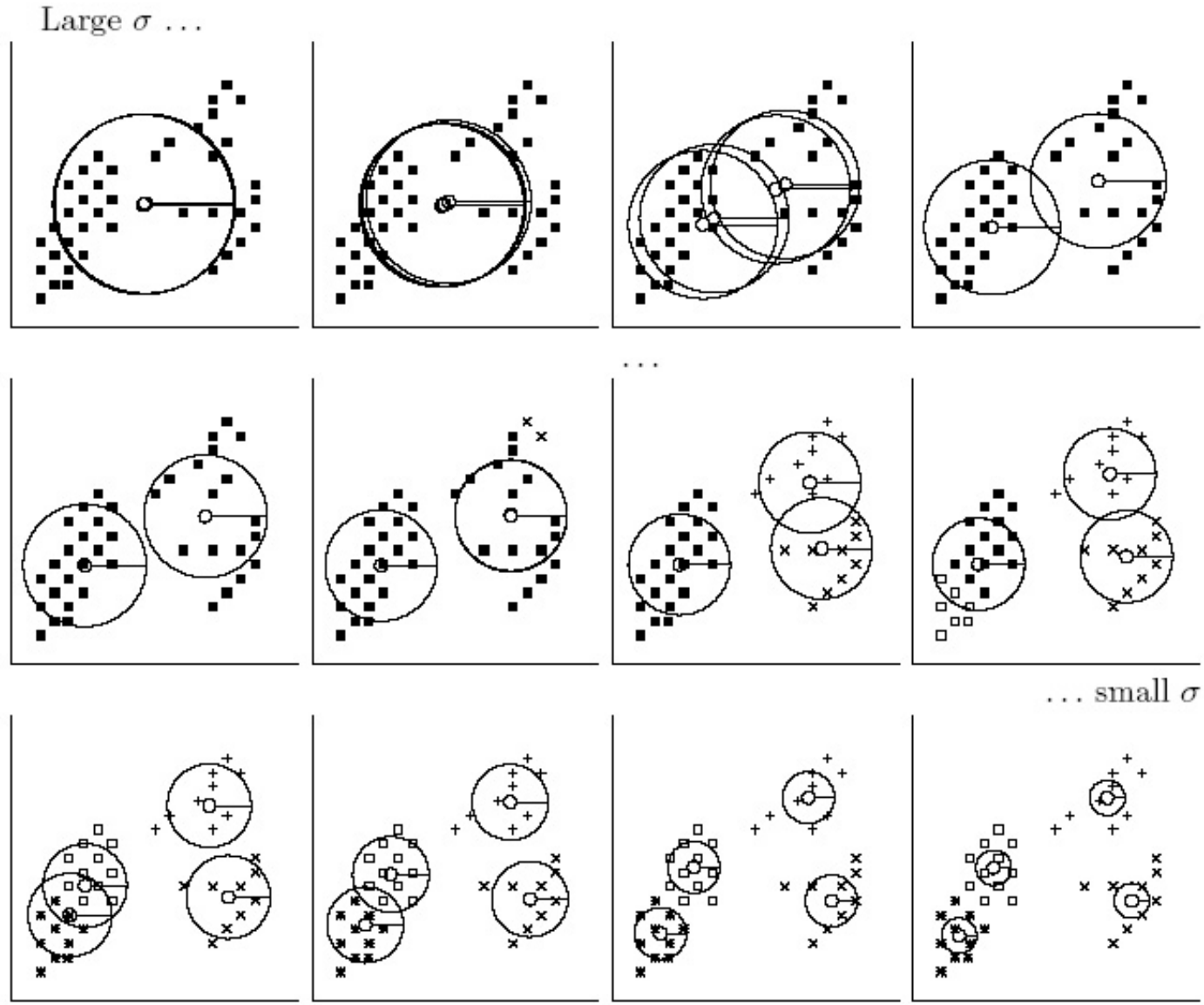
$$r_{kn} = \frac{\exp(-\beta \|m_k - x_n\|)}{\sum_i \exp(-\beta \|m_i - x_n\|)} \quad \text{mit} \quad \sum_k r_{kn} = 1$$

   β beschreibt die Reichweite der Prototypen.
3. Anpassungsschritt: Setze m_k auf den Mittelwert der ihm zugewiesenen Datenpunkte

$$m_k = \frac{\sum_n r_{kn} x_n}{\sum_n r_{kn}}$$

4. Wiederhole Zuweisungs- und Anpassungsschritt, bis sich die Zuordnung nicht mehr ändert.

Beispiel: Soft K-Means example auf 2D-Daten



[McKay 2003]