

achsenparallele Stauchung und Streckung durch Gewichte
 $w_i \geq 0$:

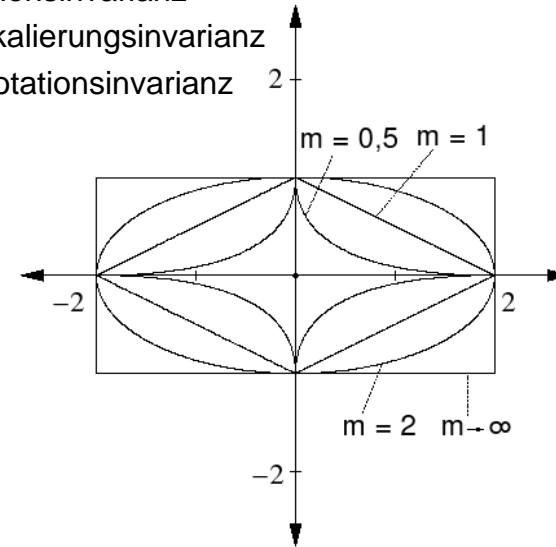
$$d_{L_m}^w : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}_0^+, d_{L_m}(p_1, p_2) \mapsto \left(\sum_{i=1}^n w_i * |p_1[i] - p_2[i]|^m \right)^{1/m}$$

Forderung:

$$\sum_{i=1}^n w_i = 1$$

Translationsinvarianz
 keine Skalierungsinvarianz
 keine Rotationsinvarianz

$w_1=0,5, w_2=1$



Matrizenschreibweise:

$$d_q(p_1, p_2) = (p_1 - p_2)^T * A * (p_1 - p_2)$$

A im n -dimensionalen Raum ist symmetrische, positiv definite Matrix $\mathbb{R}^{n \times n}$

Einheitsmatrix E : d_q identisch mit $d_{L_2}^2$

Diagonalmatrix: d_q entspricht $d_{L_2}^w$
 (Gewichte korrespondieren zu Diagonalelementen)

ansonsten: nichtuniforme Skalierung, Rotation, Spiegelung der Punkte

Symmetrische positiv definierte Matrix A <isweb>

es gilt immer:

$$\text{(Eigenwertzerlegung): } A = U * L * U^T$$

U ist orthonormale Matrix (Rotation anhand Eigenvektoren)

L ist Diagonalmatrix (Skalierung anhand Eigenwerten)

Symmetrische positiv definierte Matrix A (2) <isweb>

Berechnung der Distanz mittels $d_{L_2}^2$ auf transformierten Punkten oft relativ schnell realisierbar

$$\begin{aligned}d_q(p_1, p_2) &= (p_1 - p_2)^T A (p_1 - p_2) \\&= (p_1 - p_2)^T U L U^T (p_1 - p_2) \\&= \left(L^{1/2} U^T (p_1 - p_2) \right)^T \left(L^{1/2} U^T (p_1 - p_2) \right) \\&= \left(L^{1/2} U^T p_1 - L^{1/2} U^T p_2 \right)^T \left(L^{1/2} U^T p_1 - L^{1/2} U^T p_2 \right) \\&= d_{L_2}^2(L^{1/2} U^T p_1, L^{1/2} U^T p_2)\end{aligned}$$

Invarianzen <isweb>

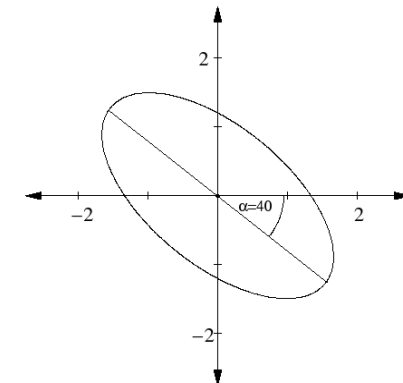
Translationsinvarianz

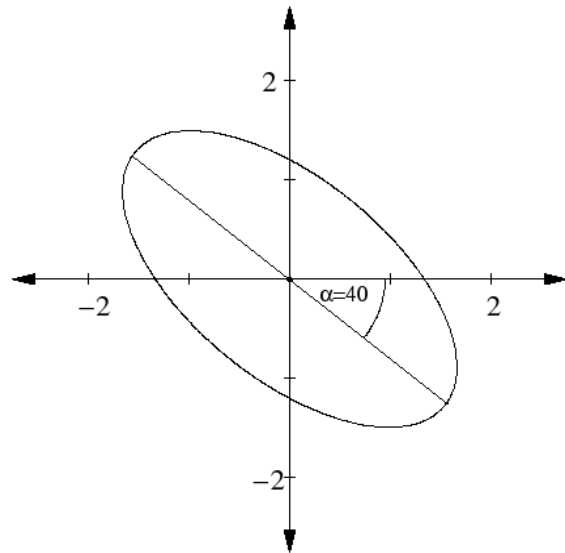
keine Skalierungsinvarianz

keine Rotationsinvarianz

Beispielmatrix <isweb>

$$\begin{aligned}A &= \begin{pmatrix} 0,5599 & 0,3693 \\ 0,3693 & 0,6901 \end{pmatrix} \\&= \begin{pmatrix} \cos 40 & \sin 40 \\ -\sin 40 & \cos 40 \end{pmatrix} * \begin{pmatrix} 0,25 & 0 \\ 0 & 1 \end{pmatrix} * \begin{pmatrix} \cos 40 & -\sin 40 \\ \sin 40 & \cos 40 \end{pmatrix}\end{aligned}$$





Einsatz der quadratischen Distanzfunktion
, wenn Distanzberechnung Kombination unterschiedlicher
Dimensionen erfordert

Grundidee kann Kovarianzmatrix auf
Dimensionen sein
→ Mahalanobis-Distanzfunktion

$$C \quad d$$

$$d_M(p_1, p_2)$$

$$d_M(p_1, p_2) = |\det C|^{1/d} (p_1 - p_2)^T * C^{-1} * (p_1 - p_2)$$

Quadratische Pseudo-Distanzfunktion

Aufgabe der Forderung nach Positiv-Definitheit für A
Ziel: *unsymmetrische Translationsinvarianz bzgl. Vektoren*
 t des Vektorraums T :

$$pd_q(p_1, p_2 + t) = pd_q(p_1, p_2)$$

Konstruktion der Matrix A aus geeigneter Orthogonalbasis und
Diagonalmatrix

Quadratische Pseudo-Distanzfunktion (2)

den U -Vektoren entsprechende Diagonalwerte von L auf Null
setzen

seien s_i mit $i = 1, \dots, m$ die durch l_i auf Null gesetzten
 U -Spaltenvektoren, dann gilt für Linearkombinationen hiervon:

$$T = \left\{ t \in \mathbb{R}^n \mid t = \sum_{i=1}^m \lambda_i * s_i : \lambda_i \in \mathbb{R} \right\}$$

$$\begin{aligned}
 &pd_q(p_1, p_2 + t) \\
 = &(p_1 - p_2 - t)^T A (p_1 - p_2 - t) \\
 = &(p_1 - p_2 - t)^T U L U^T (p_1 - p_2 - t) \\
 = &(p_1 - p_2 - t)^T U L^{1/2} L^{1/2} U^T (p_1 - p_2 - t) \\
 = &\left(L^{1/2} U^T (p_1 - p_2 - t) \right)^T \left(L^{1/2} U^T (p_1 - p_2 - t) \right) \\
 = &\left(L^{1/2} U^T (p_1 - p_2) - L^{1/2} U^T t \right)^T \left(L^{1/2} U^T p_1 - p_2 - L^{1/2} U^T t \right) \\
 = &\left(L^{1/2} U^T (p_1 - p_2) \right)^T \left(L^{1/2} U^T p_1 - p_2 \right) \\
 = &pd_q(p_1, p_2)
 \end{aligned}$$

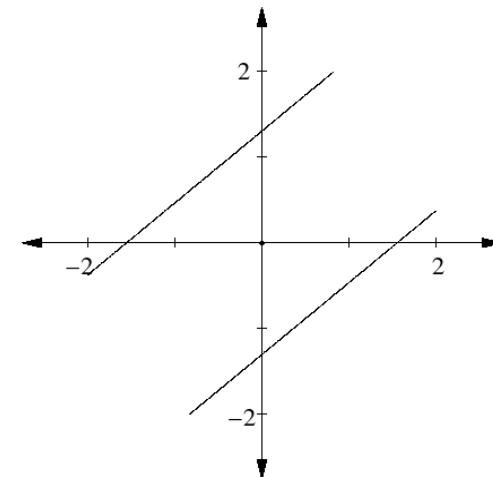
Der entscheidende Schritt: laut Def ist dieser Term 0

Konstruktion Translationsinvarianz im Winkel von 40 Grad:

$$U = \begin{pmatrix} \cos 40 & -\sin 40 \\ \sin 40 & \cos 40 \end{pmatrix} \\
 L = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$$

Die Kombination dieser Matrizen ergibt die gewünschte Matrix A:

$$U * L * U^T = \begin{pmatrix} 0,4132 & -0,4924 \\ -0,4924 & 0,5868 \end{pmatrix}$$



folgende Beobachtungen Chang/Wu03 bzgl. Unähnlichkeit im hochdimensionalen Raum:

- ♦ ähnliche Objekte liegen meist nur in wenigen Dimensionen nebeneinander
- ♦ Ähnlichkeit kann häufig nicht an bestimmten Dimensionen festgemacht werden

Problem mit Minkowski-Distanzfunktion: alle Dimensionen werden berücksichtigt

Berücksichtigung einer dynamischen Untermenge der Dimensionen

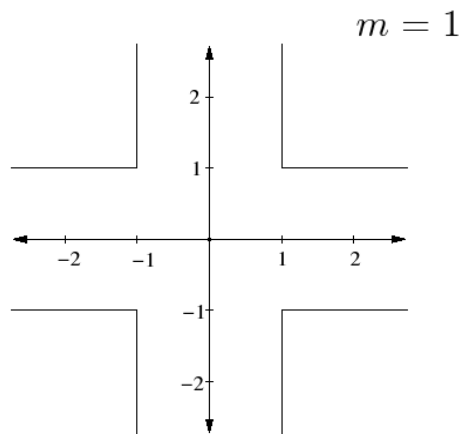
p_1 und p_2 seien zwei Punkte im n -dimensionalen Raum und $\delta_i = |p_1[i] - p_2[i]|$ der Abstand in Dimension i
nur die m kleinsten Abstände werden berücksichtigt:

$\Delta_m = \{\text{die kleinsten } m \text{ } \delta\text{-Werte aus } (\delta_1, \delta_2, \dots, \delta_n)\}$

$$d_{dp}^{m,r} = \left(\sum_{\delta_i \in \Delta_m} \delta_i^r \right)^{\frac{1}{r}}$$

Selbstidentität und Symmetrie sind erfüllt
Verletzung der Positivität und Dreiecksungleichung

zweidimensionaler Raum und



Abstand zwischen Histogrammen mit absoluten Häufigkeiten

ursprünglich in Statistik entwickelt

Untersuchung von Abhängigkeit zwischen Zufallsvariablen

basiert auf Nullhypothese: Häufigkeitsverteilungen sind gleich

also Differenz zwischen erwarteten und tatsächlichen Häufigkeiten beträgt 0

Chi-Quadrat-Semi-Pseudo-Distanzfunktion (2) <isweb>

$$spd_{\chi^2}(p_1, p_2) = \sum_{j=1}^n \frac{(p_1[j] - \hat{p}_1[j])^2}{\hat{p}_1[j]} + \sum_{j=1}^n \frac{(p_2[j] - \hat{p}_2[j])^2}{\hat{p}_2[j]} \text{ für } p_1, p_2 \in \mathbb{N}_0^n$$

erwartete Häufigkeiten:

$$\hat{p}_i[j] = \frac{(p_1[j] + p_2[j]) * \sum_{a=1}^n p_i[a]}{\sum_{a=1}^n (p_1[a] + p_2[a])}$$

Beispiel <isweb>

Test, ob Grippeimpfung Grippe verhindern kann

Befragung verschiedener Personen über Auftreten von Grippe und Impfungen

erwartete Werte sind in Klammern notiert

	keine Impfung	eine Impfung	Doppelimpfung	Σ
Grippe	24 (14,398)	9 (5,014)	13 (26,588)	46
keine Grippe	289 (298,602)	100 (103,986)	565 (551,412)	954
Σ	313	109	578	1000

wenn kein Zusammenhang zwischen Impfung und Gruppe,
dann Wert jeder Zelle abschätzbar

Beispiel Grippe(=j=1)/keine Impfung

(=i=1);
vgl. keine Impfung p_1 mit
Doppelimpfung p_2

Häufigkeit für Grippe ist
 $24+13 = 37 = p_1[j] + p_2[j]$

Wahrsch. für keine Impfung ist
 $313 = p_1[1] + p_1[2] = \sum_{a=1}^n p_1[a]$

Multiplizierte Häufigkeit für
Grippe/keine Impfung ist
 $37 \cdot 313 \sim$ Zähler

Nenner: $24+289+13+565$
 $= 313+565=878$

erwartete Häufigkeit:
 $37 \cdot 313 / 878 = 13,19$

Erwartete Wahrscheinlichkeit:
 $p_i[j] = 13,19 / 878$

$$\hat{p}_i[j] = \frac{(p_1[j] + p_2[j]) * \sum_{a=1}^n p_i[a]}{\sum_{a=1}^n (p_1[a] + p_2[a])}$$

Bemerkung: die Rechnung im Buch
berücksichtigt die drei (!)
Zufallsvariablen:

1. Keine Impfung
2. Eine Impfung
3. Doppelimpfung

Selbstidentität und Symmetrie sind erfüllt
Rotationsinvarianz
keine Positivität
keine Dreiecksungleichung

Abstand 0,1
um (1; 0,5)

