

Multivariate Analysemethoden

Vorlesung

Multivariate Distanz – Multivariate Normalverteilung
Minimum Distance Classifier – Bayes Classifier

Günter Meinhardt
Johannes Gutenberg Universität Mainz

Multivariate Klassifikation

Ziele

- Einordnen von Fällen (Versuchspersonen, Beobachtungen) in Gruppen aufgrund ihrer Werte in **mehreren** Meßvariablen.
- Maßgeblich für die Zuordnung zu eine Gruppe ist a) die **Wahrscheinlichkeit** des Auftretens des Falles in der Zielgruppe (falls ermittelbar) oder b) die **Distanz** des Falles vom charakteristischen Wert der Gruppe (Prototyp, Zentroid)

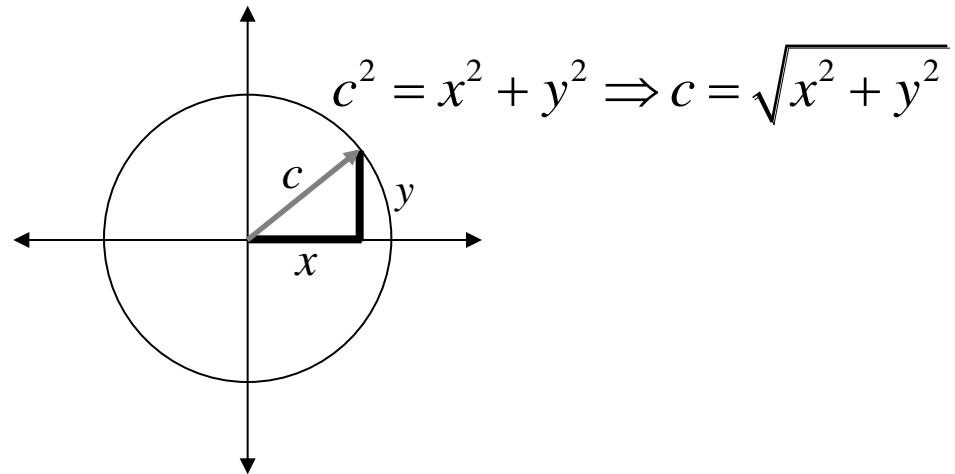
Methoden

- Deskriptive Methoden:
 - * Bestimmung von Distanzen und Wahrscheinlichkeiten auf dem **Set der beobachteten Meßvariablen**
- Analytische Methoden:
 - * Bestimmung von Distanzen und Wahrscheinlichkeiten auf **transformierten Meßvariablen** mit dem Ziel, die Separation von Gruppen zu maximieren (Diskriminanzanalytische Methoden)
- Weitere Kriterien sind Kosten von Fehlklassifikationen und die a-priori Wahrscheinlichkeit von Gruppen (**Allg. Likelihood-Ratio und Bayes-Klassifikation**)

Kreis

Iso-Distanz-Konturen in 2D

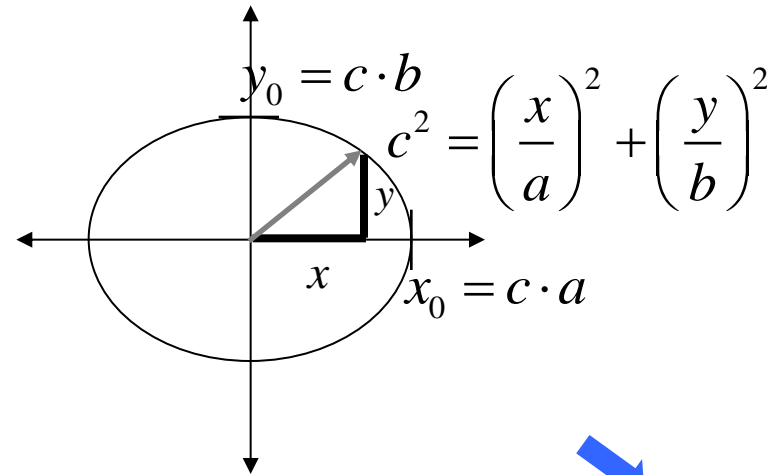
Kreis mit Radius c : Alle Punkte auf dem Kreisbogen haben euklidischen Abstand c zum Kreismittelpunkt



- Der Kreis ist die Grundform der Iso-Distanz Kontur im zweidimensionalen Raum ($p = 2$).
- Er entspricht im Variablenraum einer Iso-Distanz-Kontur für 2 unkorrelierte (orthogonale) Variablen mit derselben Skalierung.

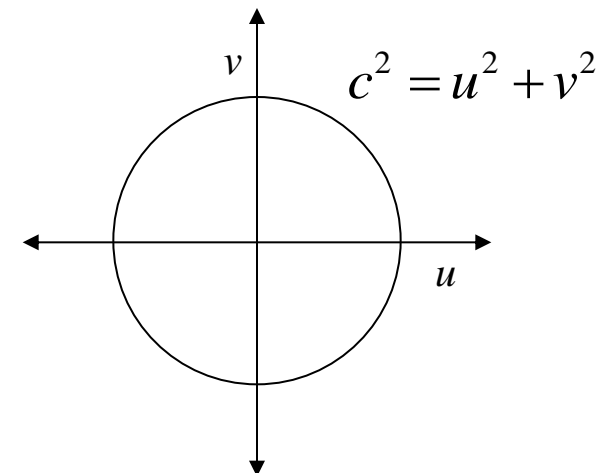
Ellipse: Skalierung

Ellipse mit Ellipsenradius c : Alle Punkte auf dem Ellipsenbogen haben, auf **Standardskala** normiert, denselben Abstand c zum Mittelpunkt



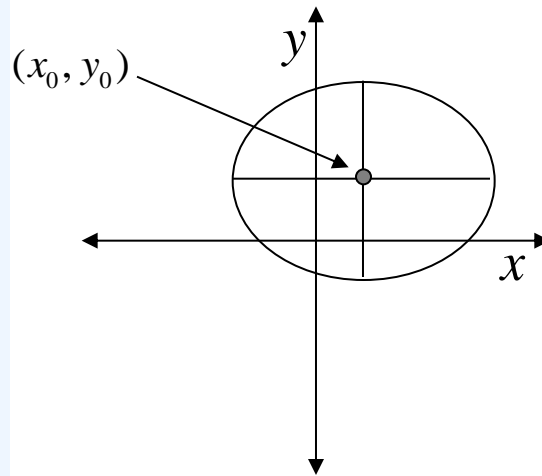
Standardskala:

$$u = \frac{x}{a} \quad v = \frac{y}{b}$$



Ellipse Translation

Translation zum Punkt (x_0, y_0) ändert an dieser Eigenschaft nichts:

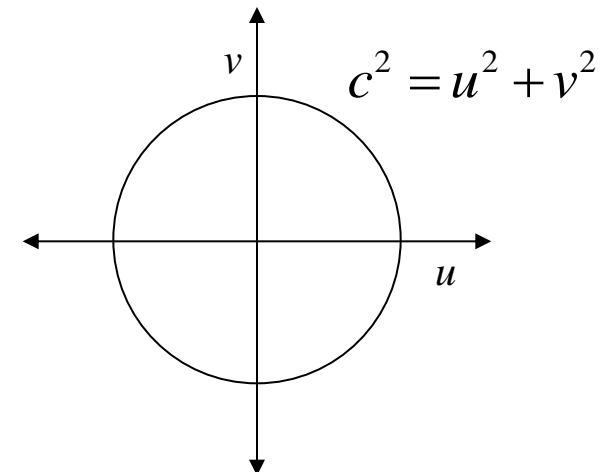


$$c^2 = \left(\frac{x - x_0}{a} \right)^2 + \left(\frac{y - y_0}{b} \right)^2$$



Standardskala:

$$u = \frac{x - x_0}{a} \quad v = \frac{y - y_0}{b}$$

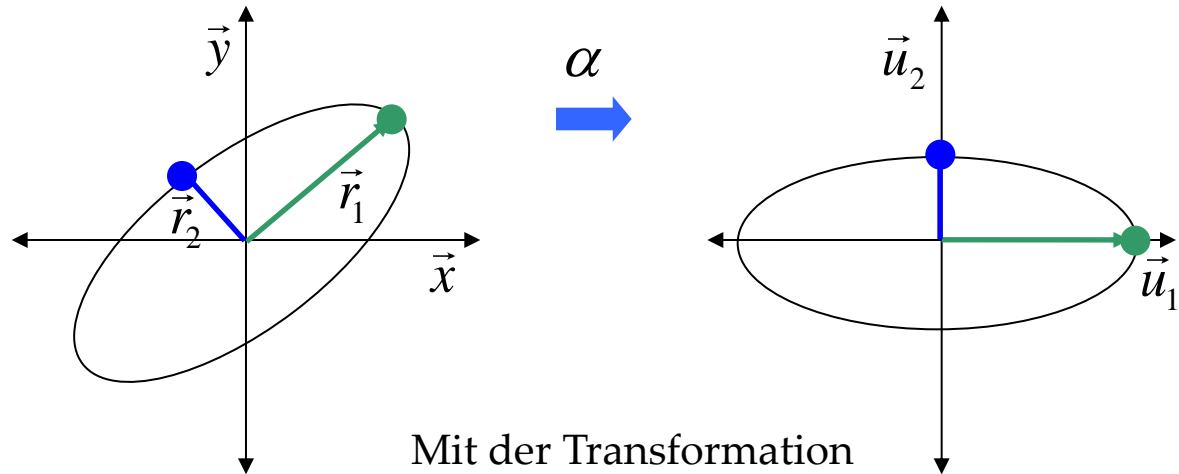


Standard- Transformation

Standard-Ellipse
 Neigung
 Korrelation ρ

Die Invarianz der Distanz im neuen Koordinatensystem mit geneigten Achsen (Korrelation der Variablen) ist über eine Rotation der Koordinaten (anticlock) erklärt:

$$c^2 = x^2 + y^2 - 2\rho xy \quad \rho = \cos \alpha$$



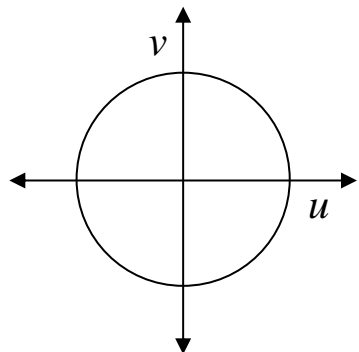
Mit der Transformation

$$u = \frac{u_1}{a} \quad v = \frac{u_2}{b}$$

Koordinaten
Korrelierte
Achsen

$$c^2 = u^2 + v^2$$

[Tafel: $\cos \alpha$]



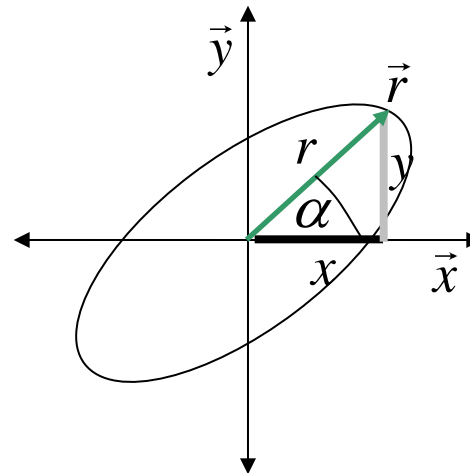
erfüllen alle Ellipsenpunkte:

$$c^2 = u^2 + v^2$$

Standard-Ellipse: Zeichen-Routine

Ellipsen sind in **kartesischen Koordinaten** unpraktisch zu zeichnen. Man geht über zur Darstellung in **Polarkoordinaten**.

$$c^2 = x^2 + y^2 - 2\rho xy$$



kartesisch

polar

$$r = (x, y) \triangleq (r, \alpha)$$

Es gelten die Transformationen:

polar \rightarrow kartesisch

$$\begin{aligned} x &= r \cos \alpha \\ y &= r \sin \alpha \end{aligned}$$

kartesisch \rightarrow polar

$$\begin{aligned} r^2 &= x^2 + y^2 \\ \alpha &= \tan^{-1} \left(\frac{y}{x} \right) \end{aligned}$$

Zum Zeichnen muß die Ellipsengleichung als Gleichung in Polarkoordinaten (Vektorlänge in Abhängigkeit des Winkels α) umgeschrieben werden

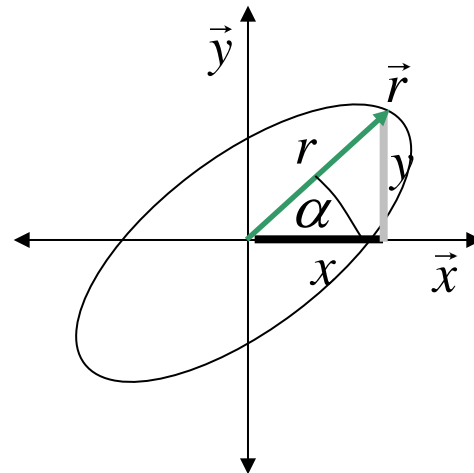
Standard-Ellipse: Zeichen-Routine

Von der Darstellung in Polarkoordinaten kann einfach in kartesische Koordinaten zurückgerechnet werden (Setzen der Ellipsenpunkte)

$$c^2 = x^2 + y^2 - 2\rho xy$$

Setze $q = \frac{y}{x}$

damit $c^2 = x^2(1 + q^2 - 2\rho q)$



$$|x| = \frac{c}{\sqrt{1 + q^2 - 2\rho q}}$$

$$r = \sqrt{x^2 + y^2} = |x| \sqrt{(1 + q^2)}$$

Verfahren



[Excel-Sheet]

1. Variiere α von $-\pi$ bis π (= ein Kreisumlauf).
2. Für jeden Winkel α berechne $q = \tan^{-1}(\alpha)$.
3. Berechne dann $|x|$
4. Berechne damit r .
5. Berechne dann x, y : $x = r \cos \alpha$
 $y = r \sin \alpha$

1 D-Normal Verteilung

Die Funktion $f(x) = e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ hat Fläche $\sqrt{2\pi}\sigma$

Die auf die Fläche 1 normierte Funktion

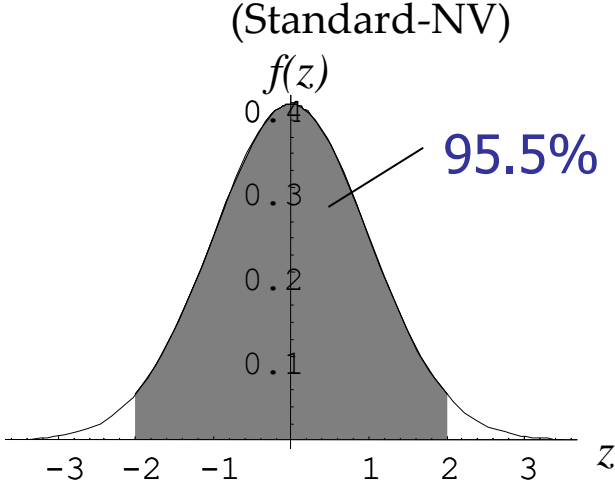
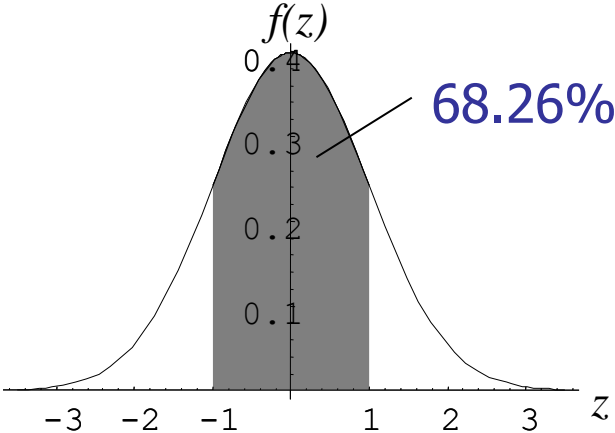
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$z = \frac{x - \mu}{\sigma}$$



$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

heißt **Normalverteilung** (Gauss-Verteilung).
 Mit ihr sind Wahrscheinlichkeiten als Flächen-
 Anteile für z - Standardvariablen definierbar.



p-variater Fall

Man bemerke daß $z^2 = \left(\frac{x - \mu}{\sigma} \right)^2 = (x - \mu) \frac{1}{\sigma^2} (x - \mu)$ ist.

Man habe nun nicht eine, sondern p Variablen:

$$\vec{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} \text{ mit Zentroid } \vec{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix} \quad (\text{jeder Messpunkt ist ein } p\text{-dimensionaler Vektor und der Zentroid ist ein } p\text{-dimensionaler Vektor})$$

Dann definiert

$$\Delta^2 = (\vec{x} - \vec{\mu})^t \Sigma^{-1} (\vec{x} - \vec{\mu}) \quad \text{mit } \Sigma^{-1} \text{ die Inverse der Varianz-Kovarianz Matrix } \Sigma.$$

die verallgemeinerte quadrierte Distanz im multivariaten Raum. Sie heißt quadrierte **Mahalanobis-Distanz**.

[Excel-Beispiel 2D]

$$\vec{x} = \begin{pmatrix} 3 \\ 2 \end{pmatrix} \quad \vec{\mu} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad \Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} \quad \Delta^2 = 4$$

p D-Normal Verteilung

Die Funktion $f(\vec{x}) = e^{-\frac{1}{2}(\vec{x}-\vec{\mu})' \Sigma^{-1}(\vec{x}-\vec{\mu})}$ hat Volumen $(2\pi)^{p/2} |\Sigma|^{1/2}$

Die auf Volumen 1 normierte Funktion

$$f(\vec{x}) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu})' \Sigma^{-1}(\vec{x}-\vec{\mu})}$$

heißt **multivariate Normalverteilung** (multivariate Gauss-Verteilung). Mit ihr sind Wahrscheinlichkeiten als **Anteile des Gesamtvolumens** eines p-dimensionalen Ellipsoids definiert.

Die in ihrem Argument auftretende Mahalanobis-Distanz erfüllt die Bedingung:

$$\Delta^2 = (\vec{x} - \vec{\mu})' \Sigma^{-1} (\vec{x} - \vec{\mu}) \leq \chi_p^2(\alpha)$$

mit α einem zu setzenden alpha-Fehler Niveau.

Alle Mahalanobisdistanzen Δ , die diese Bedingung erfüllen, erzeugen Konturen gleicher Wahrscheinlichkeit (**iso-probability contours**) mit $P = 1 - \alpha$ in der multivariaten Normalverteilung.

2 D-Normal Verteilung

Die multivariate Normalverteilung mit $p = 2$ Variablen (bivariate Normalverteilung) hat die Form

$$f(x_1, x_2) = \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2(1-\rho^2)}} e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right)\right]}$$

Die im Argument auftretende Mahalanobis-Distanz definiert eine Ellipse im zweidimensionalen Raum für jede Konstante c :

$$\begin{aligned} c^2 = \Delta^2 &= (\vec{x} - \vec{\mu})^t \mathbf{\Sigma}^{-1} (\vec{x} - \vec{\mu}) \\ &= \frac{1}{(1-\rho^2)} \left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) \right] \end{aligned}$$

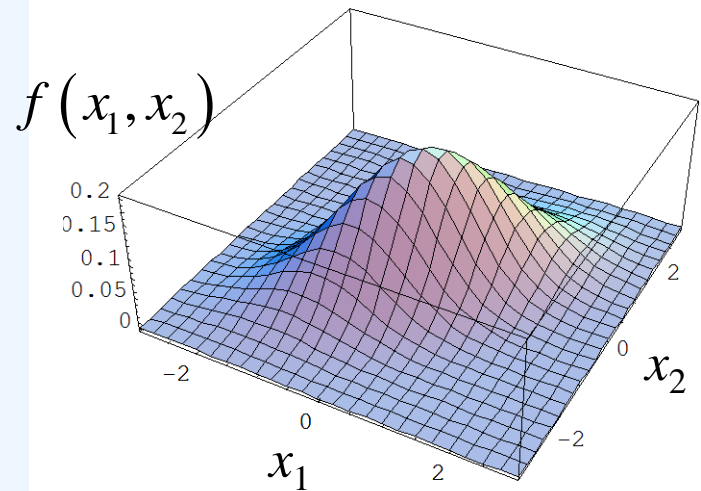
Diese ist eine Iso-Probability-Contour im obigen Sinne (s. multivariate NV, vorherige Folie)

2 D-Normal Verteilung

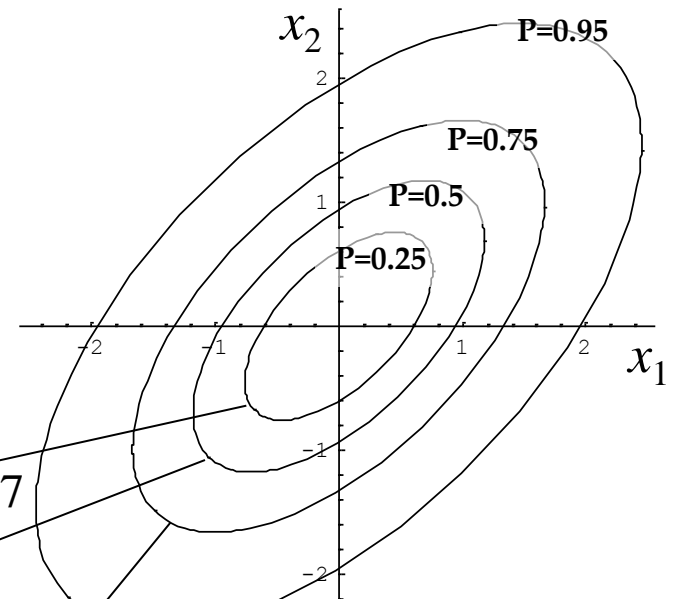
Bivariate Normalverteilung mit $p = 2$ Variablen und Korrelation $r = 0.6$

Density-Plot

$$\vec{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$



Contour-Plot



$$\Delta^2 = \chi^2_2(0.75) = 0.57$$

$$\Delta^2 = \chi^2_2(0.5) = 1.38$$

$$\Delta^2 = \chi^2_2(0.05) = 5.99$$

$$\Delta^2 = \chi^2_2(0.25) = 2.77$$

Ellipsen gleicher Wahrscheinlichkeit und zugehöriges Distanzmaß (quadrierte Mahalanobis-Distanz)

NV-2D- Ellipse: Zeichen- Routine

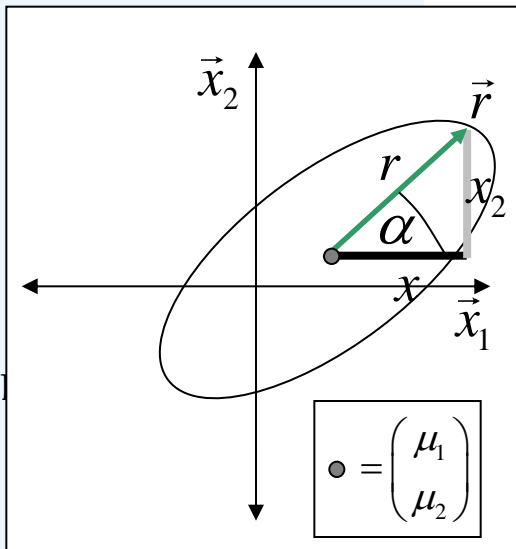
$$c^2 = \frac{1}{(1-\rho^2)} \left[\left(\frac{x_1 - \mu_1}{\sigma_1} \right)^2 + \left(\frac{x_2 - \mu_2}{\sigma_2} \right)^2 - 2\rho \left(\frac{x_1 - \mu_1}{\sigma_1} \right) \left(\frac{x_2 - \mu_2}{\sigma_2} \right) \right] \quad \text{(NV-Ellipse)}$$

Setze $q = \frac{y}{x}$ und temporär $\mu_1 = 0, \mu_2 = 0$

$$c^2 = \frac{1}{1-\rho^2} \left(\left(\frac{x}{\sigma_1} \right)^2 + \left(\frac{qx}{\sigma_2} \right)^2 - 2\rho \frac{x}{\sigma_1} \frac{qx}{\sigma_2} \right)$$

$$c^2 = \frac{1}{1-\rho^2} x^2 \left(\frac{1}{\sigma_1^2} + \frac{q^2}{\sigma_2^2} - 2 \frac{\rho q}{\sigma_1 \sigma_2} \right)$$

$$|x| = \frac{c}{\sqrt{\frac{1}{1-\rho^2} \left(\frac{1}{\sigma_1^2} + \frac{q^2}{\sigma_2^2} - 2 \frac{\rho q}{\sigma_1 \sigma_2} \right)}}$$



Und es gilt:

a) α läuft von $-\pi$ bis π (= ein Kreisumlauf)

b) $r = \sqrt{x_1^2 + x_2^2} = |x| \sqrt{1+q^2}$ c) $x_1 = r \cos \alpha + \mu_1$
 $x_2 = r \sin \alpha + \mu_2$

Verfahren

[Excel-Sheet]

p D-Normal Verteilung

Die Ellipsen der Form

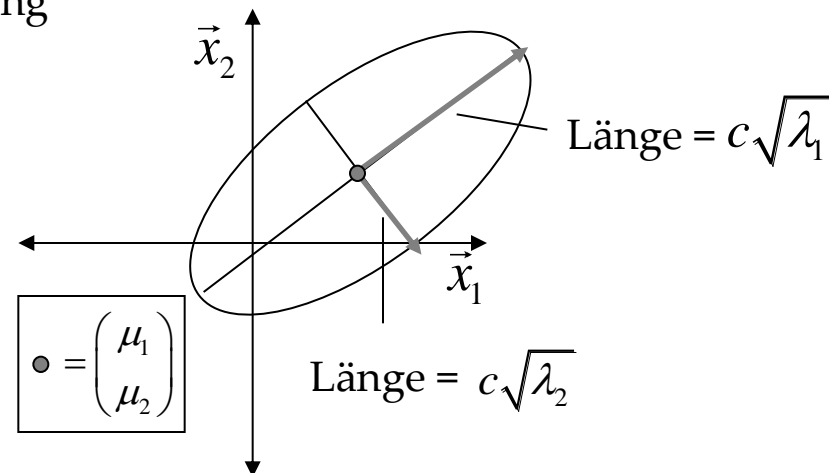
$$c^2 = (\vec{x} - \vec{\mu})^t \mathbf{\Sigma}^{-1} (\vec{x} - \vec{\mu}) \leq \chi_p^2(\alpha)$$

sind zentriert in $\vec{\mu}$ und haben Hauptachsen $\pm c\sqrt{\lambda_i}\vec{e}_i$
mit Eigenwertbedingung

$$\mathbf{\Sigma}\vec{e}_i = \lambda_i\vec{e}_i$$

Eine Eigenwertzerlegung der Varianz-Kovarianz Matrix liefert somit die Hauptachsen des p-variaten Ellipsoids der multivariaten Normalverteilung

Beispiel 2D



MDC

Mit der Mahalanobisdistanz für eine Beobachtung \vec{x} zum Zentroid der Gruppe c_j

$$\Delta_j^2 = (\vec{x} - \vec{\mu}_j)^t \Sigma_j^{-1} (\vec{x} - \vec{\mu}_j)$$

definiere die Regel:

Gruppiere \vec{x} in Gruppe c_i , wenn gilt

MDC-Regel

$$\Delta_i^2 = \min(\Delta_1^2, \Delta_2^2, \dots, \Delta_j^2, \dots, \Delta_k^2)$$

Die Performance des MDC lässt sich mit großen Stichproben für die k - Gruppen mit einer Konfusions-Matrix bewerten:

Häufigkeit zur
Einordnung
von Fall (Zeile)
in Gruppe
(Spalte)

allocated to group

	c_1	c_2	...	c_j	...	c_k
c_1	h_{11}	h_{12}	...	h_{1j}	...	h_{1k}
c_2	h_{21}	h_{22}	...	h_{2j}	...	h_{2k}
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots	\vdots
c_j	h_{j1}	h_{j2}	...	h_{jj}	...	h_{jk}
\vdots	\vdots	\vdots	\vdots	\vdots	\ddots	\vdots
c_k	h_{k1}	h_{k1}	...	h_{kj}	...	h_{kk}

Case \vec{x} is group

Confusion- matrix

	c_1	c_2	...	c_j	...	c_k
c_1	h_{11}	h_{12}	...	h_{1j}	...	h_{1k}
c_2	h_{21}	h_{22}	...	h_{2j}	...	h_{2k}
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots	\vdots
c_j	h_{j1}	h_{j2}	...	h_{jj}	...	h_{jk}
\vdots	\vdots	\vdots	\vdots	\vdots	\ddots	\vdots
c_k	h_{k1}	h_{k2}	...	h_{kj}	...	h_{kk}

Hits

Korrekte Klassifizierungen sind die Häufigkeiten auf der Diagonalen:

$$h_o = \sum_{i=1}^k h_{ii}$$

Mit den Zeilensummen h_i und N der Summe aller Häufigkeiten gilt

$$\hat{h}_{ij} = p_j \cdot h_i \quad (\text{erwartete Zellhäufigkeit})$$

mit p_j der A-priori Wahrscheinlichkeit der Gruppe c_j

p_j kann ggf. aus den empirischen Gruppenstärken über $p_j = h_i/N$ geschätzt werden, wenn keine Information über die A-priori Wahrscheinlichkeiten vorliegt.

Erwartete
Häufigkeiten
bei Zufall
(anteilige
Gleichverteilung)

**Erwartete
Confusion-
matrix**

Dann ist

$$\hat{h} = \sum_{i=1}^k \hat{h}_{ii}$$

die erwartete Hit-Häufigkeit.

Mit

$$p_e = \frac{\hat{h}}{N} \quad \sigma^2 = N \cdot p_e \cdot (1 - p_e)$$

ist h_o normalverteilt über die Approximation der Binomialverteilung

wenn $N \cdot p_e \cdot (1 - p_e) > 9$ gilt.

Dann testet der z- Test

$$z = \frac{h_o - Np_e}{\sqrt{N \cdot p_e \cdot (1 - p_e)}}$$

die Hitrate des MDC gegen den Zufall.

	c_1	c_2	...	c_j	...	c_k
c_1	\hat{h}_{11}	\hat{h}_{12}	...	\hat{h}_{1j}	...	\hat{h}_{1k}
c_2	\hat{h}_{21}	\hat{h}_{22}	...	\hat{h}_{2j}	...	\hat{h}_{2k}
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots	\vdots
c_i	\hat{h}_{i1}	\hat{h}_{i2}	...	\hat{h}_{ij}	...	\hat{h}_{ik}
\vdots	\vdots	\vdots	\vdots	\vdots	\ddots	\vdots
c_k	\hat{h}_{k1}	\hat{h}_{k2}	...	\hat{h}_{kj}	...	\hat{h}_{kk}

Hits

**Erwartete
Häufigkeiten
bei Zufall
(anteilige
Gleichverteilung)**

A-priori Wahrscheinlichkeit der Gruppen

Man habe Information über die A-priori Wahrscheinlichkeiten der Gruppen c_j :

$$P(c_1), P(c_2), \dots, P(c_j), \dots, P(c_k)$$

Dann liefert eine Klassifikation der Beobachtung \vec{x} nach ihrer A-posteriori Wahrscheinlichkeit

$$P(c_j | \vec{x})$$

eine korrektere Zuordnung als nur nach der kürzesten Distanz zum Gruppenzentrum.

Regel:

Gruppiere \vec{x} in Gruppe c_i , wenn gilt

$$P(c_i | \vec{x}) = \max \left(P(c_1 | \vec{x}), P(c_2 | \vec{x}), \dots, P(c_j | \vec{x}), \dots, P(c_k | \vec{x}) \right)$$

Max-Aposteriori WKn Classifier

Normalverteilungsannahme

Um die A-posteriori WKn zu berechnen, muss für die Likelihood-Funktionen die Annahme der multivariaten Normalverteilung gelten.

Likelihoods

Mit der multivariaten Normalverteilung haben die Likelihoods die Form

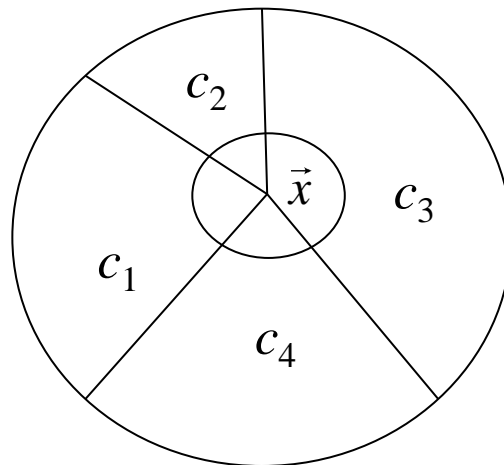
$$P(\vec{x}|c_j) = f(\vec{x}|c_j) = \frac{1}{(2\pi)^{p/2} |\Sigma_j|^{1/2}} e^{-\frac{1}{2}\Delta_j^2}$$

A-posteriori WK

mit
$$\Delta_j^2 = (\vec{x} - \vec{\mu}_j)^t \Sigma_j^{-1} (\vec{x} - \vec{\mu}_j)$$

der quadrierten Mahalanobisdistanz zum Gruppenzentrum $\vec{\mu}_j$

Klassifikations-Raum



Der Klassifikationsraum ist durch alle Gruppen **vollständig partitioniert**.

Es gilt:

$$\vec{x} = (\vec{x} \cap c_1) \cup (\vec{x} \cap c_2) \cup \dots \cup (\vec{x} \cap c_k)$$

Und wegen der Disjunktheit:

$$P(\vec{x}) = P(\vec{x} \cap c_1) + P(\vec{x} \cap c_2) + \dots + P(\vec{x} \cap c_k)$$

Normalverteilungsannahme

Likelihoods

Da

$$P(\vec{x}|c_j) = \frac{P(\vec{x} \cap c_j)}{P(c_j)}$$

(Def. der bedingten Wahrscheinlichkeit), folgt

Satz der totalen WK

$$P(\vec{x}) = P(\vec{x}|c_1)P(c_1) + P(\vec{x}|c_2)P(c_2) + \dots + P(\vec{x}|c_k)P(c_k)$$

Und damit

Satz von Bayes

$$P(c_i|\vec{x}) = \frac{P(\vec{x}|c_i)P(c_i)}{\sum_j P(\vec{x}|c_j)P(c_j)}$$

der Satz von Bayes für die A-posteriori WK der Gruppe c_i , gegeben die multivariate Beobachtung

Normalverteilungsannahme

Die approximative Gültigkeit der multivariaten NV kann durch Q-Q-Plot Methoden überprüft werden.