# Multivariate Analysemethoden

Vorlesung

*Günter Meinhardt*
*Johannes Gutenberg Universität Mainz*

**Vorlesung**

Verfahrensdarstellung in

- Überblick
- Grundprinzip
- wichtigsten mathematischen Beziehungen
- Anwendungsbeispielen
- Durchführung mit Excel und Statistica

**Übung / Tut**

Vermittlung von Hintergründen/Voraussetzungen

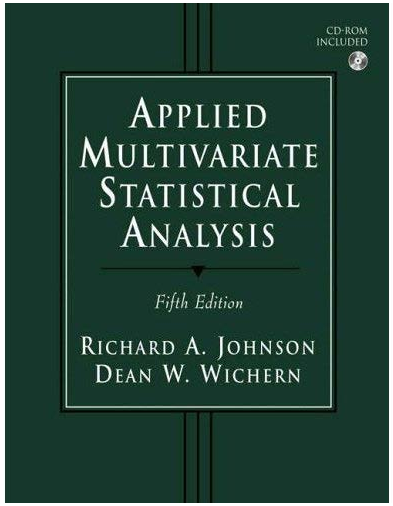
- Grundlagen (Vektoren/Matrizen)
- Wiederholung / Durcharbeiten der Beispiele
- Aufgaben und Anwendungen auf verwandte Probleme

**Prüfung**

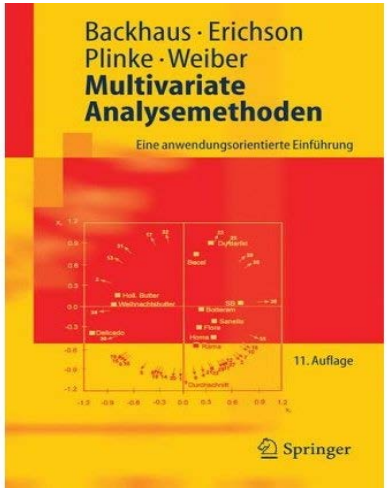
- Klausur zum Abschluss des Moduls gemeinsam mit Testtheorie

## Literatur

a)




b)




c)


d)

**Inhalte im WS 2017/18**

# Grundlagen

**Vektoren / Matrizen**

**Multivariate Distanz**

**Eigenwertzerlegung**

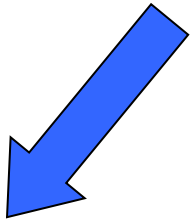
**Inhalte im WS 2017/18**

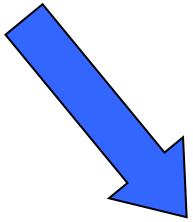
## Verfahren

- MDS
- Multiple Regression
- Faktorenanalyse
- Kanonische Korrelation
- Diskrimimanzanalyse
- Multivariate Klassifikation
- Statistisches Entscheiden

**Einteilung**

# Multivariate Analysemethoden





## Latente Variable

- Faktorenanalyse
- Diskriminanzanalyse
- MDS
- Strukturgleichungen
- Kanonische Korrelation
- ⋮

## Konkrete Variable

- Multiple/Logistische Regression
- $T^2$ / MANOVA
- Conjoint Measurement
- Kanonische Korrelation
- ⋮

**Latente Variable**

# Multidimensionale Skalierung

Problem:

Positionierung von Messobjekten in einem latenten Raum (hier: Wahrnehmungsraum)

Möglichkeiten:

**Faktorenanalyse** ⟺ **Multidimensionale Skalierung**

**Latente Variable**

# Faktor / MDS

**Demo - Beispiel mit Excel und Statistica**

## Multivariates Testen

**Grundüberlegungen zum Unterschied des Testens mit einer AV und mehreren AVs**

**Grundprinzip und Beispiel anhand einer 2 Vars – 2 Groups Diskriminanzanalyse**

# Mittelwertsprüfung bei mehreren Variablen

## Beispiel

## 10 Variablen

## 2 Gruppen

### Lebenszufriedenheit

**Arbeit**

$X_1$: Gehalt
$X_2$: Entscheidungsfreiheit
$X_3$: Qualität der Kommunikation

**Privatsphäre**

$X_4$: Ehe
$X_5$: Freunde/Beziehungen
$X_6$: Sexualität

**Person**

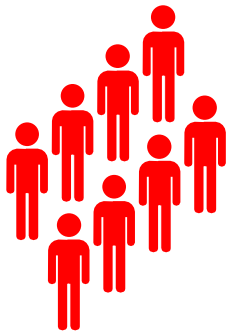
$X_7$: Lebensansprüche
$X_8$: Sinnhaftigkeit

**Aktivität**

$X_9$: Hobbies
$X_{10}$: Sport/Fitness

$$(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{10})$$

Gesunde



Herzinfarktpatienten

# Multivariate Mittelwertsvergleiche - Einzeltestungen

**Frage**

Unterscheiden sich Gesunde und Patienten im Variablen-komplex **Lebenszufriedenheit**?

**Teststrategie**

Wir testen auf jeder der 10 Skalen den Gruppenunterschied mit einem t- Test. Wenn **irgend einer** der Tests signifikant wird, sehen wir die Gruppen als verschieden an.

**Probleme**

1. **Multiples Testen:** Dieselbe Hypothese wird 10 mal geprüft.
2. **Unterstellte Unabhängigkeit:** Man behandelt die einzelnen Skalen als unabhängig voneinander.
3. **Fehlendes Konstrukt:** Lebenszufriendenheit wird nicht als Variablenkomplex mit Binnenstruktur behandelt.
4. **Mangelnde Teststärke:** Man nutzt nicht die Korrelations-struktur der Variablen für einen leistungsfähigen Test.

**Ausweg**

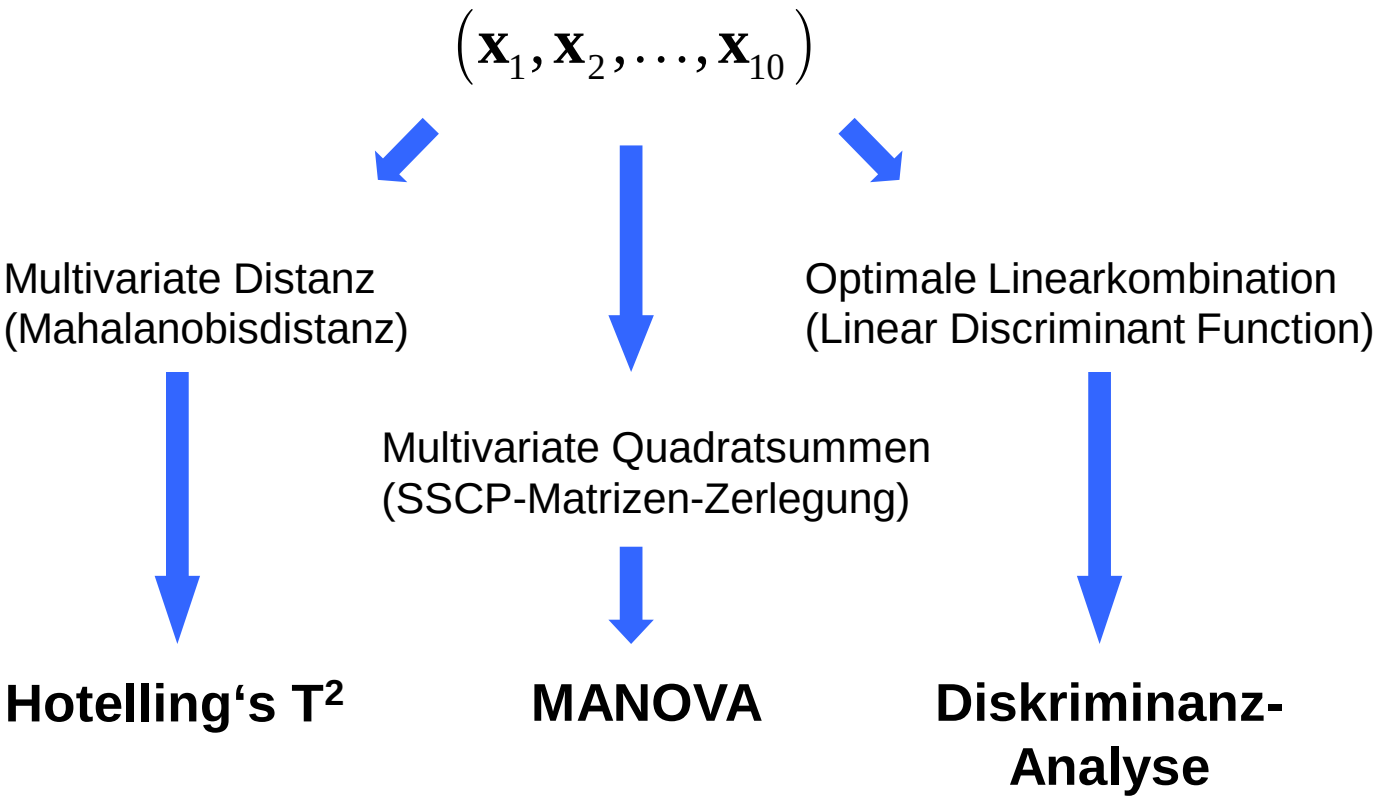
Verwendung **eines** multivariaten Tests, der die Information **aller** 10 Variablen und ihrer Korrelationsstruktur in **eine** statistische Prüfgrösse einfliessen lässt.

# Multivariate Mittelwertsvergleiche - Verfahren

**Variablen-komplex**

**Multivariates Testkonstrukt**

**Verfahren**




Alle Verfahren entscheiden über den Gruppenunterschied im **gesamten Variablenkomplex** mit **einem** statistischen Test

# Multivariates Testen - Diskriminanzanalyse

## Grundprinzip (2 Gruppen)

Für die *m* Variablen

$$\left(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\right)$$

finde eine Linearkombination zu **einer** neuen Variable

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_m x_m$$

so dass diese die Gruppen $c_1$ und $c_2$ optimal trennt.

## Kriterium der Optimierung

Das Optimierungskriterium für die Wahl der $b_j$ lautet

$$\frac{QS_{Between}}{QS_{Within}} = \frac{\text{erklärte Variation}}{\text{nicht erklärte Variation}} = \max$$

Die der $b_j$ sind so zu wählen, dass auf der neuen Variable y die Streuung zwischen den Gruppen zu der Streuung innerhalb der Gruppen ein maximales Verhältnis hat.

## 2D-Beispiel

Man möchte trennen

**Stechmücken**
$c_1$

**Blindmücken**
$c_2$

## 2 Gruppen

anhand von

## 2 Variablen

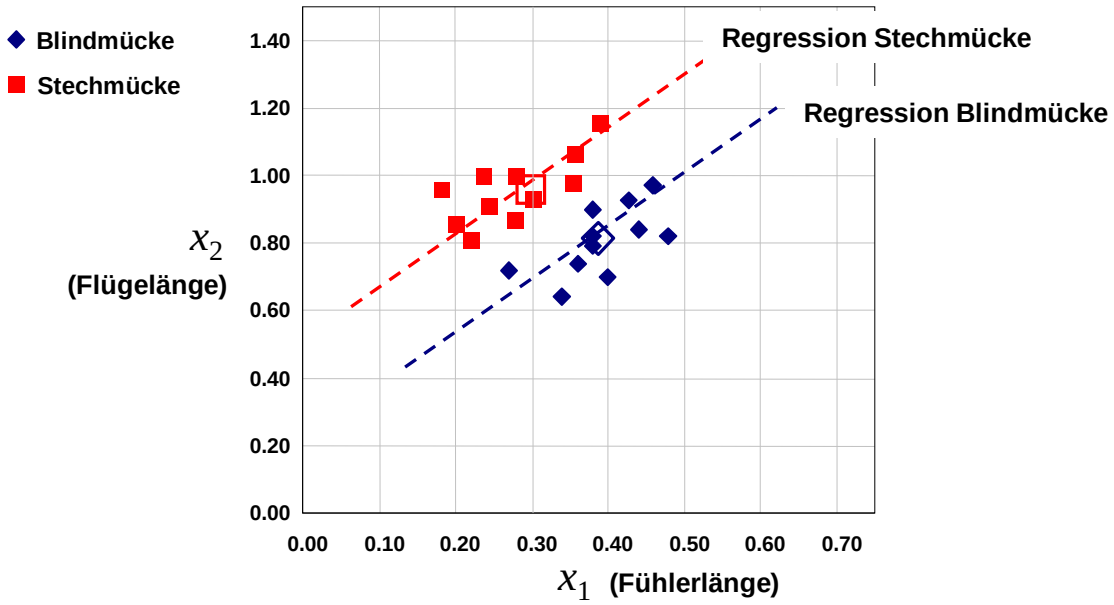
**Fühlerlänge**
$x_1$

**Flügellänge**
$x_2$

## Anforderung

- **Maximale** Gruppentrennung (Mittelwerte)
- **Minimale** Klassifikationsfehler (Fall-Klassifikation)

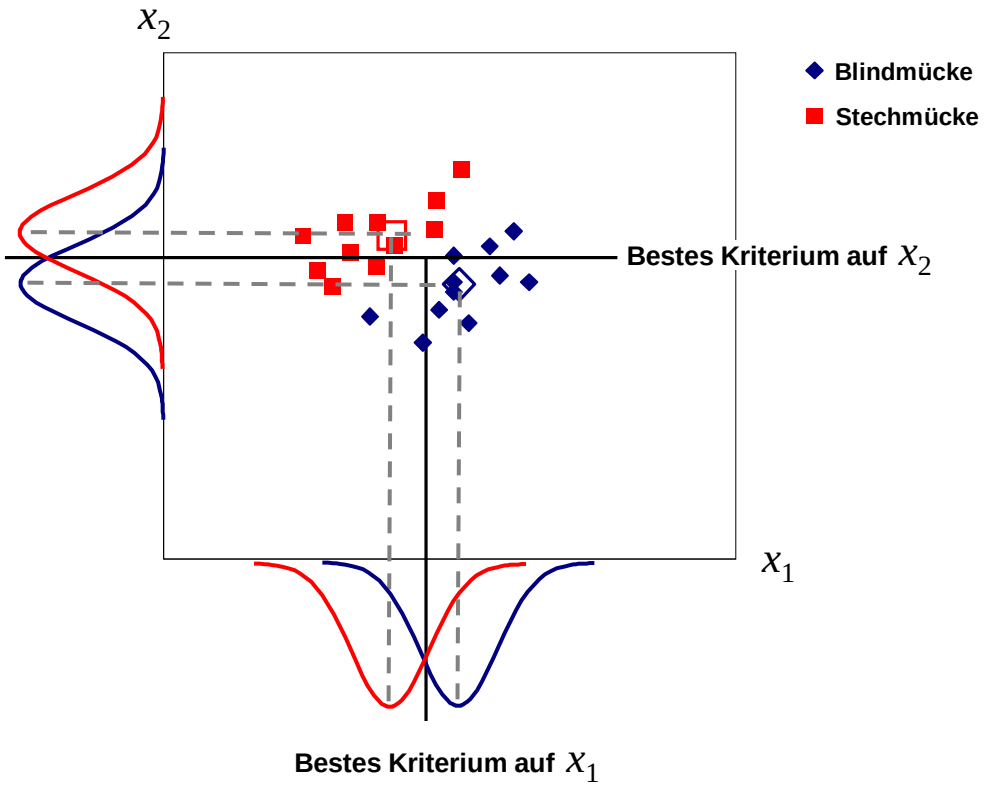
## Variablenraum




## Ausgangslage

- Klassifiziere anhand von Fühlerlänge ($X_1$) und Flügellänge ($X_2$) möglichst eindeutig in Stechmücke ($c_1$) und Blindmücke ($c_2$).
- In beiden Gruppen existiert eine Korrelation der Variablen Fühlerlänge ($X_1$) und Flügellänge ($X_2$).

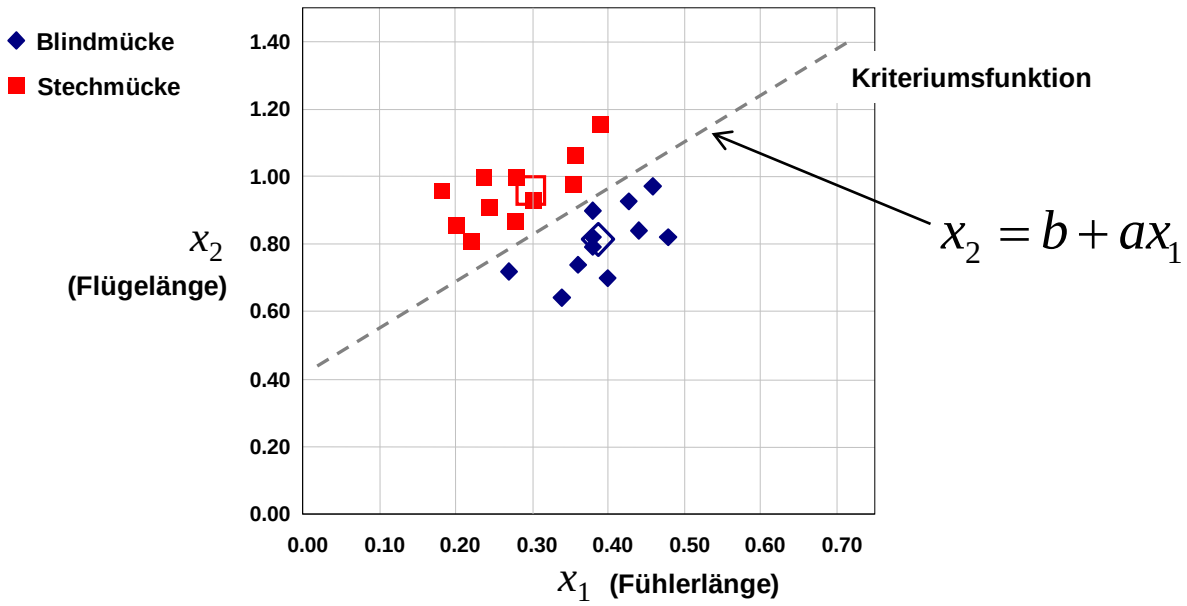
## Variablenraum




## Problem

- Klassifiziere anhand von Fühlerlänge ($X_1$) und Flügellänge ($X_2$) möglichst eindeutig in Stechmücke ($c_1$) und Blindmücke ($c_2$).
- Das geht mit einem Kriteriumswert auf jeder einzelnen Variable $X_1$ und $X_2$ offenbar nicht.

## Variablenraum




## Lösung

Eine lineare Kriteriumsfunktion teilt den Variablenraum in 2 Gebiete: Oberhalb Stechmücke ($c_1$), unterhalb Blindmücke ($c_2$).

$$x_2 = b + ax_1$$

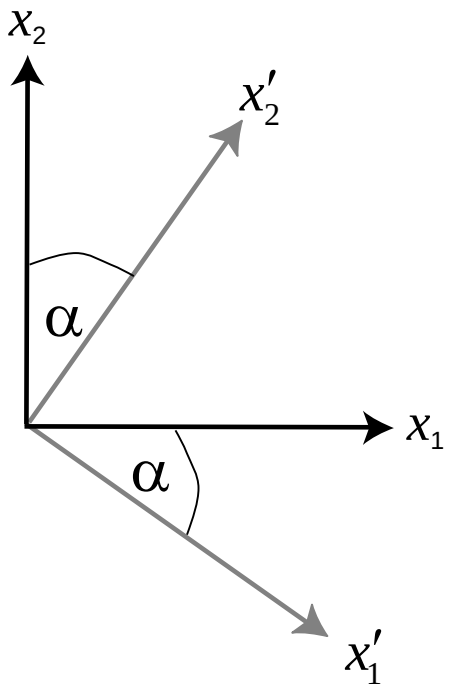
Somit folgt die Klassifikationsfunktion

$$g(x_1, x_2) = \begin{cases} c_1, & \text{wenn } x_2 - ax_1 > b \\ c_2, & \text{wenn } x_2 - ax_1 \leq b \end{cases}$$

**Einfache Lösung**

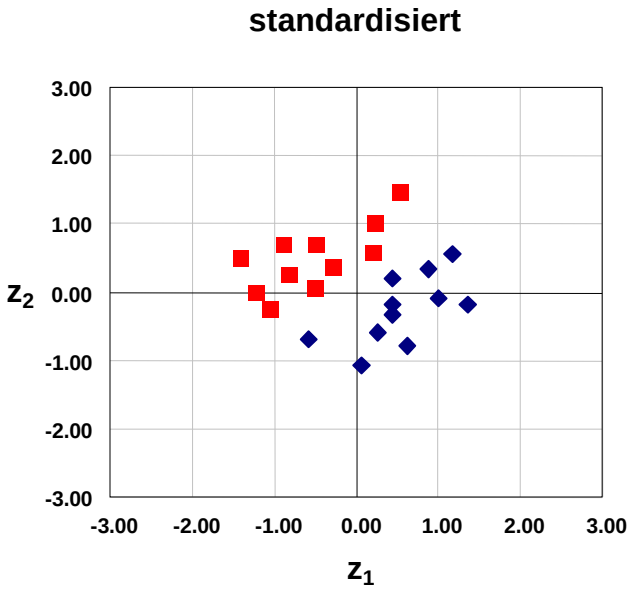
**Zentrierung & Rotation !**

Zuerst die Daten im Nullpunkt zentrieren und dann um den optimalen Winkel α drehen !



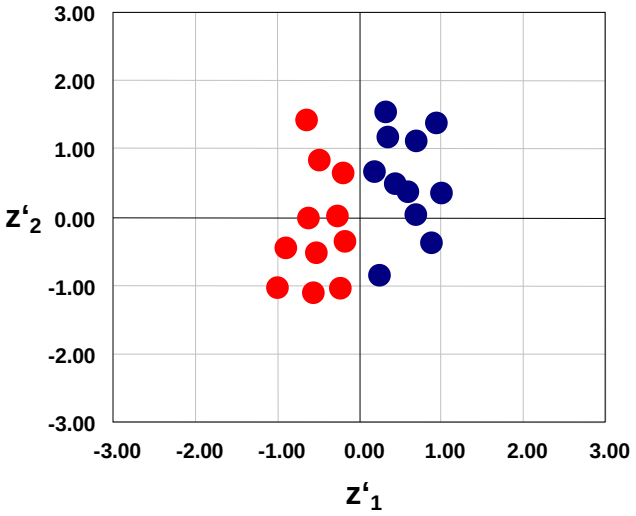

Die Varianz zwischen den Gruppen wird auf der Achse $x'_1$ maximiert, und $x'_2$ steht senkrecht $x'_1$. Eine Parallele zu $x'_2$ liefert das optimale Trennkriterium.

# z-Standard


standardisiert
3.00
2.00
1.00
0.00
-1.00
-2.00
-3.00
$z_2$
-3.00
-2.00
-1.00
0.00
1.00
2.00
3.00
$z_1$

## z-Standard

**Koordinaten rotiert um $\alpha = 46°$ (clockwise)**

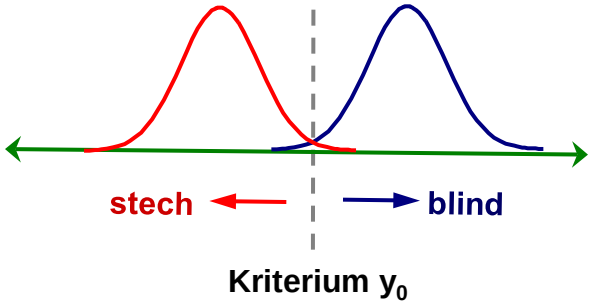



## Diskriminanz-funktion

- Die neue x- Achse $z_1$' ist die Diskriminanzfunktion y. Auf ihr läßt sich ein Kriterium zur optimalen Trennung beider Gruppen finden.
- Da eine Drehoperation auf die Diskriminanzfunktion geführt hat, ist sie darstellbar als eine Linearkombination der alten Koordinaten:

$$z_1' = b_1 z_1 + b_2 z_2$$

## y: Linear-kombination




$$\begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix}\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} z_1' \\ z_2' \end{pmatrix} \quad \Rightarrow \quad \begin{aligned} z_1\cos\alpha - z_2\sin\alpha &= z_1' \\ z_1\sin\alpha + z_2\cos\alpha &= z_2' \end{aligned}$$

Da $y = z_1'$ gilt

$$\boxed{y = b_1 z_1 + b_2 z_2}$$

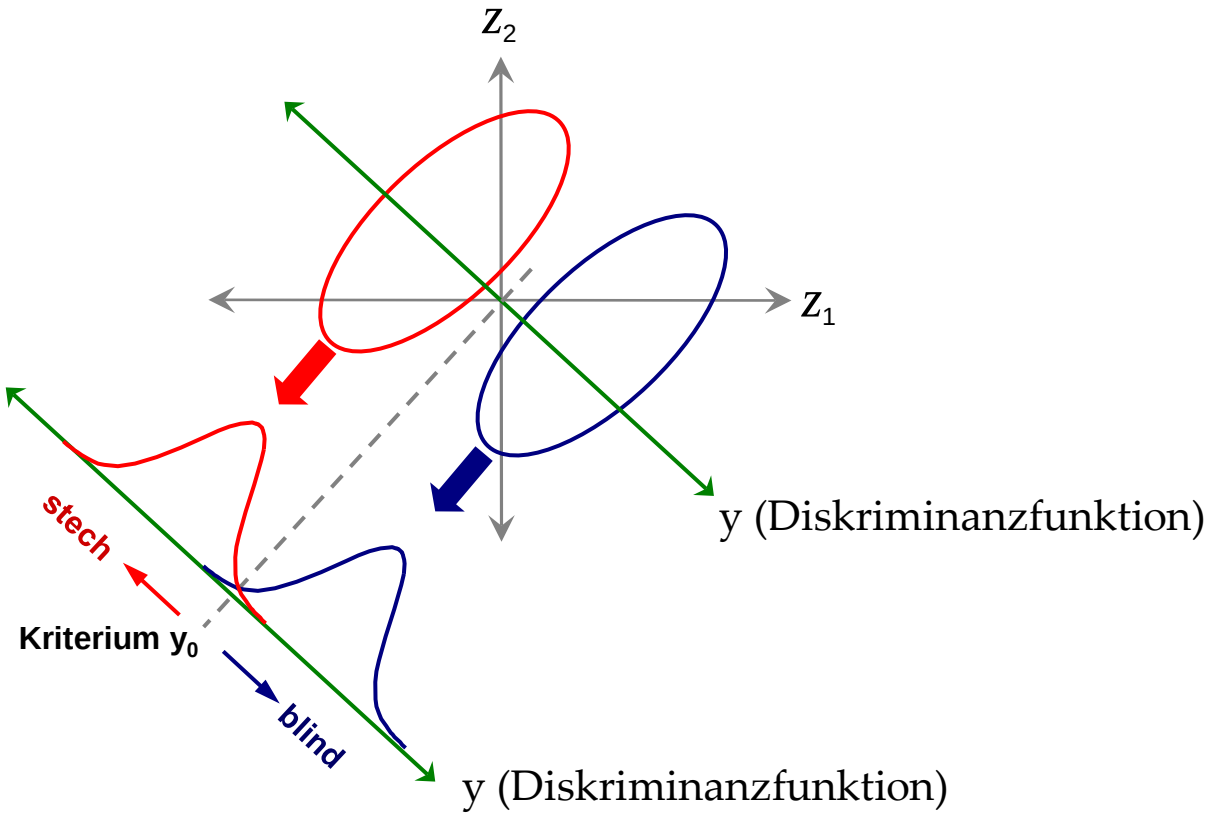
mit $b_1 = \cos\alpha$ und $b_2 = -\sin\alpha$

## Koeffizienten von y

Das Auffinden der Koeffizienten $b_1$ und $b_2$ ist also identisch mit dem Problem, den optimalen Drehwinkel $\alpha$ zu bestimmen. Hierfür braucht man ein Kriterium der gewünschten maximalen Trennung, und die Lösung des dahinter stehenden Maximierungsproblems.

[Excel-Beispiel]

## Rotation zur y - Funktion




## Klassifikation

- Case-Classification durch einfachen Vergleich mit dem Kriterium $y_0$.
- Prüfung des Gruppenunterschieds mit einem einfachen t - Test auf y.