

Wahrscheinlichkeitsrechnung und
Statistik für Biologen
Diskriminanzanalyse

Martin Hutzenthaler & Dirk Metzler

11. Juli 2012

Inhaltsverzeichnis

1 Ruf des Kleinspechts	1
2 Modell	5
2.1 Vorgehen der Diskriminanzanalyse	5
2.2 (Mehrdimensionale) Normalverteilung	6
3 Zurück zu den Rufen	6
3.1 eine Variable	7
3.2 zwei Variable	10
3.3 zehn Dimensionen	16
4 Hauptkomponentenanalyse (PCA)	18

1 Ruf des Kleinspechts

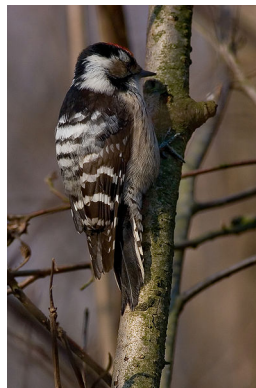


photo (c) Thermos

(Bild zeigt einen Kleinspecht (*Picoides minor*))

Man kann die Geschlechter optisch unterscheiden.

Frage: Geht es auch akustisch?

Ruf des Kleinspechts:

Längen der letzten fünf *Pausen* und *Laute*

... ki — ki — ki — ki — ki — ki — ki
 $p1$ $p2$ $p3$ $p4$ $p5$
 $l1$ $l2$ $l3$ $l4$ $l5$

Frage:

Kann man aus den Längen der Pausen und der Laute

$(p1, p2, p3, p4, p5, l1, l2, l3, l4, l5)$

das Geschlecht bestimmen?

Daten: 62 Rufe von Kleinspechten

18 Rufe von Männchen

44 Rufe von Weibchen

Daten von Dr. Kerstin Höntsch, Frankfurt (siehe <http://www.kleinspecht.de>)

aufbereitet von Dr. Brooks Ferebee, Frankfurt

Die Daten in computergerechter Form:

	G	p1	p2	p3	p4	p5	l1	l2	l3	l4	l5
1	1	0.1719	0.1581	0.1726	0.1785	0.1697	0.0740	0.0703	0.0674	0.0725	0.0660
2	1	0.1052	0.1175	0.0986	0.1008	0.1052	0.0957	0.1023	0.0950	0.0957	0.0943
3	1	0.1473	0.1407	0.1393	0.1407	0.1465	0.0754	0.0776	0.0769	0.0725	0.0653
4	1	0.1378	0.1400	0.1552	0.1828	0.1393	0.0718	0.0667	0.0645	0.0754	0.0747
5	1	0.1473	0.1371	0.1284	0.1509	0.1371	0.0740	0.0696	0.0725	0.0718	0.0718
6	1	0.1175	0.1451	0.1393	0.1407	0.1661	0.0740	0.0711	0.0754	0.0689	0.0565
7	1	0.1385	0.1262	0.1487	0.1407	0.1603	0.0653	0.0696	0.0747	0.0776	0.0725
8	1	0.1197	0.1146	0.1204	0.1182	0.1161	0.0783	0.0805	0.0783	0.0878	0.0696
9	1	0.1393	0.1269	0.1458	0.1429	0.1291	0.0761	0.0761	0.0769	0.0856	0.0725
10	1	0.1197	0.1204	0.1124	0.1146	0.1240	0.0754	0.0769	0.0848	0.0798	0.0645
11	1	0.1625	0.1589	0.1385	0.1502	0.1690	0.0638	0.0689	0.0696	0.0645	0.0529
12	1	0.1298	0.1465	0.1349	0.1400	0.1756	0.0812	0.0747	0.0747	0.0689	0.0602
13	1	0.1204	0.1226	0.1306	0.1465	0.1581	0.0761	0.0754	0.0674	0.0631	0.0689
14	1	0.1110	0.1081	0.1233	0.1248	0.1385	0.0732	0.0747	0.0732	0.0660	0.0587
15	1	0.1139	0.1313	0.1371	0.1589	0.1777	0.0689	0.0674	0.0682	0.0682	0.0711
16	1	0.1335	0.1168	0.1248	0.1313	0.1306	0.0718	0.0703	0.0689	0.0682	0.0667
17	1	0.1407	0.1407	0.1284	0.1400	0.1516	0.0725	0.0696	0.0740	0.0667	0.0696
18	1	0.1204	0.1182	0.1204	0.1269	0.1538	0.0805	0.0718	0.0769	0.0696	0.0645
19	2	0.1044	0.1204	0.1298	0.1393	0.1153	0.1110	0.1211	0.1342	0.0972	0.1037
20	2	0.1436	0.1342	0.1248	0.1581	0.1966	0.1451	0.1400	0.1335	0.1371	0.1240
21	2	0.0907	0.0943	0.0936	0.0936	0.1168	0.0921	0.0812	0.0798	0.0761	0.0674
22	2	0.0921	0.0979	0.1015	0.1015	0.1385	0.0827	0.0827	0.0754	0.0696	0.0653
23	2	0.1052	0.1168	0.1161	0.1306	0.1545	0.0776	0.0732	0.0725	0.0711	0.0609
24	2	0.0928	0.0936	0.0943	0.1066	0.1197	0.0819	0.0863	0.0812	0.0819	0.0805
25	2	0.1516	0.1494	0.1603	0.2140	0.1915	0.1414	0.1429	0.1306	0.1385	0.1044

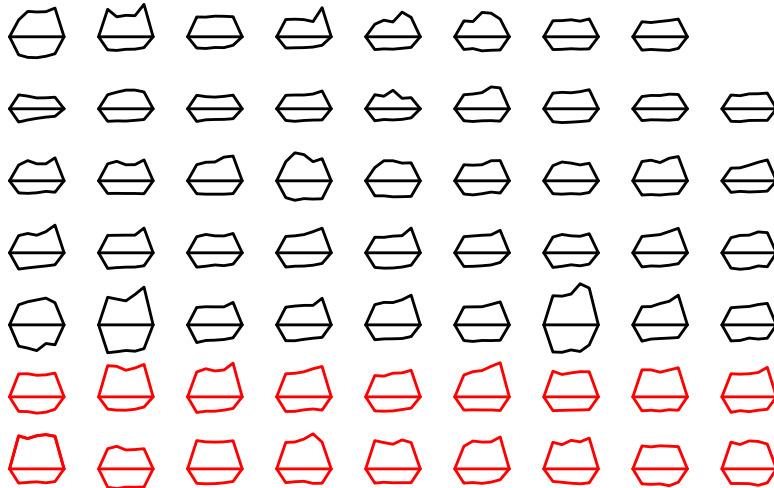
...

Gesucht:

eine dem menschlichen Gehirn gerechte Darstellung des Vektors

$(p1, p2, p3, p4, p5, l1, l2, l3, l4, l5)$

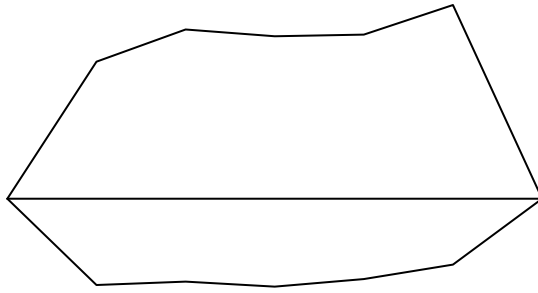
Alle 62 Rufe: *rot*=Männchen, *schwarz*=Weibchen



Mit dem Auge kann man Unterschiede erkennen:

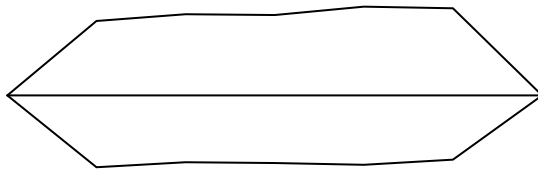
Männchen oder Weibchen?

Typisch Männchen



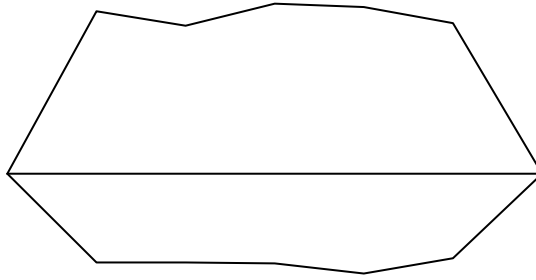
Männchen oder Weibchen?

Typisch Weibchen



Männchen oder Weibchen?

Männchen



Das Auge (das Gehirn) sieht Unterschiede.

Schafft es der Computer auch? (mit Hilfe der Mathematik)
bzw. können wir ein **reproduzierbares** Verfahren angeben?

Das Auge (das Gehirn) sieht Unterschiede.

Schafft es der Computer (mit Hilfe der Mathematik) auch?

2 Modell

Die 10 Zahlen

$(p1, p2, p3, p4, p5, l1, l2, l3, l4, l5)$

fassen wir als die Koordinaten eines Punktes im 10-dimensionalen Raum \mathbb{R}^{10} auf.

Jeder Ruf entspricht einem Zufallspunkt im \mathbb{R}^{10} :

Männchenrufe aus einer Population mit Dichte f_m

Weibchenrufe aus einer Population mit Dichte f_w

Gesucht: Eine Regel, die jeden neuen Punkt

$x = (p1, p2, p3, p4, p5, l1, l2, l3, l4, l5)$
einer der beiden Populationen zuweist.

2.1 Vorgehen der Diskriminanzanalyse

Verfahren

1. Schätze f_m und f_w

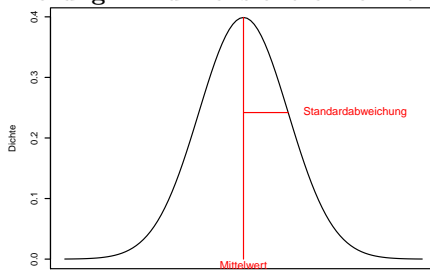
2. Ordne x der Population mit dem *größeren f -Wert* zu.

Wir benutzen für f_m und f_w *mehrdimensionale Normalverteilungen*.

Vorteil: Leicht anzupassen. Wir müssen nur Mittelwert(vektor) und Varianz (mehrdimensional: die Kovarianzmatrix) schätzen.

2.2 (Mehrdimensionale) Normalverteilung

Erinnerung: Eindimensionale Normalverteilung



Zur Beschreibung einer mehrdimensionalen Normalverteilung benötigt man

- Einen Mittelwertvektor μ
- Ein Achsenkreuz (die „Hauptachsen“)
- Standardabweichungen in den Achsenrichtungen

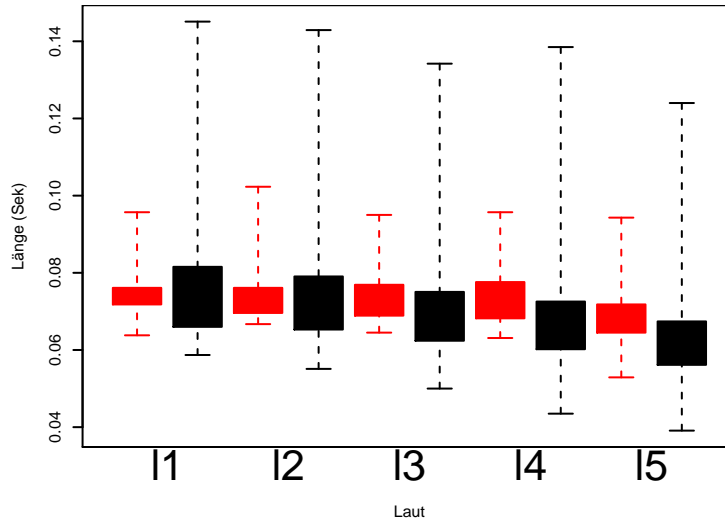
3 Zurück zu den Rufen

In unserem Problem gibt es 10 Dimensionen.

Wir beginnen eindimensional.

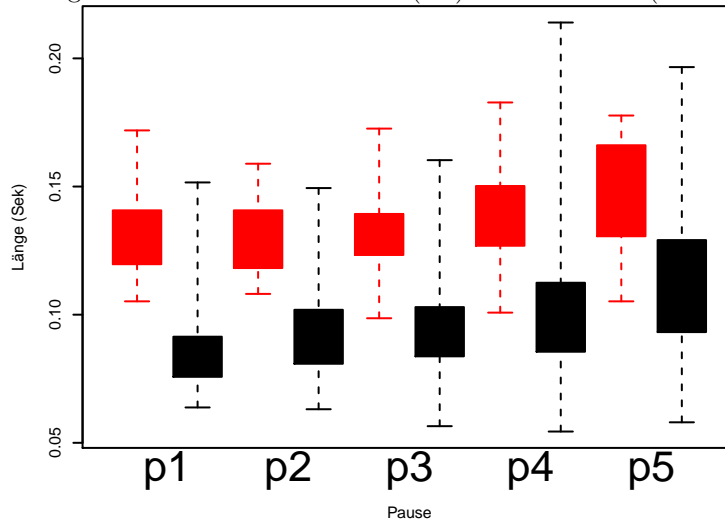
Frage: Welche *eine* der 10 Variablen sollen wir wählen?

Länge der Laute bei Männchen (rot) und Weibchen (schwarz)



Keine gute Trennung der Geschlechter

Länge der Pausen bei Männchen (rot) und Weibchen (schwarz)

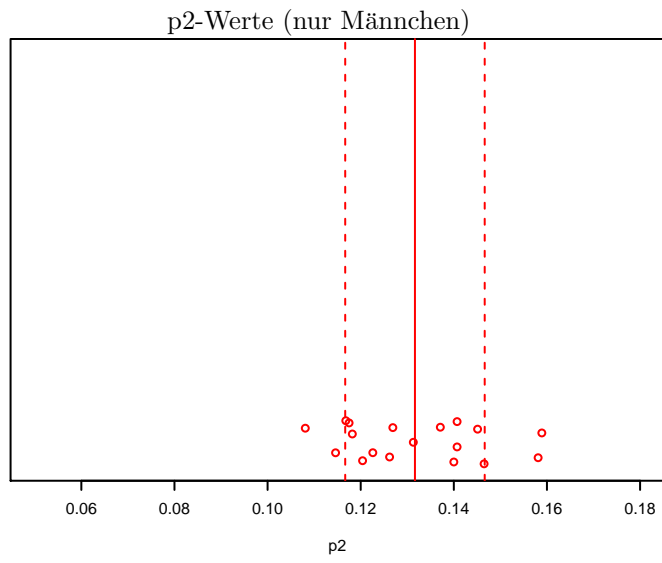
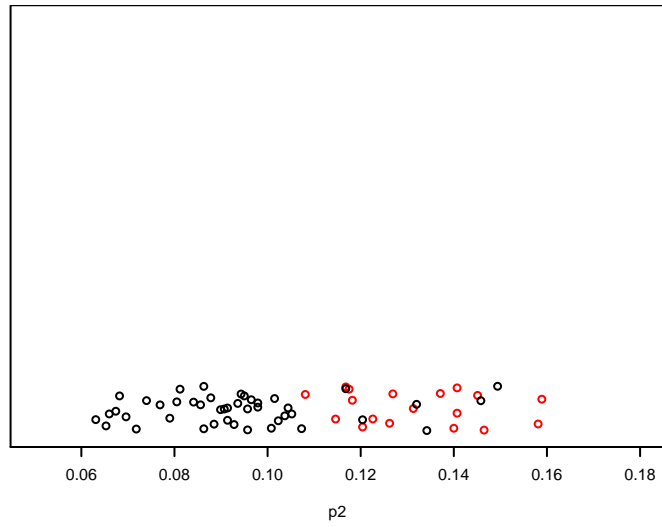


Bei den Männchen sind die Pausen typischerweise länger

3.1 eine Variable

Wie gut läßt sich das Geschlecht anhand von $p2$, der Länge der zweiten Pause, bestimmen?

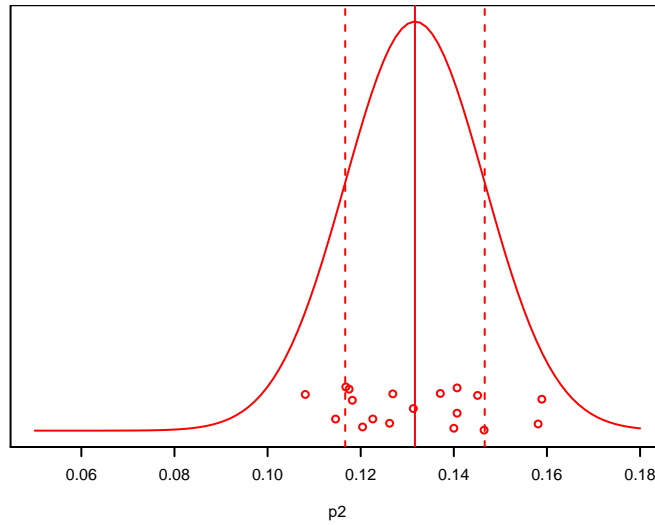
Die $p2$ -Werte (mit Jitter)



Mittelwert $\mu_m = 0,1316$, Standardabweichung $\sigma_m = 0,0150$

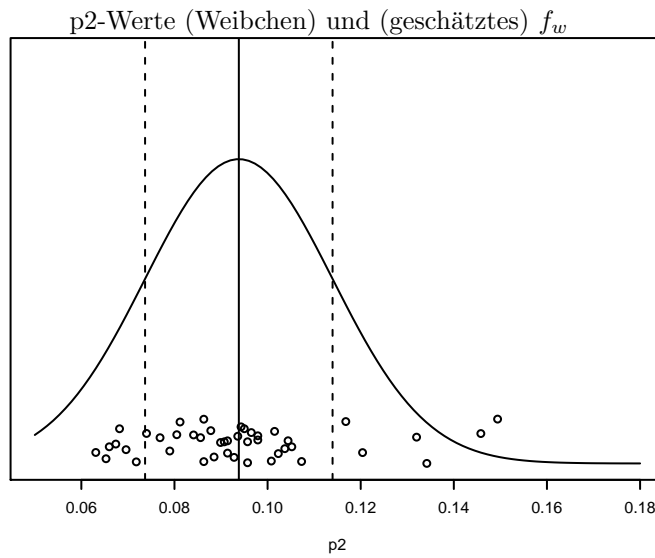
Wir approximieren f_m durch die *Normalverteilung* mit Mittelwert μ_m und Standardabweichung σ_m

p2-Werte (Männchen) und (geschätztes) f_m



p2-Werte (nur Weibchen)
 Mittelwert $\mu_w = 0,0938$, Standardabweichung $\sigma_m = 0,0201$

Wir approximieren f_w durch die *Normalverteilung* mit Mittelwert μ_w und Standardabweichung σ_w

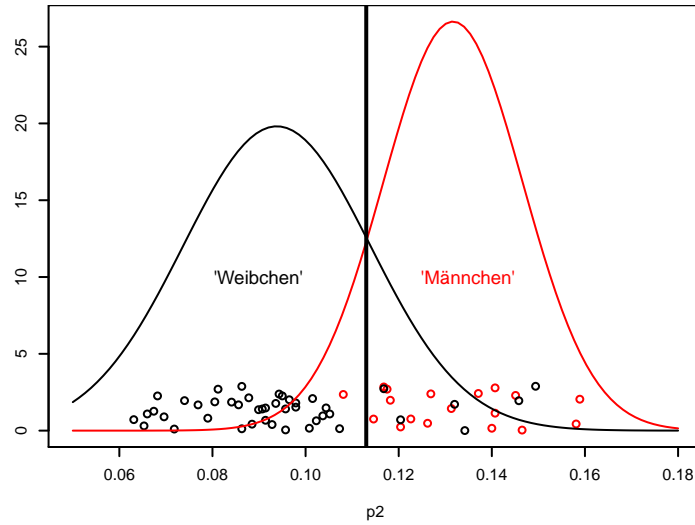


Klassifikationsregel:

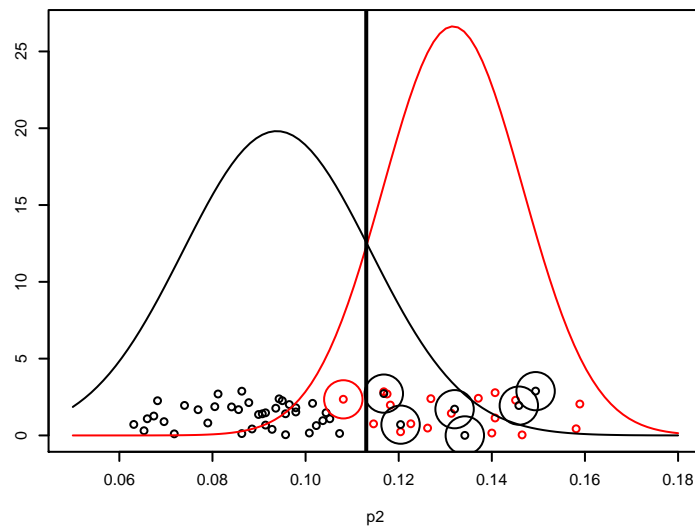
f_m größer \rightarrow „Männchen“

f_w größer \rightarrow „Weibchen“

p2-Werte



Falsch klassifiziert:
 1 Männchen 6 Weibchen

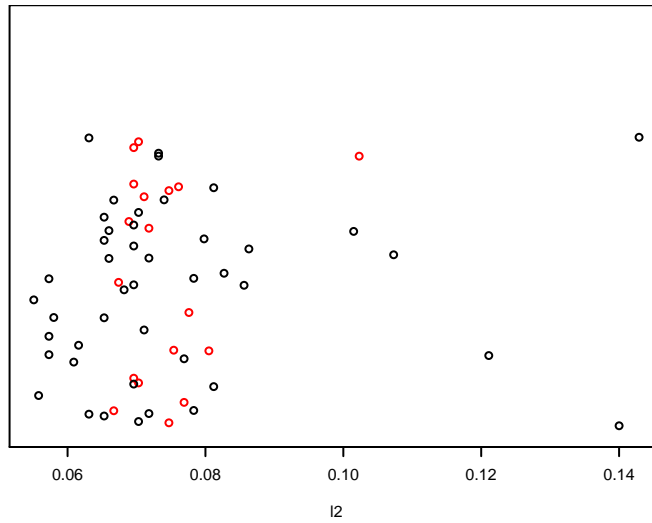


3.2 zwei Variable

Zur Verbesserung der Klassifikation nehmen wir *mehr Information hinzu*, z.B. eine weitere Variable.

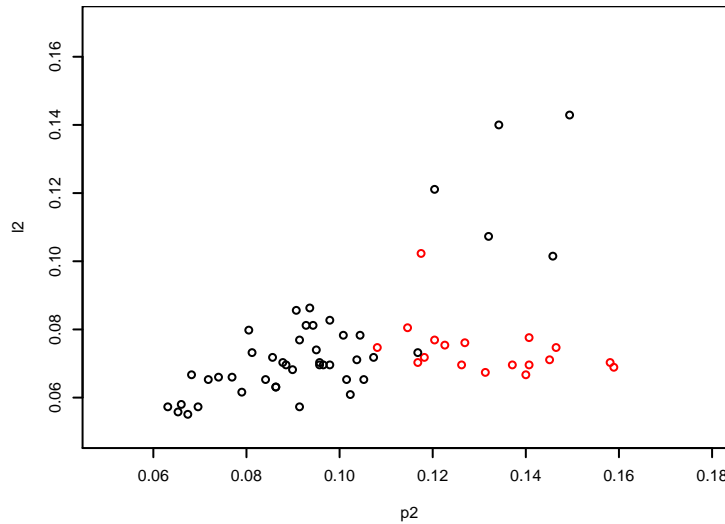
Wir betrachten:

Erste Variable = p2 Zweite Variable = l2



Beobachtung: l_2 allein trennt die Geschlechter sehr schlecht.

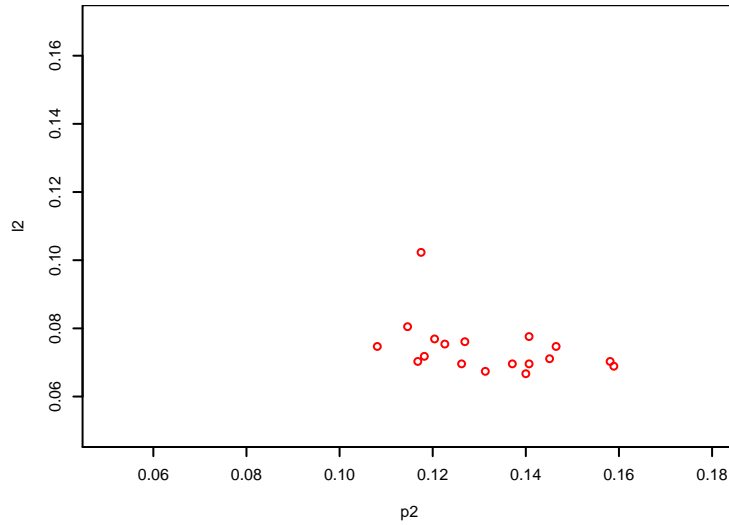
Aber: l_2 *zusammen* mit p_2 gibt zusätzliche Information:



Beispielsweise zeigt die Hinzunahme von l_2 , dass die 5 Punkte oben rechts besser zu den Weibchen passen.

Wir approximieren die Verteilungen von (p_2, l_2) bei Männchen und bei Weibchen durch zweidimensionale Normalverteilungen.

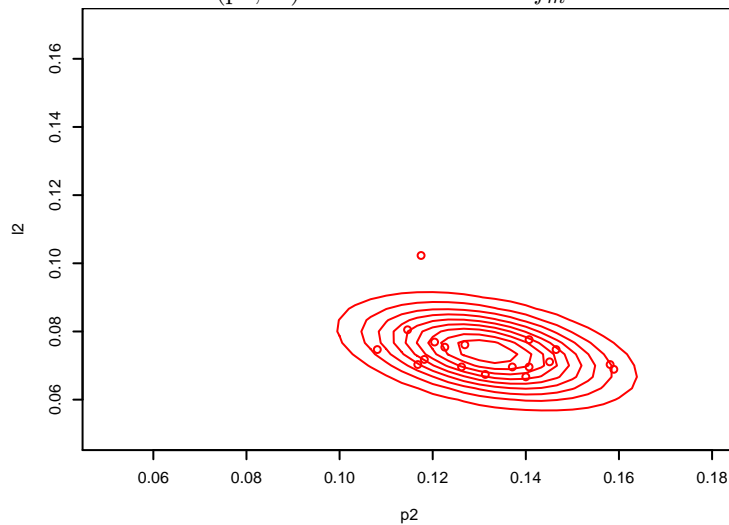
(p_2, l_2) , Männchen



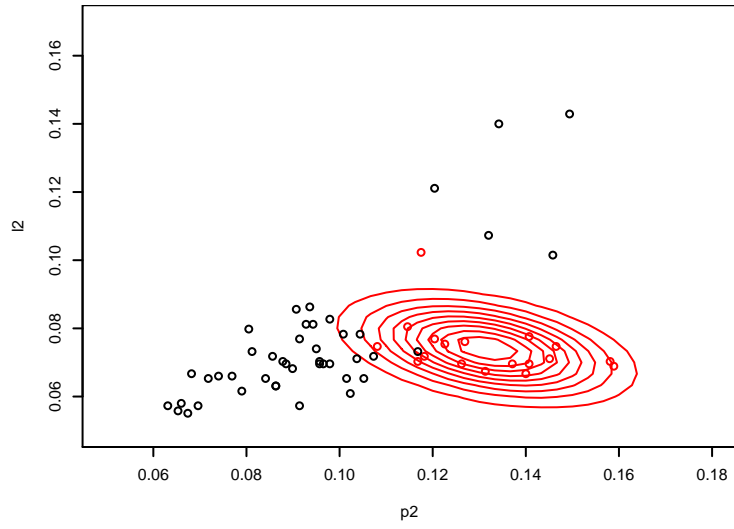
Wie im eindimensionalen Fall schätzen wir
den (zweidimensionalen) *Mittelwert*
und die (zweidimensionale) Varianz (d.h. die sog. *Kovarianzmatrix*)

und approximieren f_m durch eine *zweidimensionale Normalverteilung* mit dem geschätzten
Mittelwert und der geschätzten Varianz.

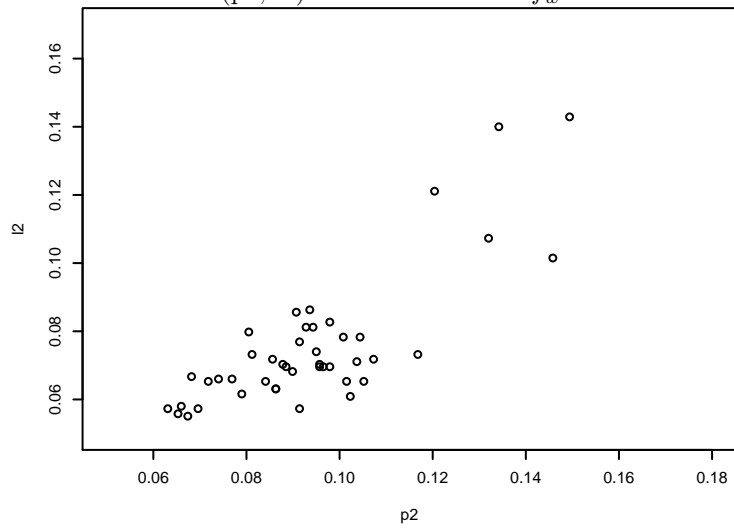
(p_2, l_2) für Männchen und f_m

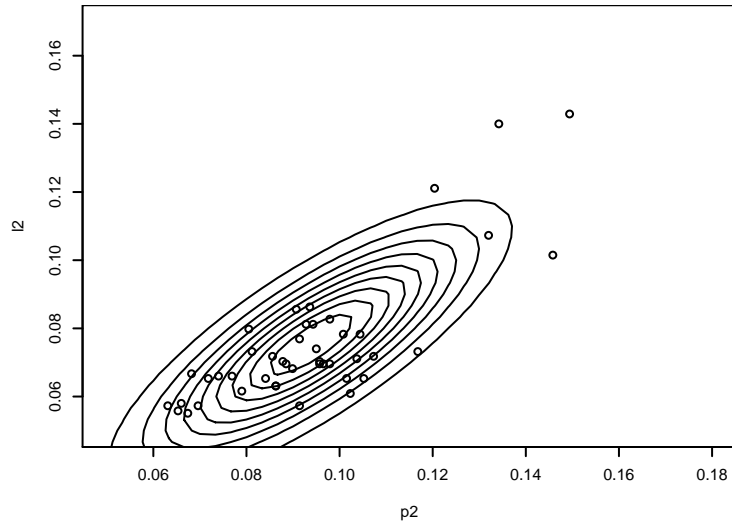


Viele der Weibchen passen schlecht zu f_m :

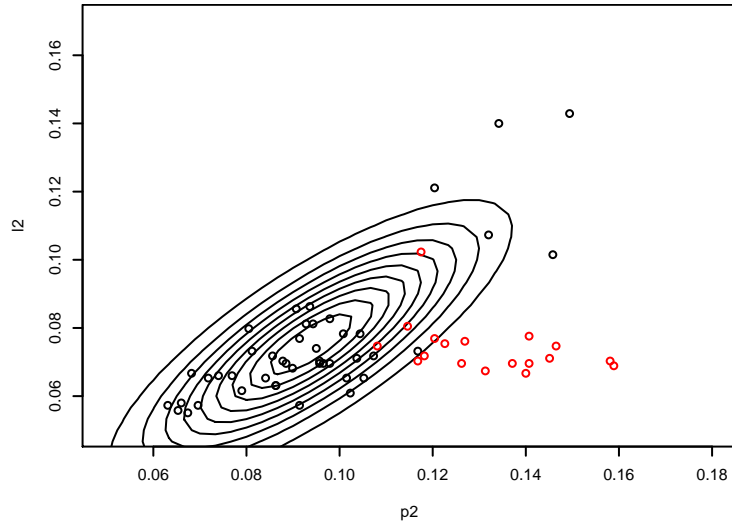


Analog für die Weibchen:
 (p_2, l_2) für Weibchen und f_w





Viele der Männchen passen schlecht zu f_w :

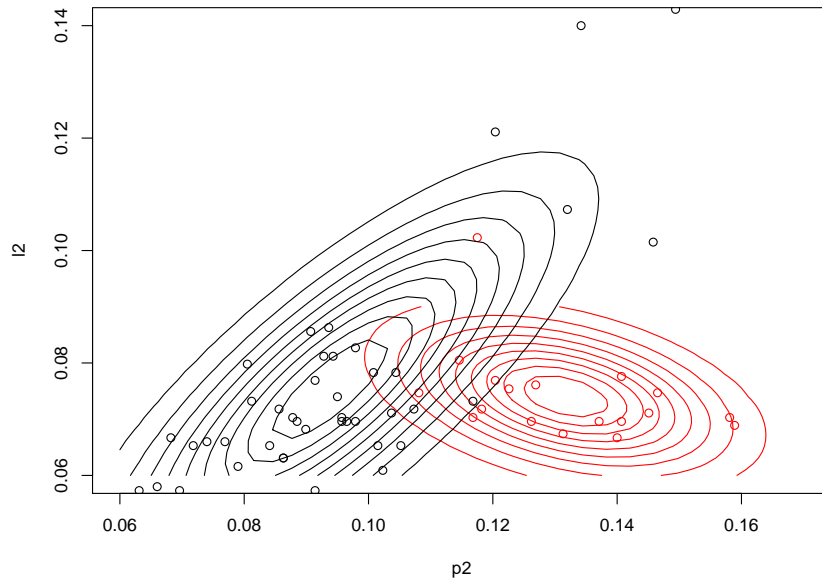


Klassifikation:

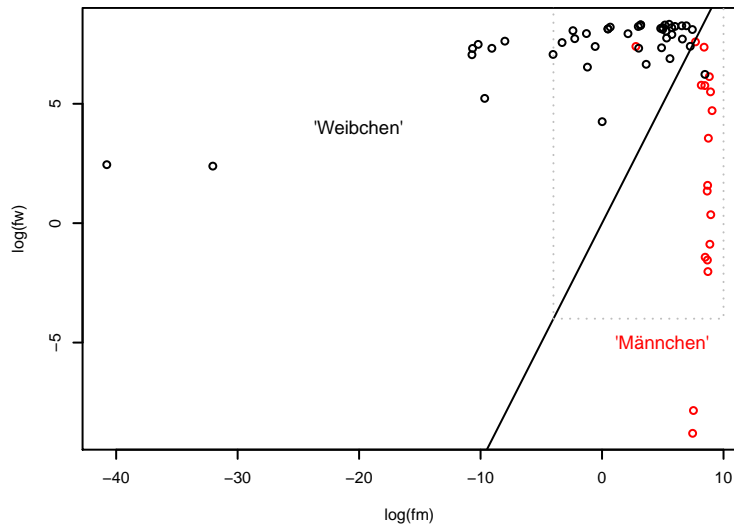
Für jeden Punkt berechnen wir $f_m(x)$ und $f_w(x)$.

$f_m(x)$ größer \rightarrow „Männchen“

$f_w(x)$ größer \rightarrow „Weibchen“

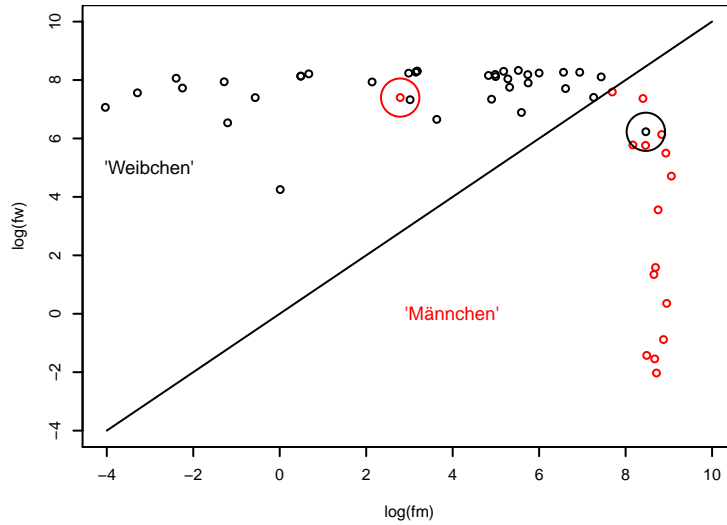


$\log(f_w)$ gegen $\log(f_m)$ und Diagonale:

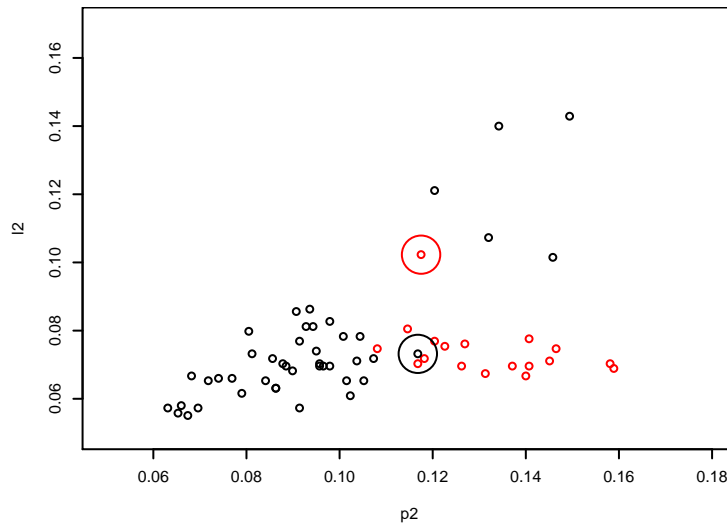


$\log(f_w)$ gegen $\log(f_m)$ und Diagonale, Ausschnittvergrößerung:

Falsch klassifiziert: 1 Männchen, 1 Weibchen (und eigentlich 2 „unentschieden“)



Welche Fälle wurden falsch zugeordnet?



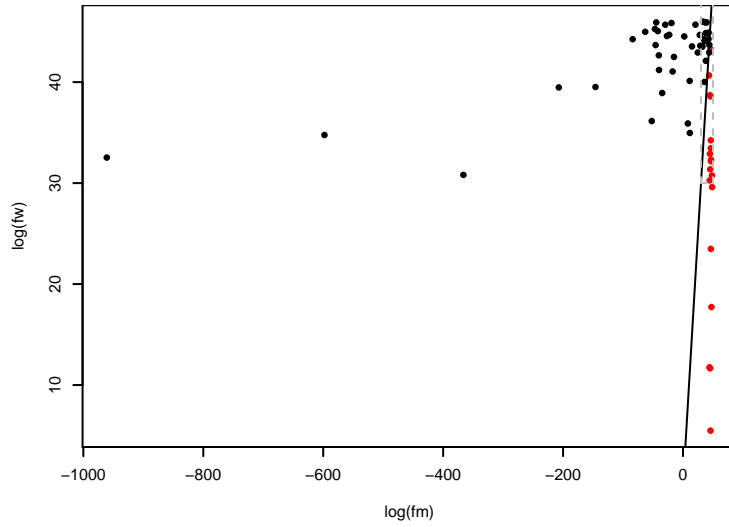
Wenn man nur p_2 und l_2 kennt, ist es sehr verständlich, dass diese Fälle falsch klassifiziert werden.

3.3 zehn Dimensionen

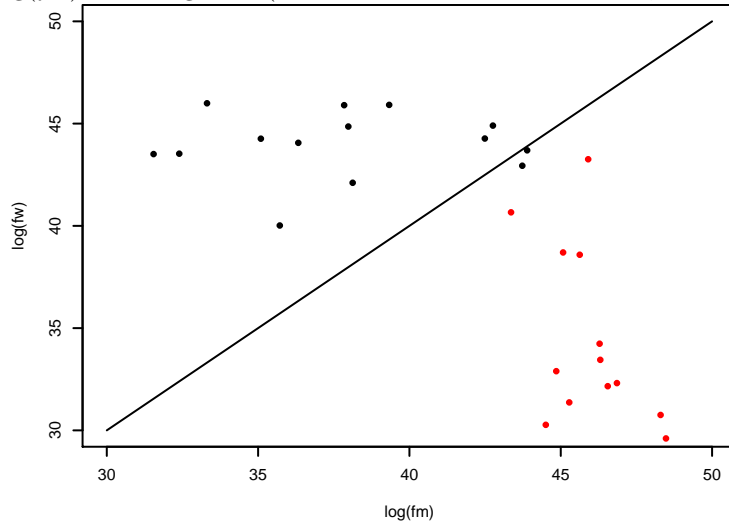
Wir verfahren genauso mit allen Variablen ($p_1, p_2, p_3, p_4, p_5, l_1, l_2, l_3, l_4, l_5$) gemeinsam — mathematisch analog, allerdings geometrisch sehr schwierig darzustellen.

Ergebnis:

$\log(f_w)$ gegen $\log(f_m)$ und Diagonale (basierend auf allen 10 Variablen):

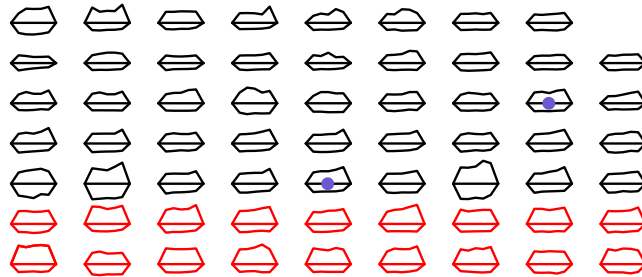


$\log(f_w)$ gegen $\log(f_m)$ und Diagonale (basierend auf allen 10 Variablen, Ausschnittvergrößerung):



Die zwei mit (p2,l2) falsch klassifizierten Fälle wurden nun richtig klassifiziert. Allerdings wurden zwei Weibchen (knapp) falsch klassifiziert.

Falsch klassifiziert



Die beiden falsch klassifizierten Rufe: sie sehen ziemlich „männlich“ aus.

Warnhinweis

Der Anteil der falsch klassifizierten wurde hier nur für Daten geschätzt, die auch für die Anpassung der Klassifizierung verwendet wurden.

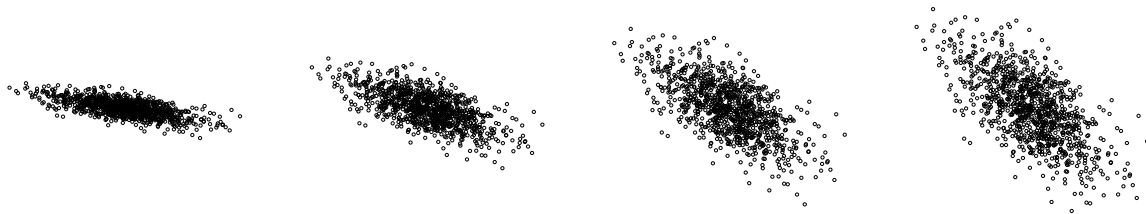
- Der Klassifikationsfehler könnte zu optimistisch geschätzt werden.
- Mögliche Lösungen: Schätze Klassifikationsfehler auf unabhängigen Daten oder Kreuzvalidierung.
- Dieser Effekt ist umso größer je mehr Variablen für die Klassifikation verwendet werden wegen Überanpassung, engl. *overfitting*.

4 Hauptkomponentenanalyse (PCA)

Wir wollen multi-dimensionale Daten visualisieren, um gewisse Muster zu finden.

Wie visualisieren wir, welche multi-dimensionale Datenpunkte nah bei einander liegen?

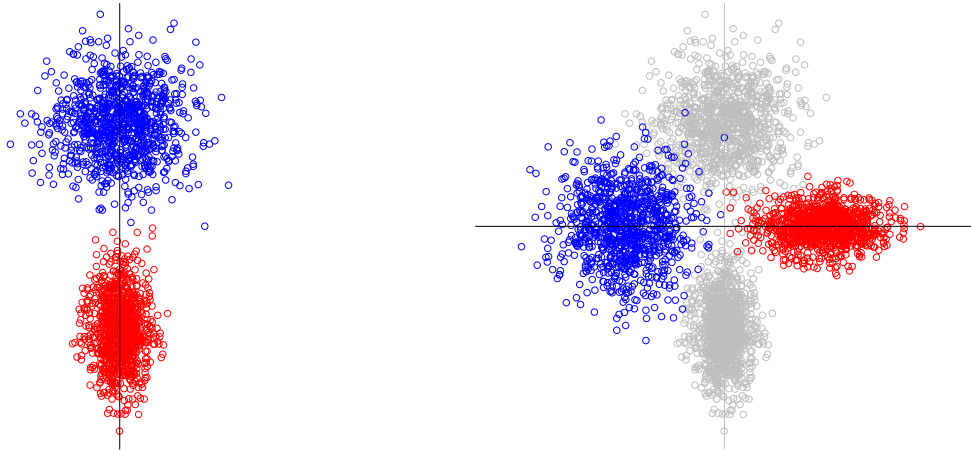
Beispiel: 2-dimensionale Daten in 3 Dimensionen (Vorstellung: Wolke rotiert in 3 Dimensionen)



Um einen guten Blick auf die Daten zu haben wollen wir die Komponenten darstellen, die die meiste Variation beitragen.

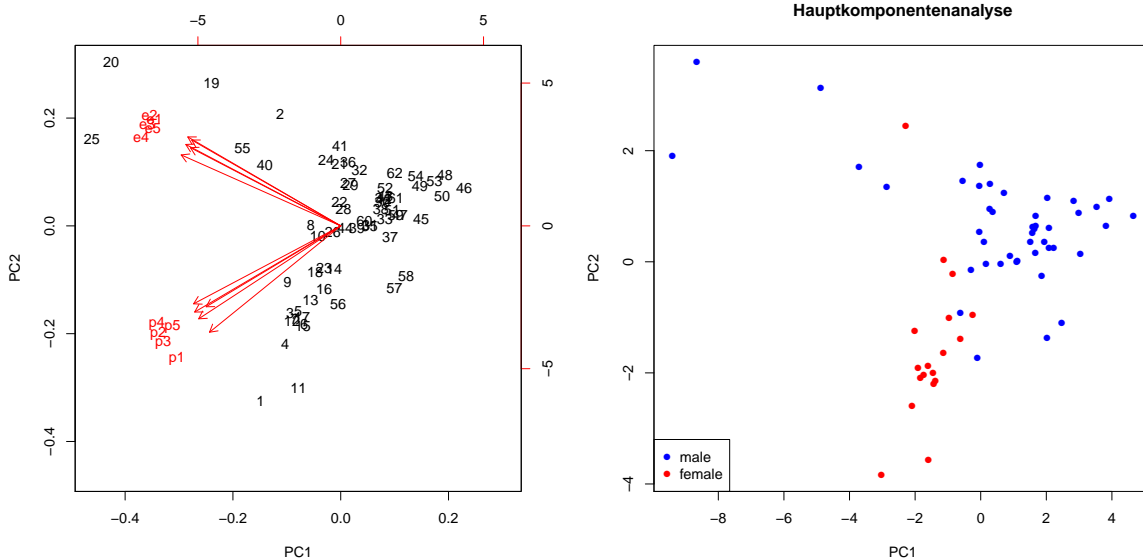
Die Achse mit der größten Variation wird in die x-Achse rotiert, die Achse mit der zweit größten Variation wird in die y-Achse rotiert.

Beispiel: 2-dimensionale Daten

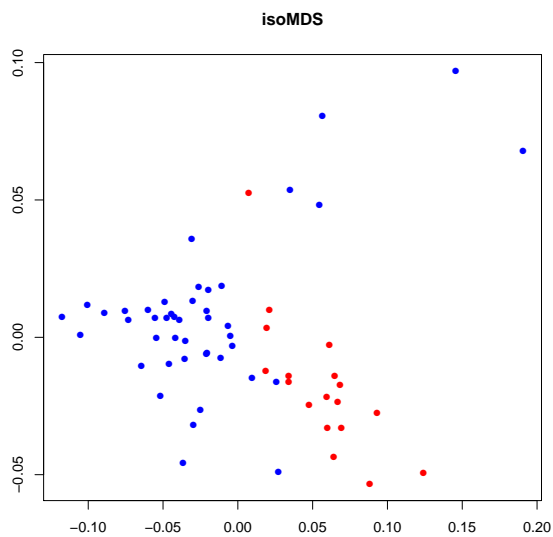


Die **Hauptkomponentenanalyse** (engl. **principal component analysis, PCA**) findet die Achse mit dem größten Beitrag zur gesamten Variation.

PCA für die Kleinspechtrufe



Alternative: Multidimensionale Skalierung



PCA vs. MDS

- PCA ist eine lineare Transformation, d.h. die Hauptkomponenten sind Linearkombinationen der ursprünglichen Variablen.
- MDS ist eine nichtlineare Transformation.
- Dadurch kann MDS in der Ebene Punkte finden, deren Abstände die Abstände im hochdimensionalen Parameterraum besser widerspiegeln.

- Bei PCA lassen sich die Hauptkomponenten aber besser interpretieren und für nachfolgende Analysen verwenden (z.B. lineare Regression).

```

kiki <- read.table("kiki.bb62",h=T)
str(kiki)
pca <- prcomp( ~ p1+p2+p3+p4+p5+e1+e2+e3+e4+e5,
               data=kiki,scale.=TRUE)
biplot(pca)
plot(pca$x[, "PC1"],pca$x[, "PC2"],col=2*as.numeric(kiki$G),
     pch=16,xlab="PC1",ylab="PC2",
     main="Hauptkomponentenanalyse")
legend("bottomleft",col=c("blue","red"),pch=16,
      legend=c("male","female"))

library(MASS)
D <- dist(as.matrix(kiki[4:13]))
mds <- isoMDS(D)
plot(mds$points,pch=16,col=2*as.numeric(kiki$G),
     xlab="",ylab="",main="isoMDS")

```

