Wahrscheinlichkeitsrechnung und
Statistik für Biologen
Diskriminanzanalyse

Dirk Metzler

24. März 2026

Inhaltsverzeichnis

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*

					p1			*p2*			*p3*			*p4*			*p5*			
⋯	ki	—	ki	—	ki	—	ki	—	ki	—	ki	—	ki	—	ki	—	ki			
					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, Senckenberg Gesellschaft, Frankfurt (siehe `http://www.kleinspecht.de`)

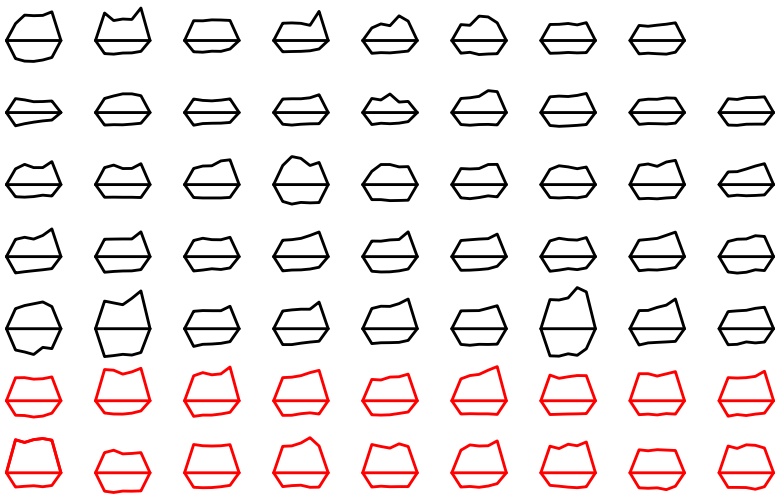
aufbereitet von Dr. Brooks Ferebee, Goethe-Universität, Frankfurt

Gesucht:

eine dem menschlichen Gehirn gerechte Darstellung des Vektors

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

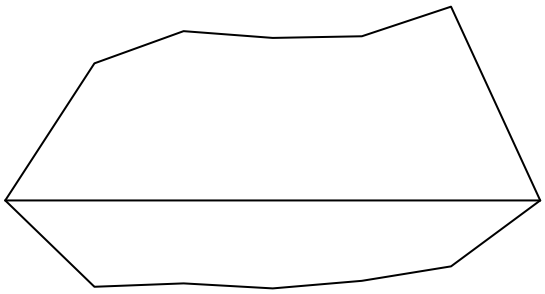
Alle 62 Rufe: *rot*=Männchen, schwarz=Weibchen



Mit dem Auge kann man Unterschiede erkennen:

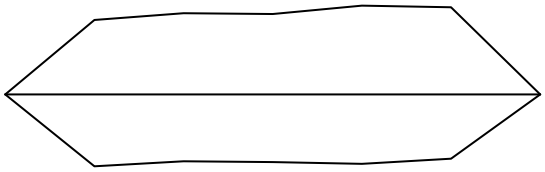
Männchen oder Weibchen?

Typisch Männchen



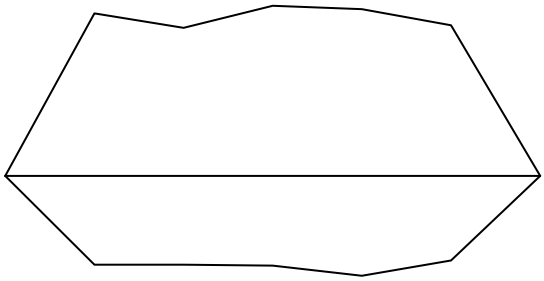
Männchen oder Weibchen?

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

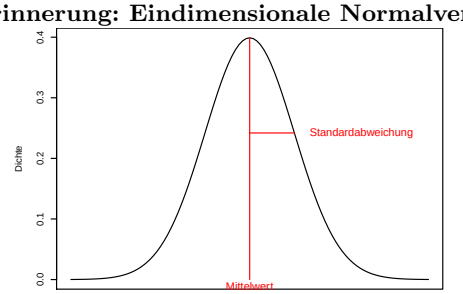
1. Schätze f_m und f_w
2. Ordne x der Population mit dem *größeren f-Wert zu.*

Wir benutzen für f_m und f_w *mehrdimensionale Normalverteilungen.*

Vorteil: Leicht anzupassen. Wir müssen nur Mittelwert(svektor) 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

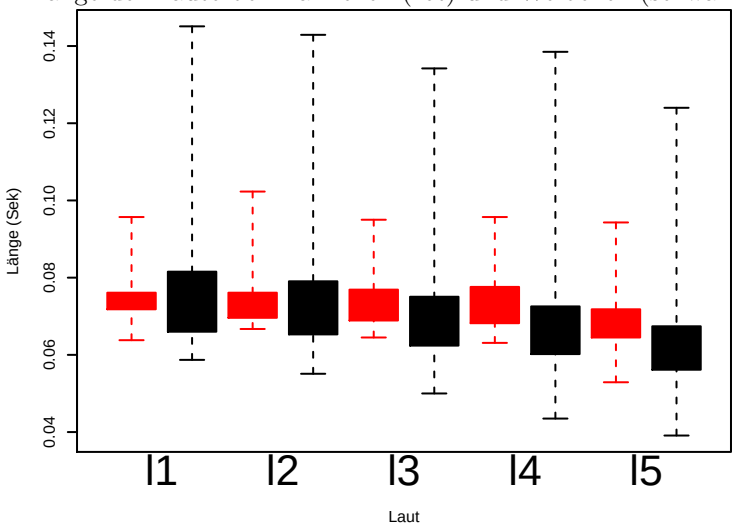
3 Zurück zu den Rufen

In unserem Problem gibt es 10 Dimensionen.

Wir beginnen eindimensional.

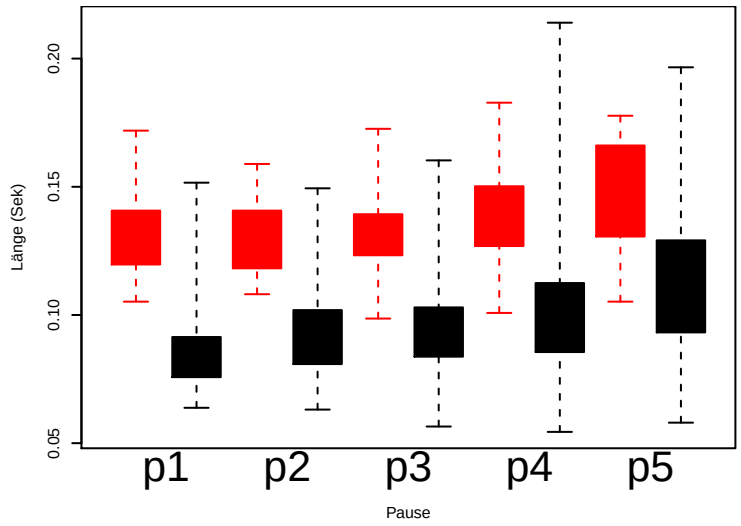
Frage: Welche *eine* der 10 Variablen sollen wir wählen?

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



Keine gute Trennung der Geschlechter

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

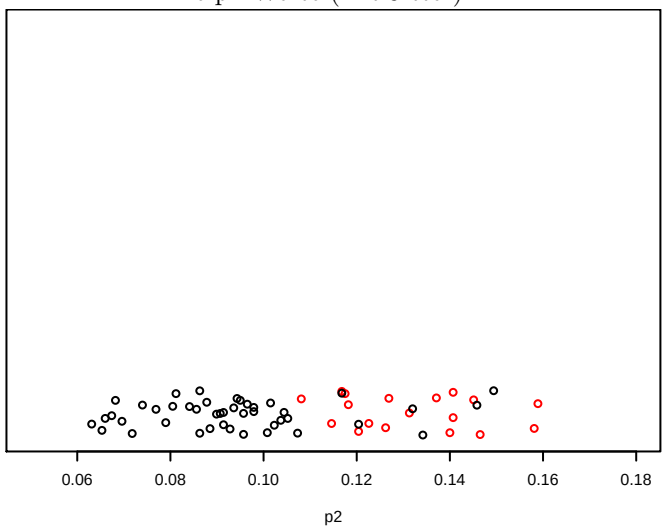


Bei den Männchen sind die Pausen typischerweise länger

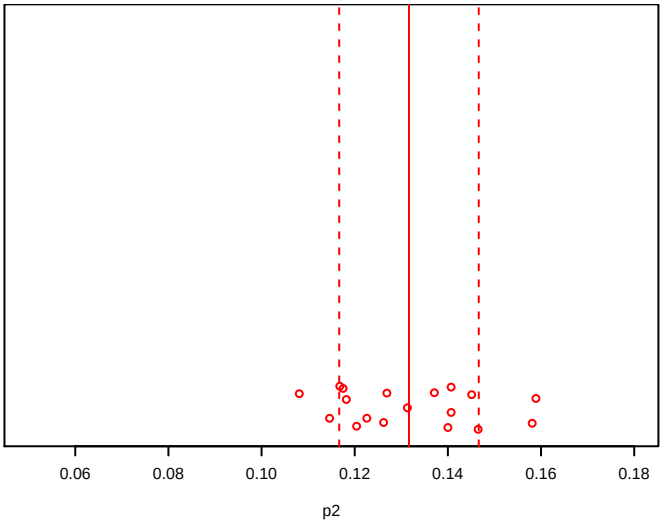
3.1 eine Variable

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

Die p2-Werte (mit Jitter)



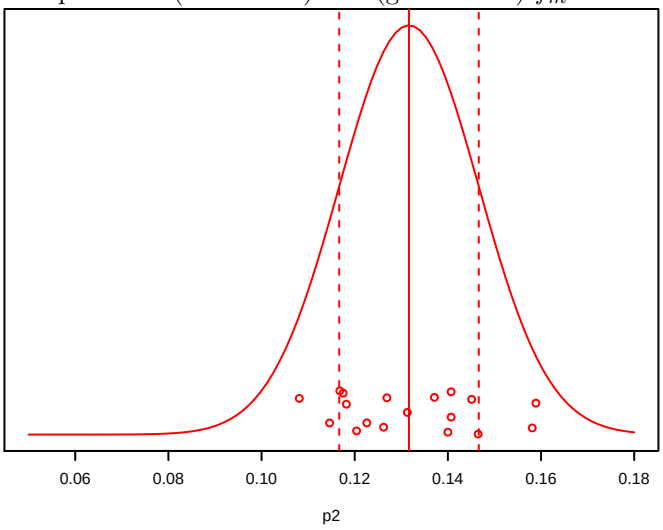
p2-Werte (nur Männchen)



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

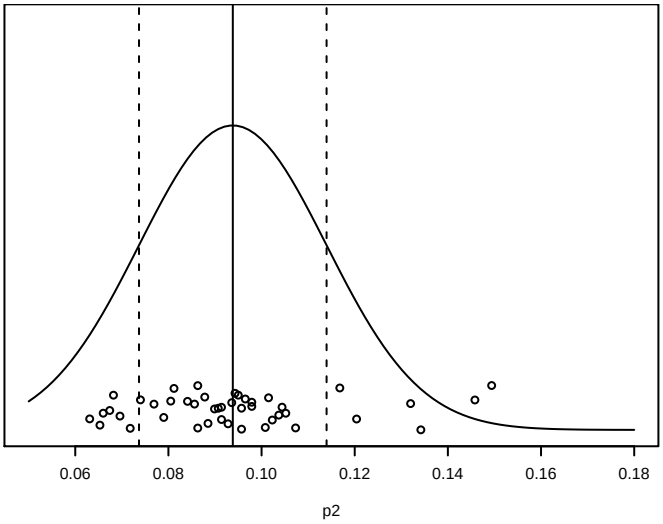
p2-Werte (Männchen) und (geschätztes) f_m



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

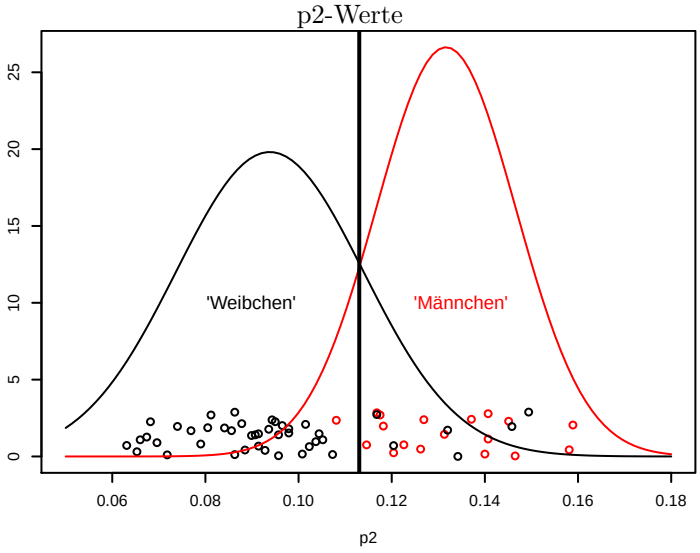
p2-Werte (Weibchen) und (geschätztes) f_w



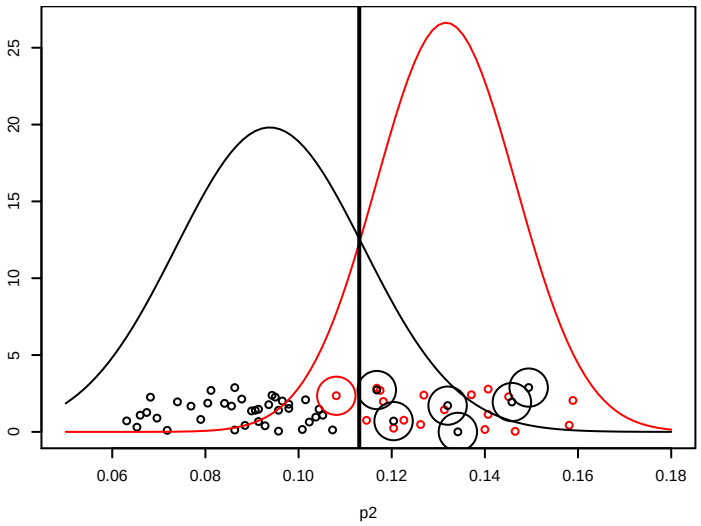
Klassifikationsregel:

f_m größer $\longrightarrow$ „Männchen“

f_w größer $\longrightarrow$ „Weibchen“



Falsch klassifiziert:
1 Männchen 6 Weibchen

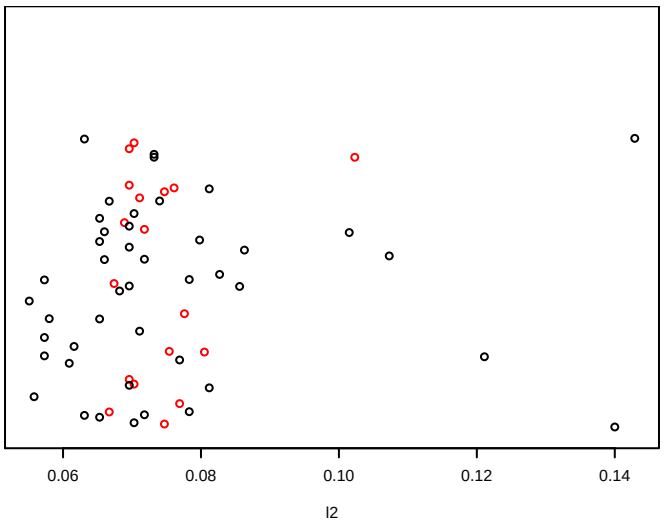


3.2 zwei Variable

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

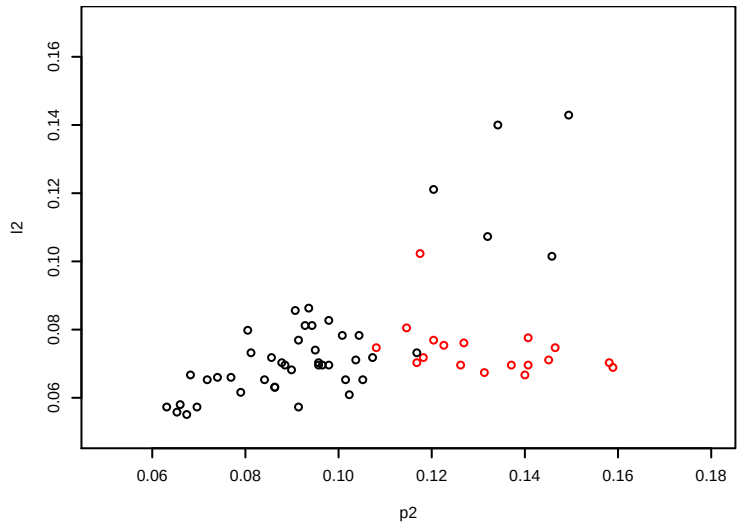
Wir betrachten:

Erste Variable = p2 Zweite Variable = l2



Beobachtung: l2 allein trennt die Geschlechter sehr schlecht.

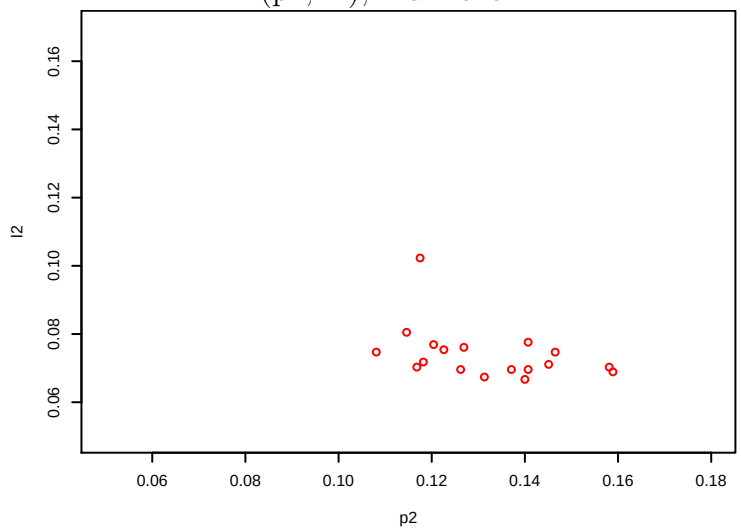
Aber: l2 *zusammen* mit p2 gibt zusätzliche Information:



Beispielsweise zeigt die Hinzunahme von l2, dass die 5 Punkte oben rechts besser zu den Weibchen passen.

Wir approximieren die Verteilungen von (p2, l2) bei Männchen und bei Weibchen durch zweidimensionale Normalverteilungen.

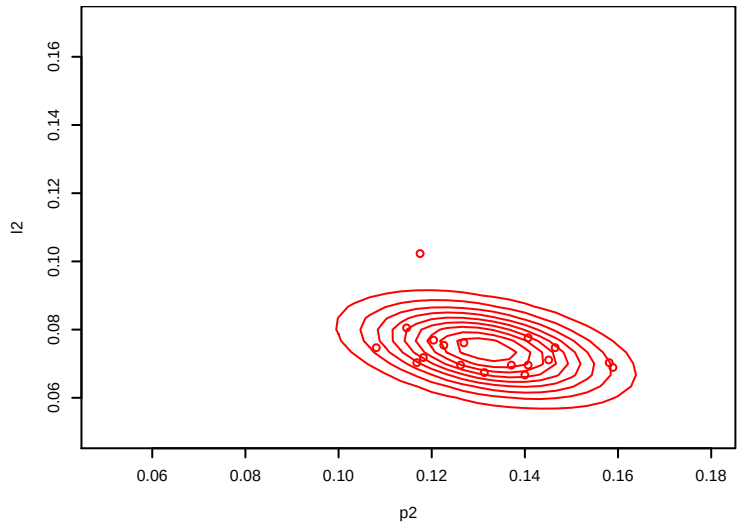
(p2, l2), Männchen



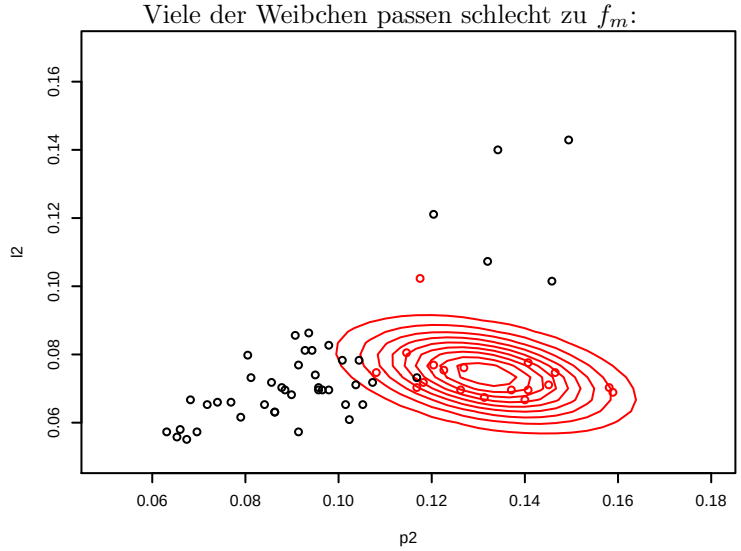
Wie im eindimensionalen Fall schätzen wir
den (zweidimensionalen) *Mittelwert*
und die (zweidimensionale) Varianz (d.h. die sog. *Kovarianzmatrix*)

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

(p2, l2) für Männchen und f_m

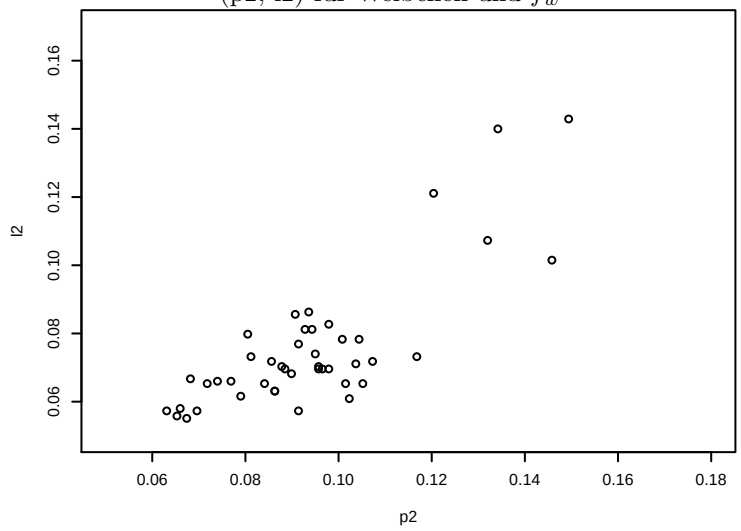


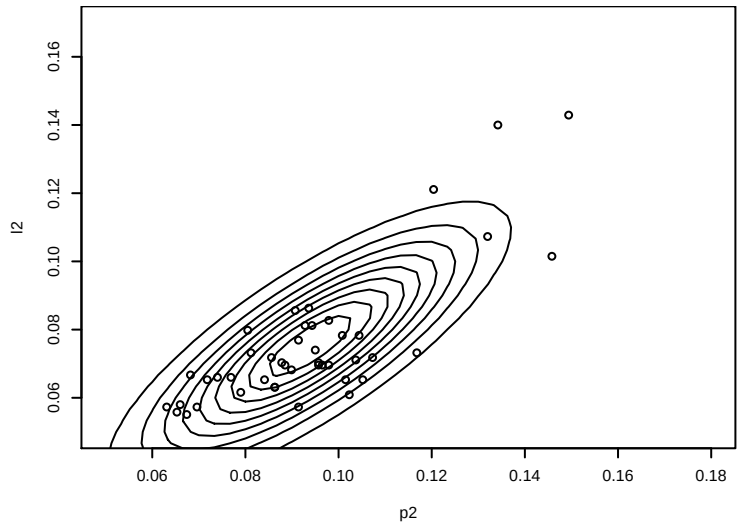
Viele der Weibchen passen schlecht zu f_m:



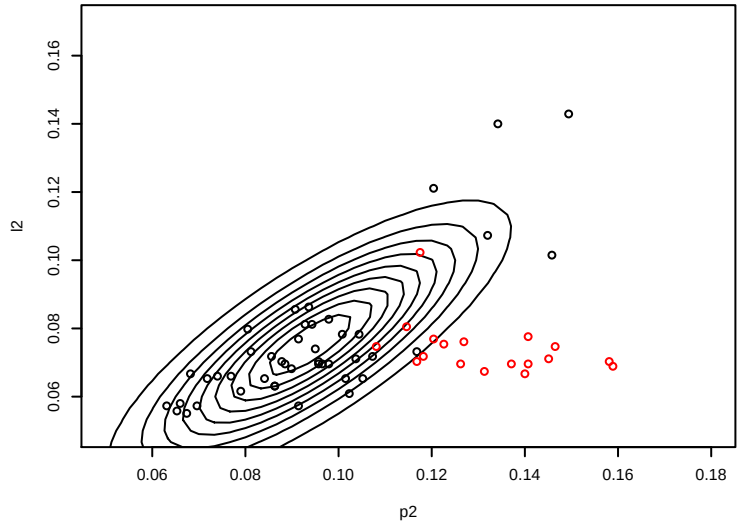
Analog für die Weibchen:

(p2, l2) für Weibchen und f_w

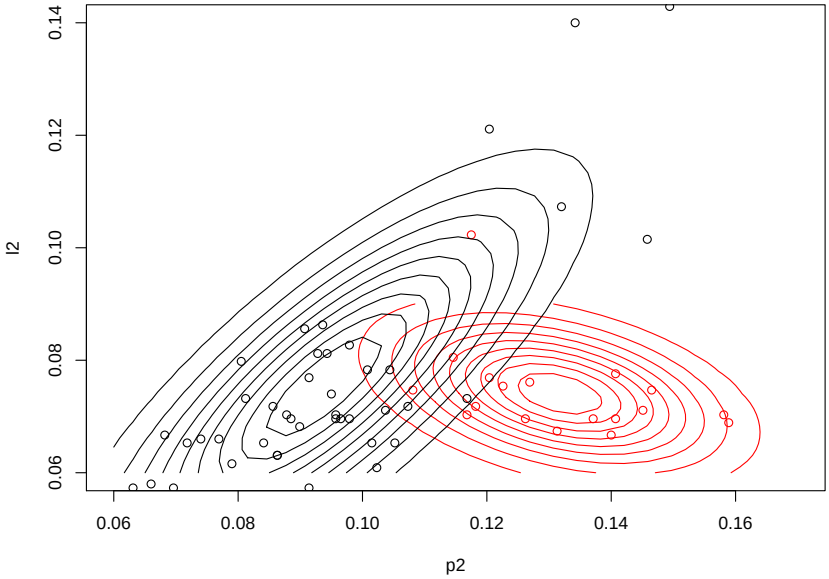




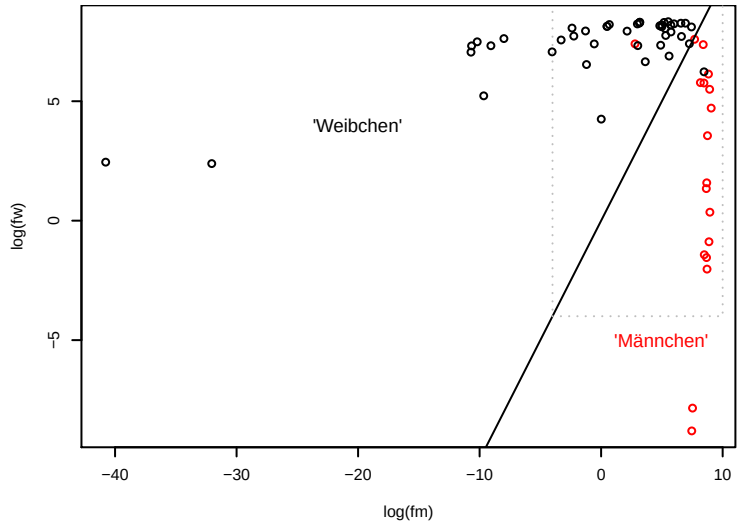
Viele der Männchen passen schlecht zu f_w:



Klassifikation:
Für jeden Punkt berechnen wir $f_m(x)$ und $f_w(x)$.
$f_m(x)$ größer $\longrightarrow$ „Männchen“
$f_w(x)$ größer $\longrightarrow$ „Weibchen“

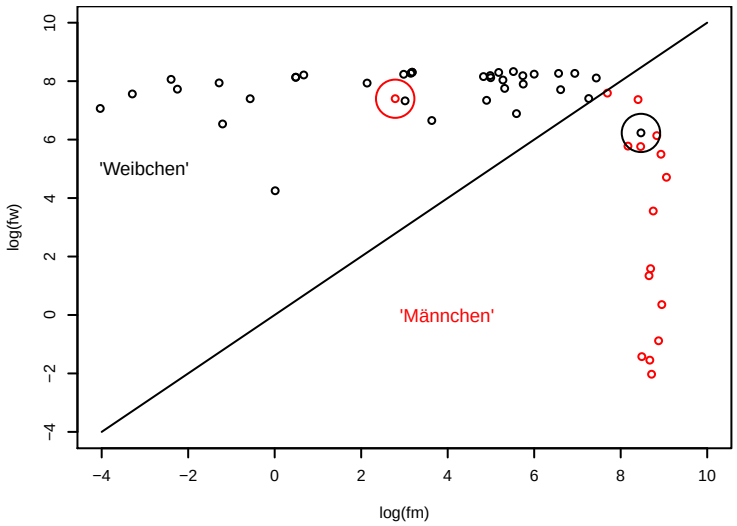


$\log(f_w)$ gegen $\log(f_m)$ und Diagonale:

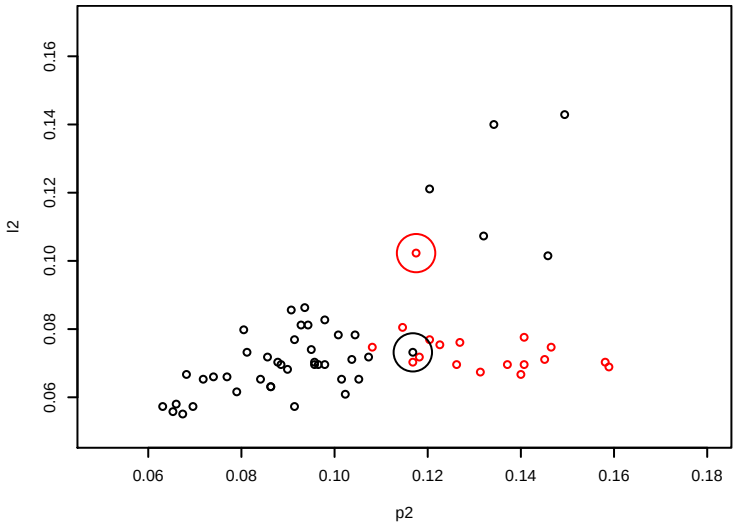


$\log(f_w)$ gegen $\log(f_m)$ und Diagonale, Ausschnittvergrößerung:

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



Welche Fälle wurden falsch zugeordnet?



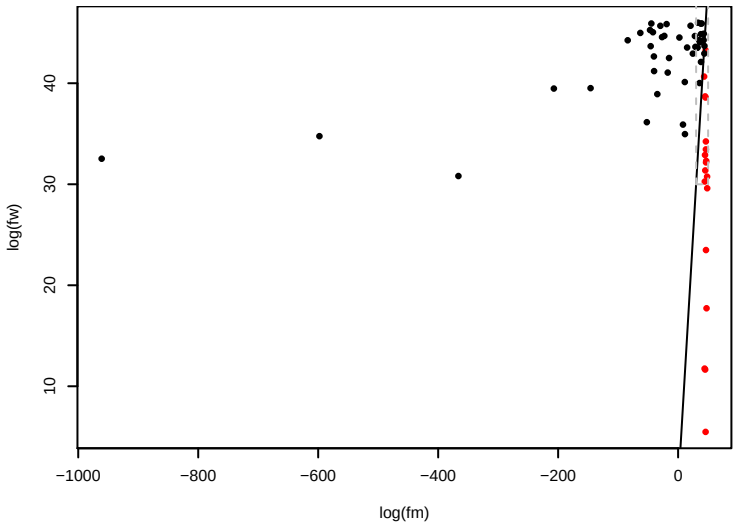
Wenn man nur p2 und l2 kennt, ist es sehr verständlich, dass diese Fälle falsch klassifiziert werden.

3.3 zehn Dimensionen

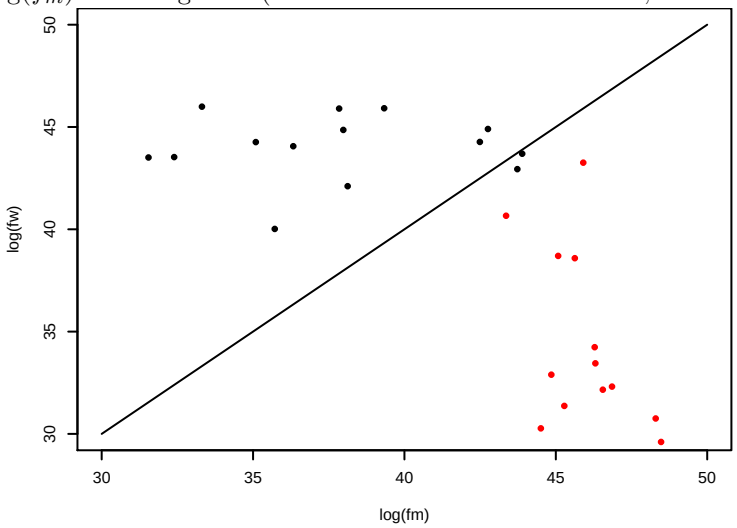
Wir verfahren genause mit allen Variablen (p1, p2, p3, p4, p5, l1, l2, l3, l4, l5) gemeinsam — mathematisch analog, allerdings geometrisch sehr schwierig darzustellen.

Ergebnis:

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

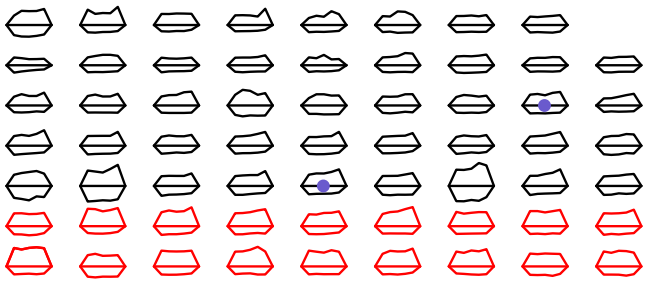


$\log(f_w)$ gegen $\log(f_m)$ und Diagonale (basierend auf allen 10 Variablen, Ausschnittvergrößerung):



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.

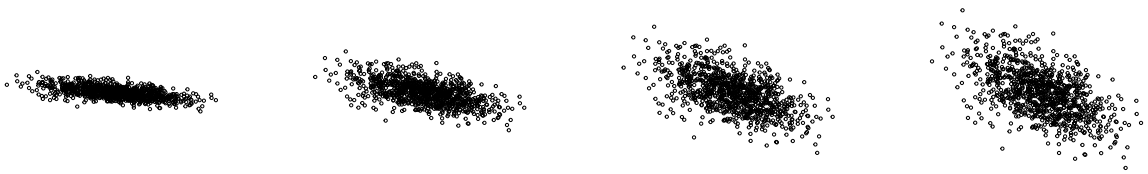
- Der Klassifikationsfehler könnte zu optimistisch geschätzt werden.
- Mögliche Lösungen: Schätze Klassifikationsfehler auf unabhängigen Daten oder Kreuzvalidierung.
- Dieser Effekt ist umso größer je mehr Variablen für die Klassifikation verwendet werden wegen Überanpassung, engl. *overfitting*.

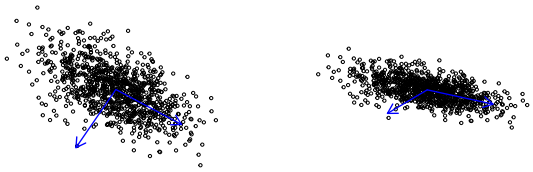
4 Hauptkomponentenanalyse (PCA)

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)

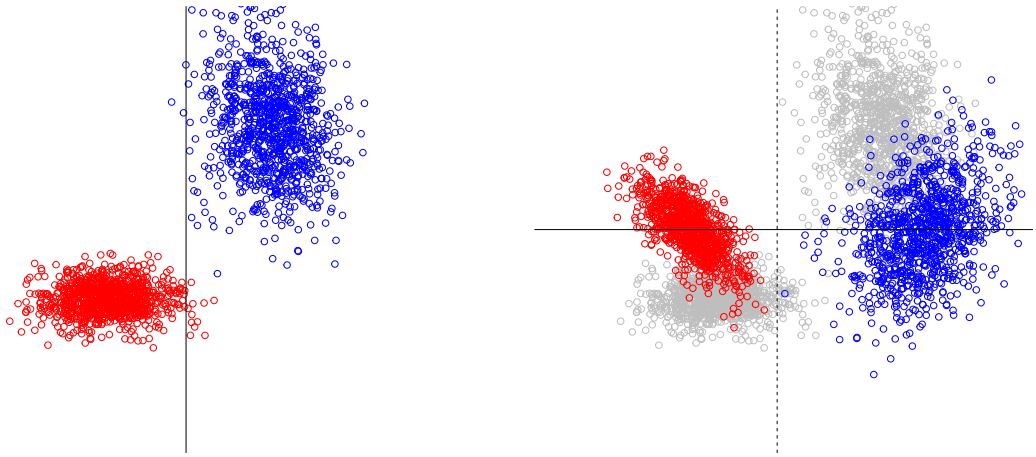




Um einen guten Blick auf die Daten zu haben wollen wir die Komponenten darstellen, die die meiste Variation beitragen.

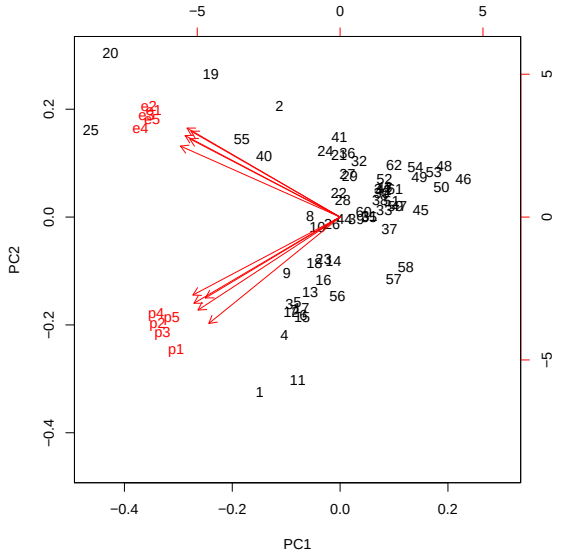
Die Achse mit der größten Variation wird in die x-Achse rotiert, die Achse mit der zweit größten Variation wird in die y-Achse rotiert.

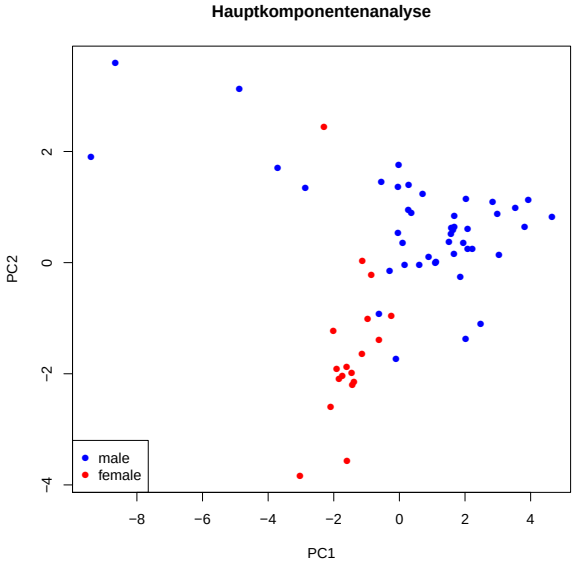
Beispiel: 2-dimensionale Daten



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

PCA für die Kleinspechtrufe





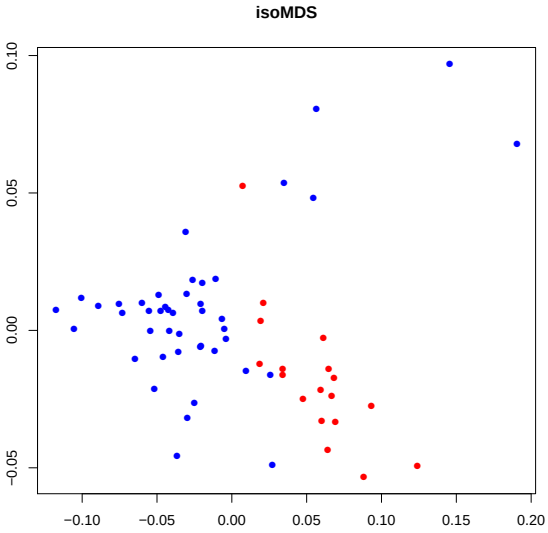
PCA für die Kleinspechtrufe

```
> pca$rotation
           PC1        PC2         PC3
p1 -0.2822046 -0.3934227  0.06673009
p2 -0.3142081 -0.3187375  0.21781852
p3 -0.3055310 -0.3438220  0.39481762
p4 -0.3164843 -0.2883903 -0.01334022
p5 -0.2895334 -0.2992765 -0.76584748
e1 -0.3205125  0.3176608 -0.29541900
e2 -0.3290219  0.3291692 -0.10654413
e3 -0.3332567  0.3012729  0.03026653
e4 -0.3431864  0.2625826  0.11413181
e5 -0.3232216  0.2893101  0.30488479

            PC4          PC5          PC6
p1  0.48824844  0.54477120  0.35844178
p2  0.37248001 -0.44405252 -0.21766281
p3 -0.14311992 -0.07383095 -0.31532958
p4 -0.72829058  0.30998300 -0.07417985
p5 -0.04742929 -0.37590860  0.20433598
e1  0.10981182  0.38559134 -0.23181948
e2  0.14627053  0.01080058 -0.19306531
e3  0.01495723 -0.15546815 -0.32355609
e4  0.01731206  0.10213805  0.16126133
e5 -0.18908978 -0.28247848  0.67508930

           PC7          PC8         PC9
p1 -0.01935566  0.305839783 -0.05971310
p2  0.51843943 -0.281675873 -0.09853561
p3 -0.58382109 -0.041285935  0.39513786
p4  0.37828512  0.022507247 -0.19687500
```

```
p5 -0.23342261  0.008714212  0.01370418
e1  0.13890983 -0.302606996  0.50720476
e2  0.10860831  0.120990852  0.01915132
e3 -0.09258674  0.647509138 -0.28340220
e4 -0.36541954 -0.538467391 -0.58072851
e5  0.13862775  0.096998416  0.34125903

          PC10
p1  0.002299543
p2 -0.105374067
p3  0.092215466
p4  0.046103449
p5 -0.004663970
e1 -0.355651192
e2  0.828524772
e3 -0.400023478
e4 -0.052783430
e5 -0.049940991
```

```
> pca$x[,1:3]
           PC1          PC2          PC3
1  -3.03118736 -3.836949298  0.57243527
2  -2.29104639  2.445940366  0.72045947
3  -1.90818565 -1.918234661  0.36236514
4  -2.09448178 -2.586937634  1.06487215
5  -1.60947269 -1.877439205  0.58648187
6  -1.38377685 -2.145820570 -0.29407050
7  -1.74059359 -2.042148100  0.44633190
8  -1.12504147  0.027894999  0.65728703
9  -2.01482060 -1.235413232  0.98584518
10 -0.85790979 -0.218484221  0.31909771
11 -1.59313759 -3.562727743 -0.06513507
12 -1.83852072 -2.089209406 -0.63314484
13 -1.13503074 -1.634330801 -0.28475954
14 -0.24859616 -0.952493266 -0.17989555
15 -1.43206135 -2.199854538 -0.34107832
16 -0.62209171 -1.386622917  0.38971370
17 -1.45181103 -1.992807142  0.20203001
18 -0.96753200 -1.009539876 -0.45545190
19 -4.87496964  3.134254777  0.83516149
20 -8.65940737  3.598744777 -1.07592369
21 -0.04054120  1.367898098 -0.39474360
22 -0.04178728  0.530202105 -0.70114424
23 -0.61952138 -0.915773298 -0.61614360
24 -0.55147698  1.457954699  0.01367081
25 -9.39969100  1.907459184 -0.63011106
26 -0.29565292 -0.142951087 -0.95619024
27  0.27937402  0.951000599 -0.08546215
28  0.10244905  0.362558330 -1.06868312
29  0.36578656  0.892760681 -0.53464476
30  1.57819294  0.521129767  0.03659239
31  1.11387202  0.014438176 -0.66999184
32  0.71431633  1.234477107 -0.70350063
33  1.66890799  0.156868442 -0.53095225
34  1.58637653  0.620291656  0.03903341
35  1.09325119 -0.006766015 -0.58717439
36  0.28362879  1.404990503  0.05550391
37  1.86096484 -0.252900633 -0.16981464
38  1.51138174  0.367027447  0.19599092
39  0.61238840 -0.041842184 -0.18129817
40 -2.86100061  1.343826138  1.41617944
41 -0.01960435  1.752542648  0.98797651
42  1.63264811  0.586543903 -0.13813615
43  1.68109381  0.652043259  0.53438917
44  0.16398811 -0.040304081 -0.14938059
45  3.04062674  0.146458892 -0.38521007
46  4.65593306  0.824050945 -0.14932694
47  2.22850201  0.249900140  0.48911699
48  3.92700625  1.125514262  0.02783022
49  2.99083426  0.877287974 -0.04849626
50  3.82424810  0.648418430  0.91628634
51  1.94602458  0.354049439 -0.08820379
52  1.67947317  0.833654206  0.11141847
53  3.53948039  0.979694233  0.29048270
54  2.84017729  1.096855332  0.24397110
55 -3.70637464  1.700057316  0.12296383
56 -0.09829316 -1.725842328 -1.01034234
57  2.02500754 -1.366958758  0.30145712
58  2.47481026 -1.094673138 -1.23801573
59  2.08529515  0.247482514 -0.21332675
60  0.89203474  0.103470181  1.14875794
61  2.07875245  0.608927029  0.38512280
62  2.03679155  1.150355579  0.12092772
```

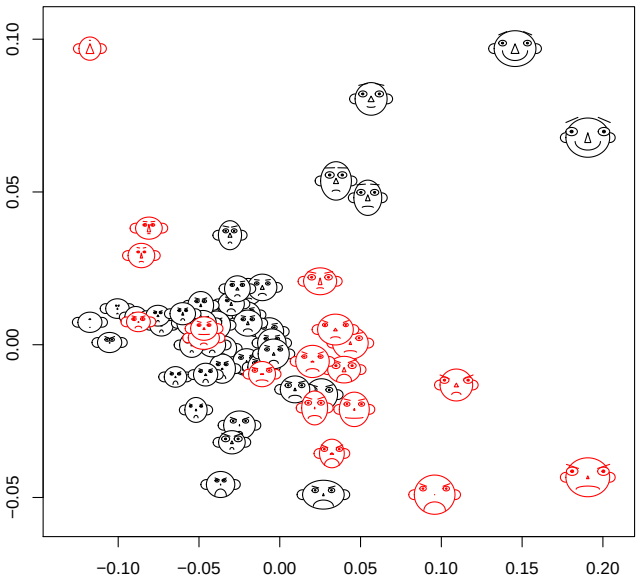
Alternative: Multidimensionale Skalierung



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")
```


0.10
0.05
0.00
−0.05
−0.10
−0.05
0.00
0.05
0.10
0.15
0.20