Robuste Statistik

Musterlösung zu Blatt 9

Aufgabe 9.1:

Lade zunächst die benötigten Pakete. Das erste ist für die Berechnung der Tiefen, das zweite ist eine Grafik-Paket, welches ein Grafiktool in `R` liefert.

```
library(mrfDepth)
library(ggplot2)
```

a) Erstelle nun die Daten und das Gitter. Bei dem Gitter ist wichtig, dass alle Punkte des Gitters im Originaldatensatz vorhanden sind und dass das Gitter hinreichend fein ist.

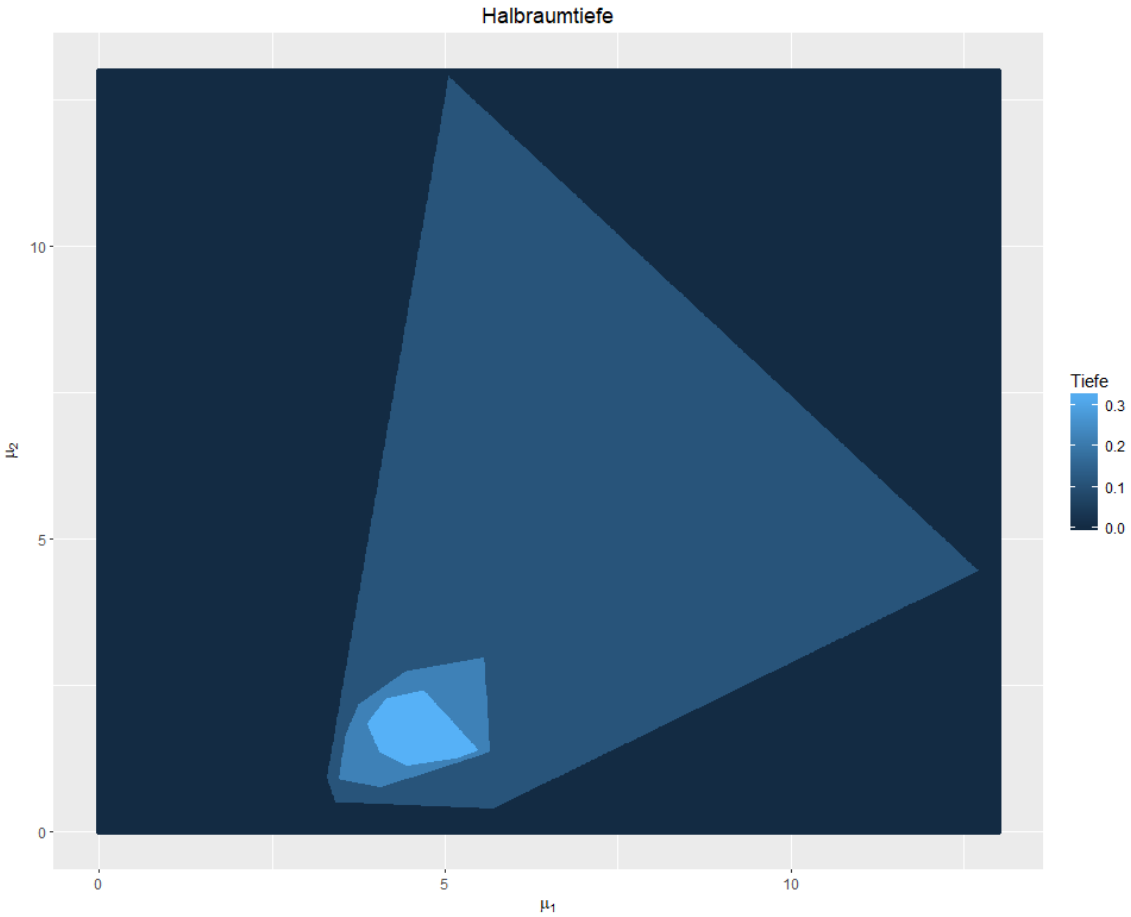
```
# Daten
data <- list(
  y1 = c(3.43, 0.55),
  y2 = c(3.77, 2.22),
  y3 = c(5.56, 1.36),
  y4 = c(4.07, 1.4),
  y5 = c(5.07, 12.95),
  y6 = c(5.72, 0.44),
  y7 = c(4.46, 2.79),
  y8 = c(12.73, 4.5),
  y9 = c(3.31, 0.97)
)

# Gitter
grid <- expand.grid(mu1 = seq(0, 13, 0.01), mu2 = seq(0, 13, 0.01))
```

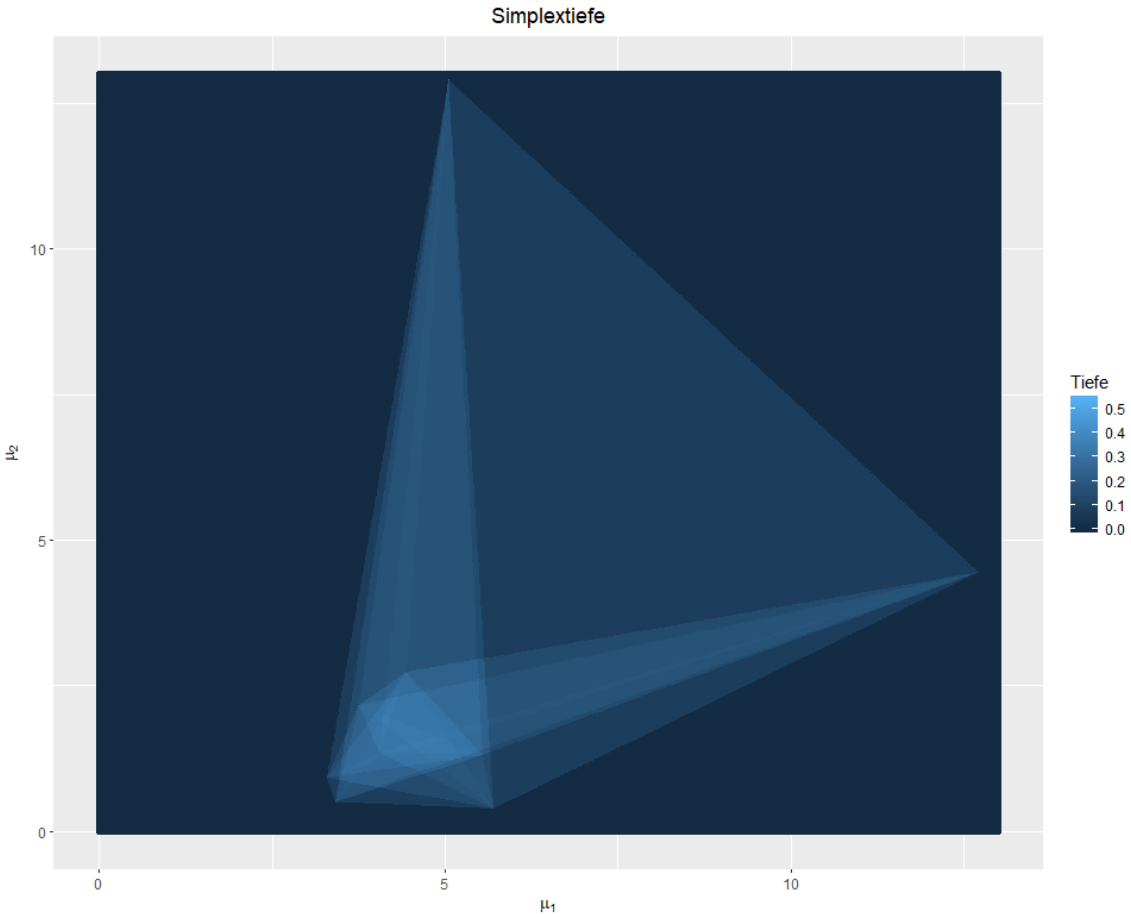
Hinreichend fein heißt, dass das Gitter mindestens die zweite Nachkommastelle berücksichtigt, da der Datensatz mit dieser Genauigkeit arbeitet und die minimalen und maximalen Werte des Datenbereiches in beiden Koordinaten berücksichtigt (1 Punkt). Es ist aber auch legitim zu begründen, dass man lediglich die erste Nachkommastelle betrachtet, da der Informationsverlust nicht sehr groß ist und wir so deutlich Rechenzeit bei der Erstellung der Grafiken sparen.

Berechne nun die Tiefen für jeden Punkt des Gitters. Dafür wird das Paket `mrfDepth` verwendet. Das Paket `depth` weißt Schwächen auf, da es z.B. bei der Berechnung der Simplex-Tiefe zu falschen Ergebnissen kommen kann. (Danke nochmal für die Info, Markus!) Die Funktion `hdepth()` liefert dabei die Halbraum-Tiefe und `sdepth()` liefert die Simplex-Tiefe (1 Punkt).

```
# Halbraum- und Simplex-Tiefe
res_halbraum <- hdepth(t(as.data.frame(data)), grid)$depthZ
res_simplex <- sdepth(t(as.data.frame(data)), grid)$depthZ
```

Stellt man die Ergebnisse nun mithilfe von `ggplot2` grafisch dar, erhält man folgende Grafiken (2 Punkte):





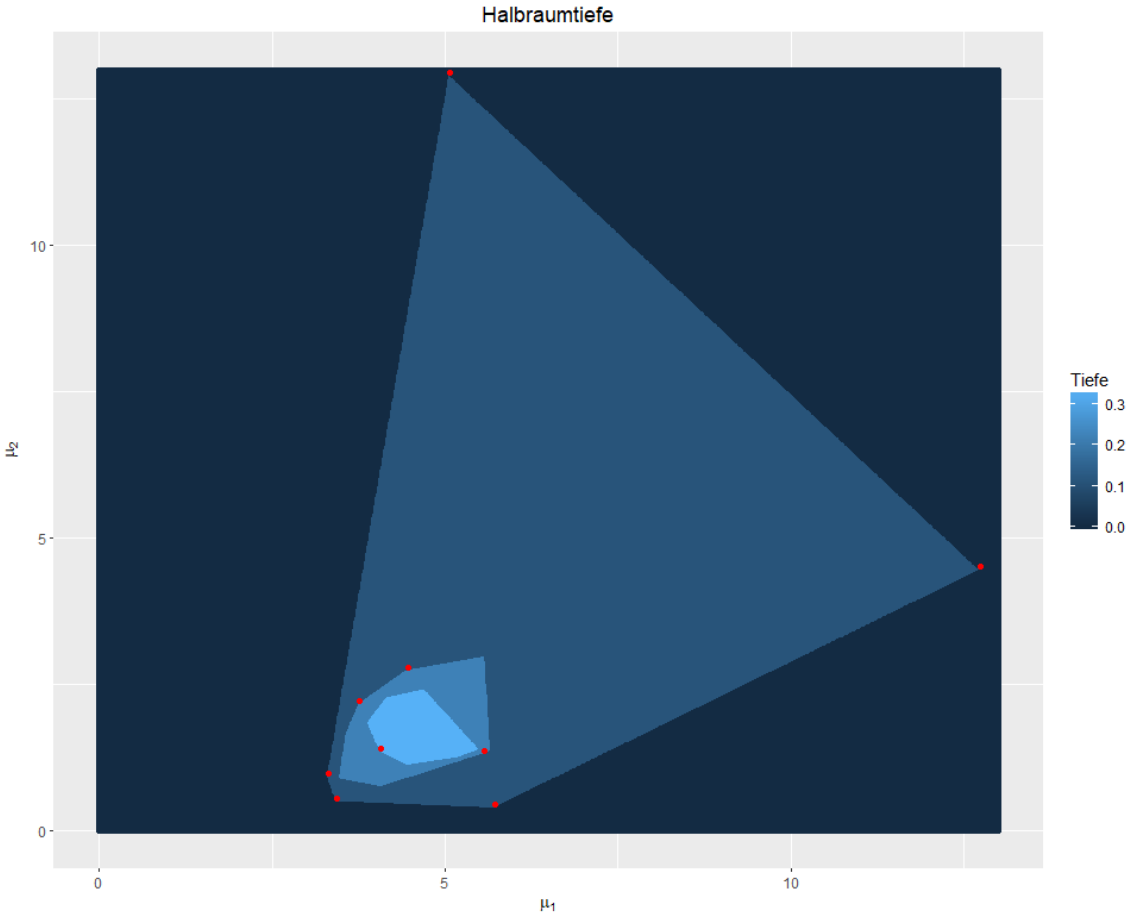
Vorsicht: Es kann zu langen Ladezeiten kommen, falls man das Gitter zu fein rastert, sodass man gegebenenfalls die Grafikdateien komprimieren sollte. Ich habe sie nicht über RStudio in einer pdf-Datei, sondern als png-Datei abgespeichert. Dadurch habe statt 63,5 MB (bei pdf) nur 10 KB (bei png) Speicherplatz benötigt. Dies sollte man generell berücksichtigen, wenn man umfangreiche Grafiken wie Spektrogramme o.ä. produziert!

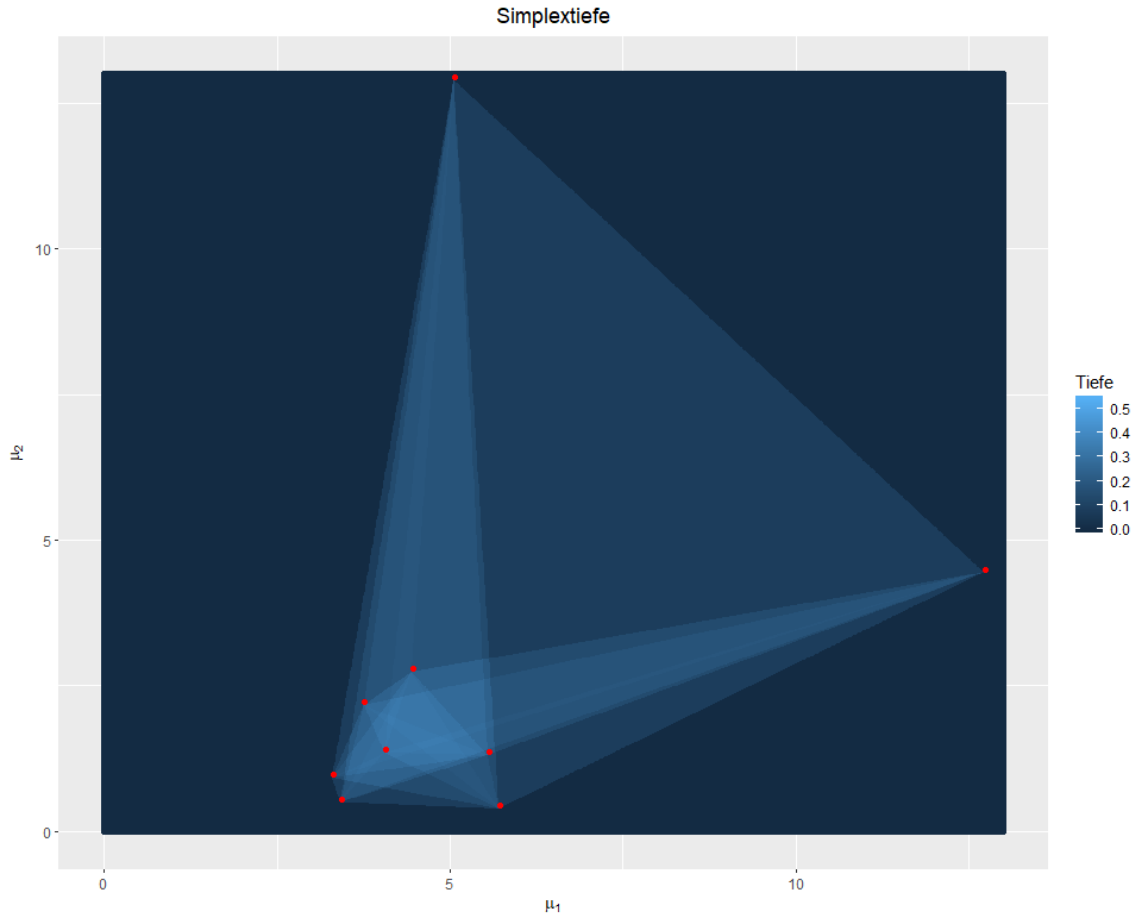
In beiden Grafiken ist die Lage der Datenpunkte gut zu erkennen. Während bei der Halbraum-Tiefe nur wenige verschiedene Tiefen-Werte in gewissen Plateau-Levels angenommen werden, ist dies bei der Simplex-Tiefe deutlich differenzierter. Hier sind beispielsweise auch die Verbindungsgeraden zwischen einzelnen Datenpunkten und die für Simplex-Tiefen typische sternförmige Struktur erkennbar.

b) Für die Berechnung der Tiefe der originalen Datenpunkte können wir erneut die oben genannten Funktionen verwenden (1 Punkt).

```
hdepth(t(as.data.frame(data)))$depthZ
# 0.1111111 0.2222222 0.2222222 0.3333333 0.1111111 0.1111111 0.2222222 0.1111111 0.1111111
sdepth(t(as.data.frame(data)))$depthZ
#  0.3333333 0.4523810 0.4523810 0.5476190 0.3333333 0.3333333 0.4761905 0.3333333 0.3333333
```

Datenpunkte mit höheren Tiefen sind anschaulich ausgedrückt tiefer in der Datenwolke. Der Begriff *Tiefe* wird durch beide Definitionen unterschiedlich charakterisiert. Die nächste Grafik soll die Ergebnisse noch im Detail veranschaulichen.

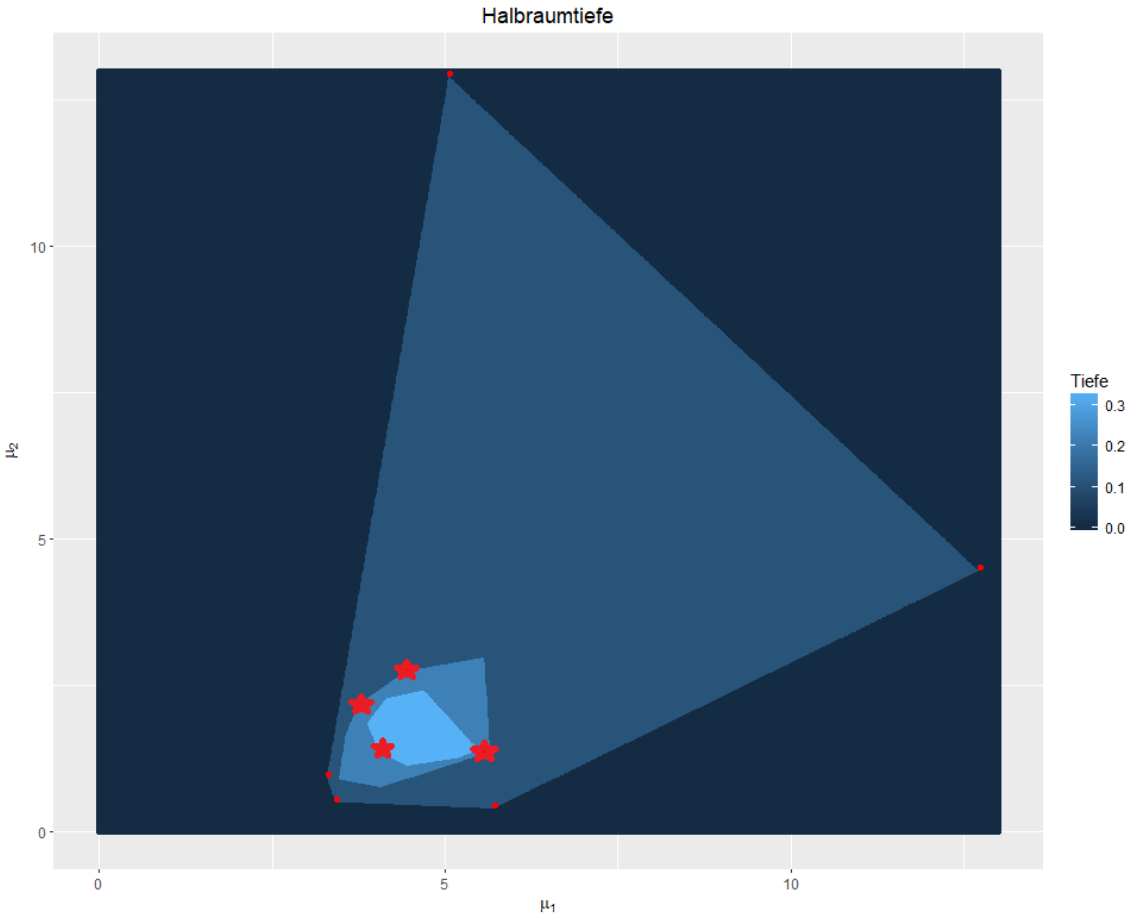


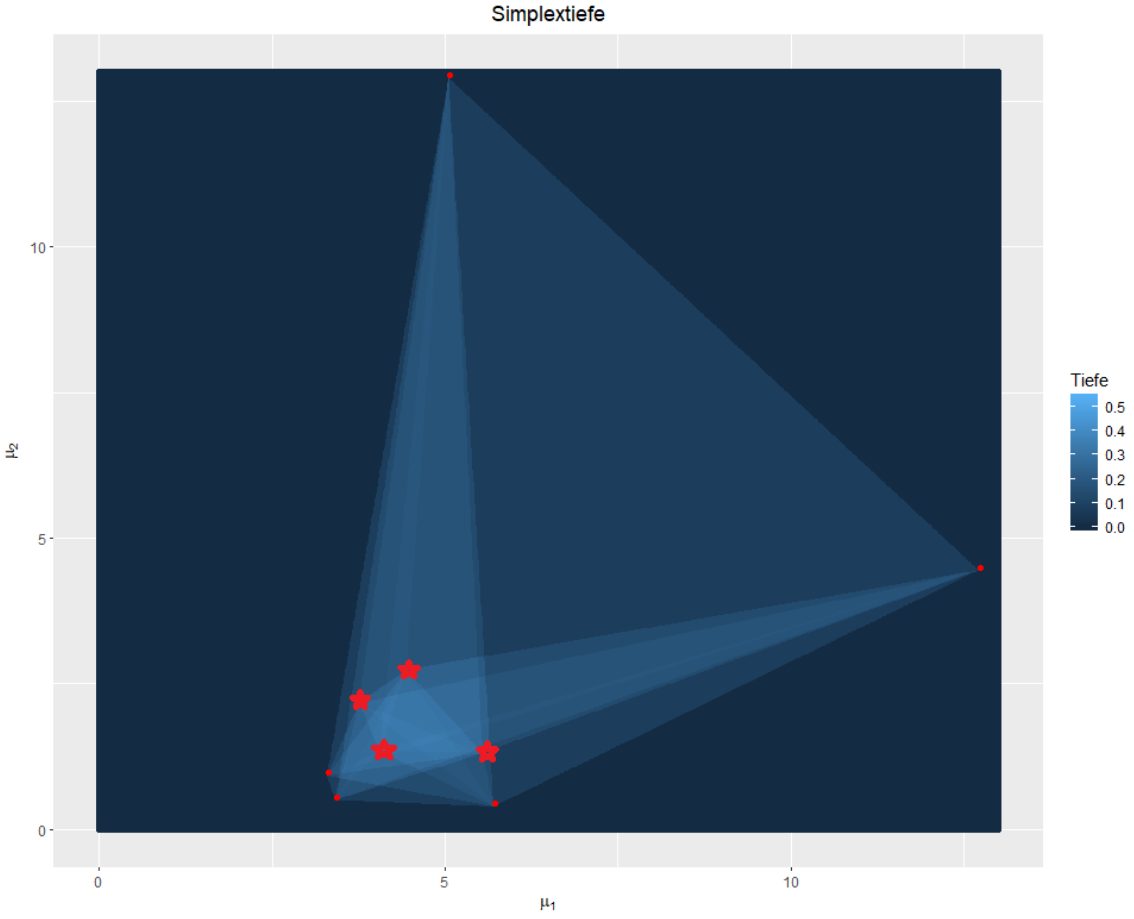


Eine andere Möglichkeit die Aufgabe zu lösen, ist folgende Vorgehensweise mittels einer Kreuzvalidierung. Bei dieser Methode bestimmen wir von y_i die Tiefe bzgl. $(y_j)_{j\in\mathfrak{J}}$ mit $\mathfrak{J} = \{1,\dots,9\}\setminus\{i\}$. Dadurch ist die Tiefe von y_i nicht so stark an y_i angepasst, sodass sich auch Tiefen mit dem Wert Null ergeben können. Dieses *Train-and-Test-Prinzip* (oder statt *Train* auch *Learn*) sollte man generell im Hinterkopf haben, um Überanpassungen zu vermeiden. Hierbei ist $(y_j)_{j\in\mathfrak{J}}$ der Lern-Datensatz und y_i ein Testdatum.

```
hdepthcross <- numeric(9)
sdepthcross <- numeric(9)
for(i in 1:9){
hdepthcross[i] <- hdepth(t(X[,-i]), t(X[,i]))$depthZ
sdepthcross[i] <- sdepth(t(X[,-i]), t(X[,i]))$depthZ
}
hdepthcross
# 0.000 0.125 0.125 0.250 0.000 0.000 0.125 0.000 0.000
round(sdepthcross, digits = 3)
# 0.000 0.179 0.179 0.321 0.000 0.000 0.214 0.000 0.000
```

Die Datenpunkte als Sterne dargestellt haben für beide Tiefebegriffe bei der Kreuzvalidierung noch eine positive Tiefe. Das sind die Datenpunkte, die nicht am Rand der konvexen Hülle der Daten liegen.

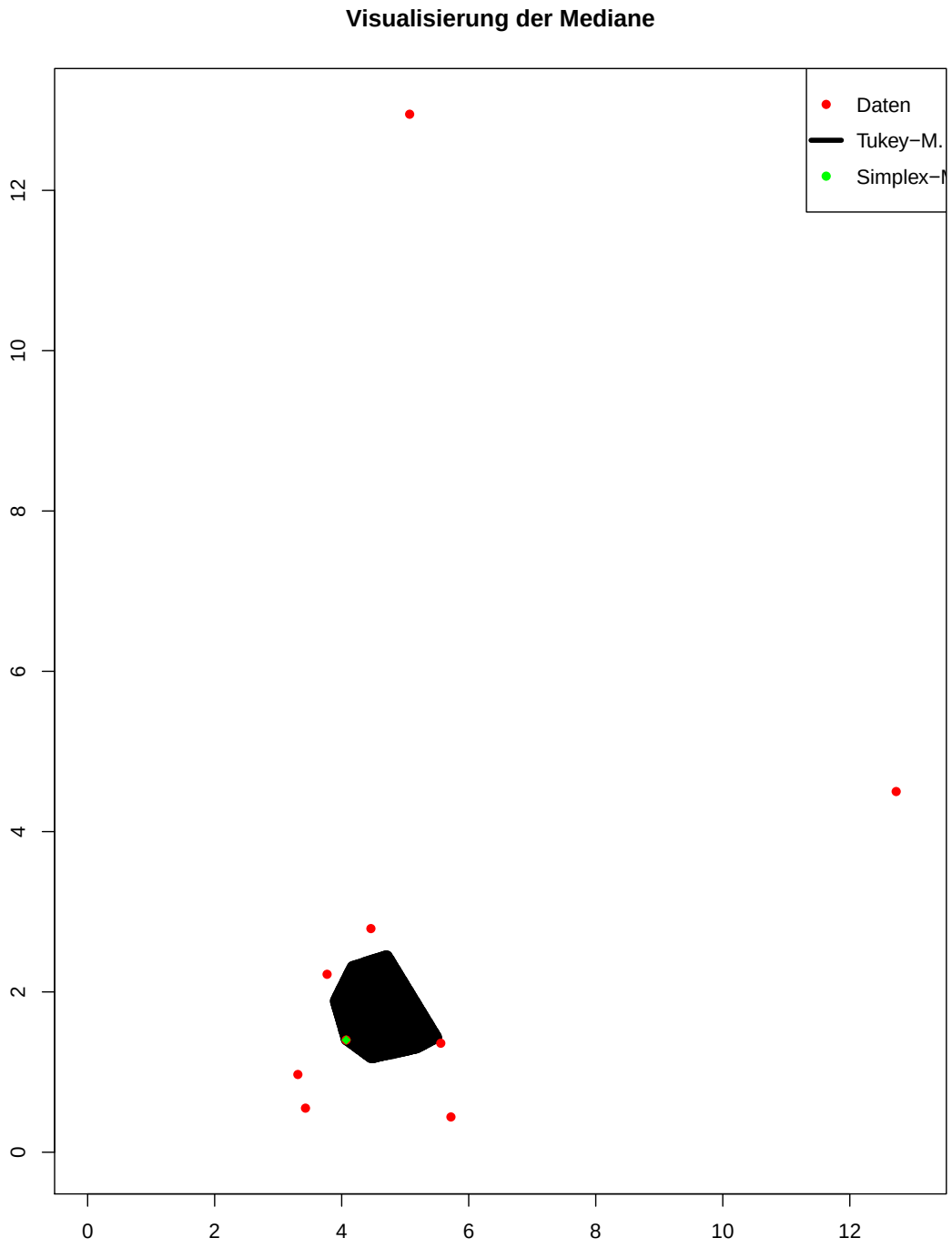




c) Wir ermitteln nun die Punkte mit maximaler Tiefe, also den Halbraum-Median (oder Tukey-Median) und den Simplex-Median (1 Punkt):

```
hs_median <- cbind(grid, res_halbraum)[which(res_halbraum == max(res_halbraum)), ]
sim_median <- cbind(grid, res_simplex)[which(res_simplex == max(res_simplex)), ]
```

Es zeigt sich, dass der Simplex-Median eindeutig ist, während der Halbraum-Median eine Fläche ist. Diese Fläche lässt sich mit etwas Aufwand identifizieren, in dem man von passenden Schnittpunkten von Verbindungsstrecken zwischen den Datenpunkten die Konvexkombination bestimmt.



Die Werte der maximalen Tiefen sind jedoch unterschiedlich zwischen der Halbraum-Tiefe und der Simplex-Tiefe. Der Wert der Simplex-Tiefe ist mit 0.55 größer als der Wert der Halbraum-Tiefe mit 0.33. Dies ist aber nicht aussagekräftig, da die Werte der Tiefen zwischen den beiden Verfahren nicht miteinander vergleichbar sind, da sie je nach gegebenen Daten auch unterschiedliche Maximalwerte erreichen können.

Aufgabe 9.2:

Gegeben sind folgende Punkte:

$$\boldsymbol{y}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \boldsymbol{y}_2 = \begin{pmatrix} 4 \\ 7 \end{pmatrix}, \quad \boldsymbol{y}_3 = \begin{pmatrix} 6 \\ 9 \end{pmatrix}, \quad \boldsymbol{y}_4 = \begin{pmatrix} 5 \\ 14 \end{pmatrix}, \quad \boldsymbol{y}_5 = \begin{pmatrix} 10 \\ 5 \end{pmatrix}.$$

Hierfür soll per Hand die Simplex-Tiefe für gegebene Punkte $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_8$ des $\mathbb{R}^2$ berechnet werden. Ursprünglich sollte in dieser Aufgabe die Simplex-Tiefe für alle Punkte aus $\mathbb{R}^2$ bestimmt werden, was was im nachfolgenden zunächst präsentiert wird.

Es werden die Punkte zunächst in den $\mathbb{R}^2$ gezeichnet und miteinander verbunden:

Nun muss man sich überlegen, welche unterschiedlichen Fälle man bei der Berechnung der verschiedenen Tiefen beachten muss. Dies sind:

- Das innere jedes Dreiecks (Punkte 1 bis 7 in obiger Grafik),
- Alle Punkte des $\mathbb{R}^2$, welche nicht mindestens in einem Simplex enthalten sind (Punkt 8 in obiger Grafik),
- Die Punkte $\boldsymbol{y}_1$ bis $\boldsymbol{y}_5$ (Punkte 9 bis 13 in obiger Grafik),
- Schnittpunkte zwischen Geraden (Punkt 14 in obiger Grafik),
- Die Verbindungsgeraden zwischen den einzelnen Punkten (Punkte 15 bis 26 in obiger Grafik).

Die Punkte und Geraden müssen deshalb gesondert betrachtet werden, weil sie jeweils auf den Rand mehrerer Simplizes liegen und somit zu mehreren Simplizes gehören.

Um diese 26 Mengen nun mathematisch voneinander abgrenzen zu können, stellen wir zunächst Geradengleichungen der einzelnen Verbindungslinien auf und berechnen den einen Schnittpunkt:

Die Geradengleichungen bekommt man schnell durch die Zweipunkteform. Im folgenden beschreibt $f_{ij}(x)$ die Gerade, welche durch $\boldsymbol{y}_i$ und $\boldsymbol{y}_j$ geht.

$$f_{12}(x) = \frac{7}{3}x - \frac{7}{3}, \quad f_{13}(x) = \frac{9}{5}x - \frac{9}{5}, \quad f_{14}(x) = \frac{7}{2}x - \frac{7}{2}, \quad f_{15}(x) = \frac{5}{9}x - \frac{5}{9}, \quad f_{23}(x) = x + 3,$$

$$f_{24}(x) = 7x - 21, \quad f_{25}(x) = -\frac{1}{3}x + \frac{25}{3}, \quad f_{34}(x) = -5x + 39, \quad f_{35}(x) = -x + 15, \quad f_{45}(x) = -\frac{9}{5}x + 23$$

Der Schnittpunkt errechnet sich nun durch Gleichsetzen von $f_{25}(x)$ und $f_{34}(x)$:

$$\begin{aligned} & \frac{9}{5}x - \frac{9}{5} = -\frac{1}{3} + \frac{25}{3} \\ \Leftrightarrow \quad & 27x - 27 = -5x + 125 \\ \Leftrightarrow \quad & 32x = 152 \\ \Leftrightarrow \quad & x = \frac{19}{4} \end{aligned}$$

Und somit gilt: $y = \frac{9}{5} \cdot \frac{19}{4} - \frac{9}{5} = \frac{27}{4}$. Der Schnittpunkt ist also $\begin{pmatrix} \frac{19}{4} \\ \frac{27}{4} \end{pmatrix}$.

Bestimme nun die 26 verschiedenen Mengen und nummeriere diese durch, genauso wie es auf obiger Grafik gemacht wurde:

1: $\left\{ \begin{pmatrix} x \\ y \end{pmatrix} : y > \frac{7}{3}x - \frac{7}{3} \wedge y < \frac{7}{2}x - \frac{7}{2} \wedge y > 7x - 21 \right\}$

2: $\left\{ \begin{pmatrix} x \\ y \end{pmatrix} : y < \frac{7}{3}x - \frac{7}{3} \wedge y > \frac{9}{5}x - \frac{9}{5} \wedge y < -\frac{1}{3}x + \frac{25}{3} \right\}$

3: $\left\{ \begin{pmatrix} x \\ y \end{pmatrix} : y < \frac{9}{5}x - \frac{9}{5} \wedge y > \frac{5}{9}x - \frac{5}{9} \wedge y < -\frac{1}{3}x + \frac{25}{3} \right\}$

4: $\left\{ \begin{pmatrix} x \\ y \end{pmatrix} : y > x + 3 \wedge y < 7x - 21 \wedge y < -5x + 39 \right\}$

5: $\left\{ \begin{pmatrix} x \\ y \end{pmatrix} : y < x + 3 \wedge y > \frac{9}{5}x - \frac{9}{5} \wedge y > -\frac{1}{3}x + \frac{25}{3} \right\}$

6: $\left\{ \begin{pmatrix} x \\ y \end{pmatrix} : y < \frac{9}{5}x - \frac{9}{5} \wedge y > -\frac{1}{3}x + \frac{25}{3} \wedge y < -x + 15 \right\}$

7: $\left\{ \begin{pmatrix} x \\ y \end{pmatrix} : y > -5x + 39 \wedge y > -x + 15 \wedge y < -\frac{9}{5}x + 23 \right\}$

8: $\left\{ \begin{pmatrix} x \\ y \end{pmatrix} : y > \frac{7}{2}x - \frac{7}{2} \wedge y < \frac{5}{9}x - \frac{5}{9} \wedge y > -\frac{9}{5}x + 23 \right\}$

9: $\left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right\}$

10: $\left\{ \begin{pmatrix} 4 \\ 7 \end{pmatrix} \right\}$

11: $\left\{ \begin{pmatrix} 6 \\ 9 \end{pmatrix} \right\}$

12: $\left\{ \begin{pmatrix} 5 \\ 14 \end{pmatrix} \right\}$

13: $\left\{ \begin{pmatrix} 10 \\ 5 \end{pmatrix} \right\}$

14: $\left\{ \begin{pmatrix} \frac{19}{4} \\ \frac{27}{4} \end{pmatrix} \right\}$

15: $\left\{ \begin{pmatrix} x \\ y \end{pmatrix} : x \in (1,\, 4) \wedge y = \frac{7}{3}x - \frac{7}{3} \right\}$

16: $\left\{ \begin{pmatrix} x \\ y \end{pmatrix} : x \in \left(1,\, \frac{19}{4}\right) \wedge y = \frac{9}{5}x - \frac{9}{5} \right\}$

17: $\left\{ \begin{pmatrix} x \\ y \end{pmatrix} : x \in \left(\frac{19}{4},\, 6\right) \wedge y = \frac{9}{5}x - \frac{9}{5} \right\}$

18: $\left\{ \begin{pmatrix} x \\ y \end{pmatrix} : x \in (1,\, 5) \wedge y = \frac{7}{2}x - \frac{7}{2} \right\}$

19: $\left\{ \begin{pmatrix} x \\ y \end{pmatrix} : x \in (1,\, 10) \wedge y = \frac{5}{9}x - \frac{5}{9} \right\}$

20: $\left\{ \begin{pmatrix} x \\ y \end{pmatrix} : x \in (4,\, 6) \wedge y = x + 3 \right\}$

21: $\left\{ \begin{pmatrix} x \\ y \end{pmatrix} : x \in (4,\, 5) \wedge y = 7x - 21 \right\}$

22: $\left\{ \begin{pmatrix} x \\ y \end{pmatrix} : x \in \left(4,\, \frac{19}{4}\right) \wedge y = -\frac{1}{3}x + \frac{25}{3} \right\}$

23: $\left\{ \begin{pmatrix} x \\ y \end{pmatrix} : x \in \left(\frac{19}{4},\, 10\right) \wedge y = -\frac{1}{3}x + \frac{25}{3} \right\}$

24: $\left\{ \begin{pmatrix} x \\ y \end{pmatrix} : x \in (5,\, 6) \wedge y = -5x + 39 \right\}$

25: $\left\{ \begin{pmatrix} x \\ y \end{pmatrix} : x \in (6,\, 10) \wedge y = -x + 15 \right\}$

26: $\left\{ \begin{pmatrix} x \\ y \end{pmatrix} : x \in (5,\, 10) \wedge y = -\frac{9}{5}x + 23 \right\}$

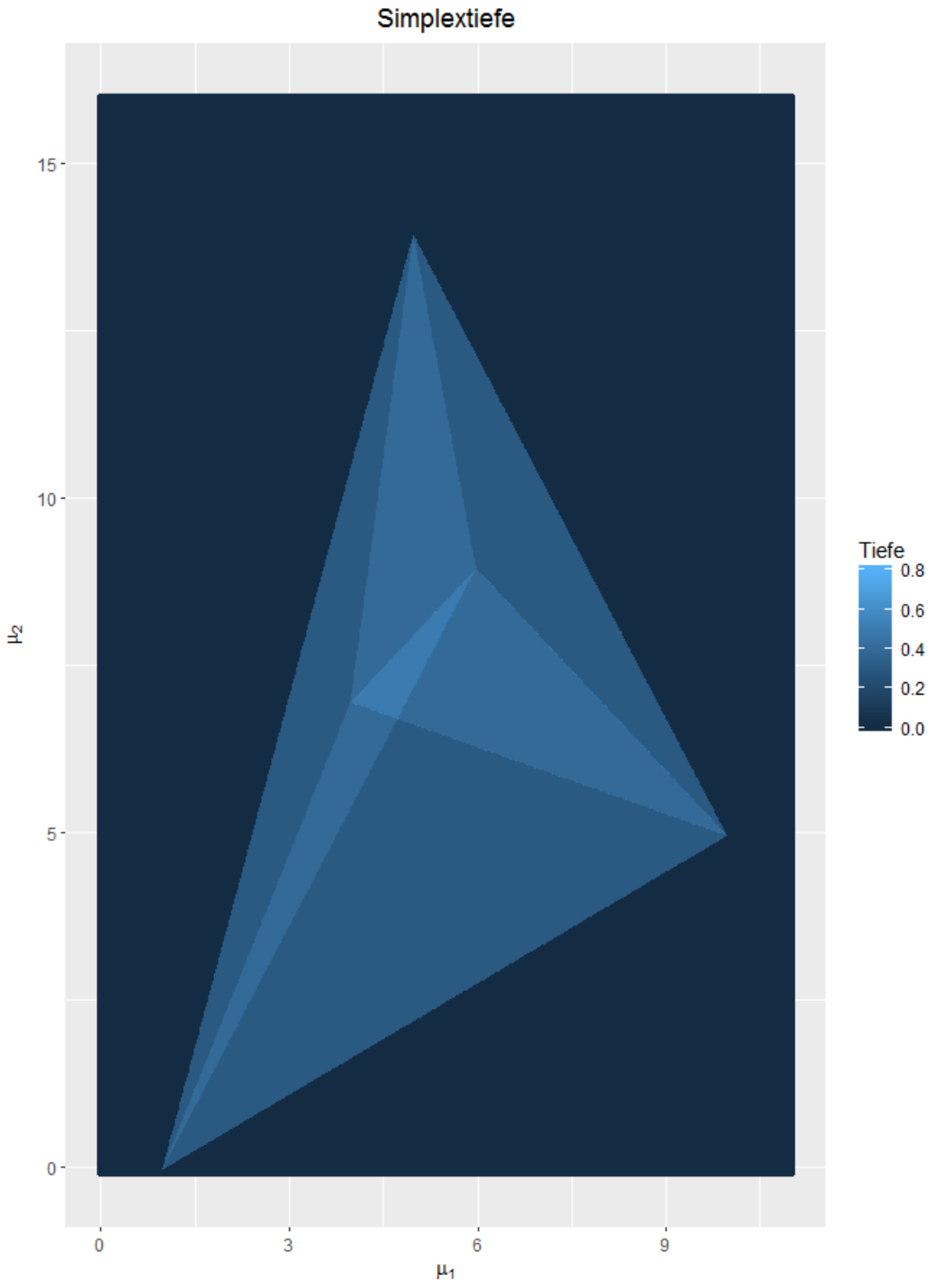
Nun können wir am besten tabellarisch gucken, welche der 26 Mengen von welchen aufgespannten Simplizes erfasst wird. Dazu seien in folgender Tabelle in den Zeilen die 26 durchnummerierten Mengen geschrieben und in den Spalten die aufgespannten Simplizes:

	y_1 y_2 y_3	y_1 y_2 y_4	y_1 y_2 y_5	y_1 y_3 y_4	y_1 y_3 y_5	y_1 y_4 y_5	y_2 y_3 y_4	y_2 y_3 y_5	y_2 y_4 y_5	y_3 y_4 y_5	Tiefe
1	✗	✓	✗	✓	✗	✓	✗	✗	✗	✗	0.3
2	✓	✗	✓	✓	✗	✓	✗	✗	✗	✗	0.4
3	✗	✗	✓	✗	✓	✓	✗	✗	✗	✗	0.3
4	✗	✗	✗	✓	✗	✓	✓	✗	✓	✗	0.4
5	✓	✗	✗	✓	✗	✓	✗	✓	✓	✓	0.5
6	✗	✗	✗	✗	✓	✓	✗	✓	✓	✗	0.4
7	✗	✗	✗	✗	✗	✓	✗	✗	✓	✓	0.3
8	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	0
9	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	0.6
10	✓	✓	✓	✓	✗	✓	✓	✓	✓	✗	0.8
11	✓	✗	✗	✓	✓	✓	✓	✓	✓	✓	0.8
12	✗	✓	✗	✓	✗	✓	✓	✗	✓	✓	0.6
13	✗	✗	✓	✗	✓	✓	✗	✓	✓	✓	0.6
14	✓	✗	✓	✓	✓	✓	✗	✓	✓	✗	0.7
15	✓	✓	✓	✓	✗	✓	✗	✗	✗	✗	0.5
16	✓	✗	✓	✓	✓	✓	✗	✗	✗	✗	0.5
17	✓	✗	✗	✓	✓	✓	✗	✓	✓	✗	0.6
18	✗	✓	✗	✓	✗	✓	✗	✗	✗	✗	0.3
19	✗	✗	✓	✗	✓	✓	✗	✗	✗	✗	0.3
20	✓	✗	✗	✓	✗	✓	✓	✓	✓	✗	0.6
21	✗	✓	✗	✓	✗	✓	✓	✗	✓	✗	0.5
22	✓	✗	✓	✓	✗	✓	✗	✓	✓	✗	0.6
23	✗	✗	✓	✗	✓	✓	✗	✓	✓	✗	0.5
24	✗	✗	✗	✓	✗	✓	✓	✗	✓	✓	0.5
25	✗	✗	✗	✗	✓	✓	✗	✓	✓	✓	0.5
26	✗	✗	✗	✗	✗	✓	✗	✗	✓	✓	0.3

Es ist zu erkennen, dass die Tiefe 0.8 die maximale Tiefe ist. An den Punkten mit Tiefe 0.8 liegt also der Simplex-Median. Die Menge der Simplex-Mediane ist somit

$$\left\{ \begin{pmatrix} 4 \\ 7 \end{pmatrix}, \begin{pmatrix} 6 \\ 9 \end{pmatrix} \right\}$$

Berechnet man die Tiefen mit `R` und der Funktion `sdepth`, so zeigt sich folgendes Bild:



Wir haben einen lediglich einen Spezialfall dieser Aufgabe durchgerechnet. Man kann die gegebenen Punkte $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_8$ mit jeweils einem der 26 Fälle identifizieren, wie es die nachfolgende Tabelle zeigt (je 0.5 Punkte pro Punkt).

Punkt $\boldsymbol{p}_i$ mit i = ...	1	2	3	4	5	6	7	8
gehört zum Fall	9	18	8	1	6	20	5	10
mit einer Tiefe von	0.6	0.3	0	0.3	0.4	0.6	0.5	0.8