1. Übungsblatt

Ksenia Klassen
ksenia.klassen@udo.edu

Dag-Björn Hering
dag.hering@udo.edu

Henning Ptaszyk
henning.ptaszyk@udo.edu

29. November 2016

1 Aufgabe 1

1.1 a)

1.2 b)

1.3 c)

1.4 d)

2 Aufgabe2

2.1 a)

In der Abbildung 1 sind die beiden Populationen $P0$ und $P1$ sowie die drei Projektionsgeraden:

$$g_1(x) = 0 \tag{1}$$

$$g_2(x) = -\frac{3}{4}x \tag{2}$$

$$g_3(x) = -\frac{5}{4}x \tag{3}$$

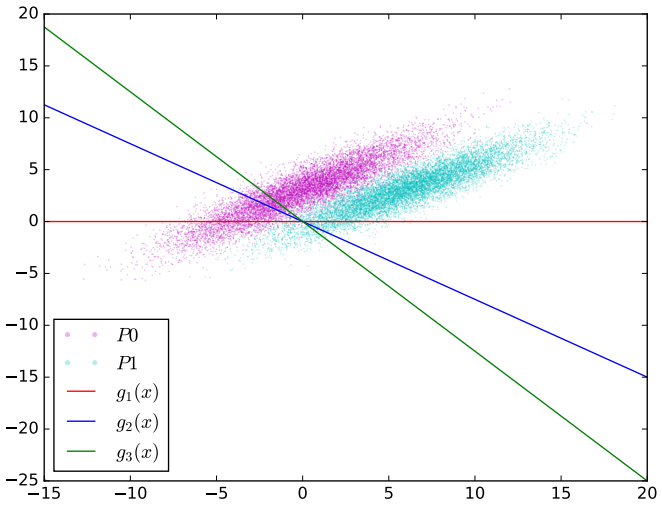


Abbildung 1: Zweidimensionaler Scatterplot der Populationen und die Projektionsgeraden.

2.2 b)

Um die Populationen $P0$ und $P1$ jeweils auf die Geraden zu projezieren, muss zu nächst der Richtungsvektor der Geraden bestimmt und normiert werden.

Richtungsvektor $g_1(x)$:

$$\vec{v}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}. \tag{4}$$

Dieser Vektor ist schon normiert, folglich muss nur noch die Richtung umgedreht werden damit die projizierte Population $P0$ rechts von der Population $P1$ liegt.

$$\Rightarrow 1 = \begin{pmatrix} -1 \\ 0 \end{pmatrix} \tag{5}$$

Richtungsvektor $g_2(x)$:

$$\vec{v}_2 = \begin{pmatrix} 1 \\ -\frac{3}{4} \end{pmatrix}. \tag{6}$$

Normierung:

$$\left| n \cdot \begin{pmatrix} 1 \\ -\frac{3}{4} \end{pmatrix} \right| \stackrel{!}{=} 1 \tag{7}$$

$$n\sqrt{1^2 + \left(\frac{3}{4}\right)^2} \stackrel{!}{=} 1 \tag{8}$$

$$n = \frac{4}{5} \tag{9}$$

$$\Rightarrow \vec{v}_2 = \begin{pmatrix} \frac{4}{5} \\ -\frac{6}{10} \end{pmatrix} \tag{10}$$

Wegen der Reihenfolge muss der Vektor wieder umgedreht werden.

$$\Rightarrow \vec{v}_2 = \begin{pmatrix} -\frac{4}{5} \\ \frac{6}{10} \end{pmatrix} \tag{11}$$

Richtungsvektor $g_3(x)$:

$$\vec{v}_3 = \begin{pmatrix} 1 \\ -\frac{5}{4} \end{pmatrix}. \tag{12}$$

Normierung:

$$\left| n \cdot \begin{pmatrix} 1 \\ -\frac{5}{4} \end{pmatrix} \right| \stackrel{!}{=} 1 \tag{13}$$

$$n\sqrt{1^2 + \left(\frac{5}{4}\right)^2} \stackrel{!}{=} 1 \tag{14}$$

$$n = \frac{4}{\sqrt{41}} \tag{15}$$

$$\Rightarrow \vec{v}_3 = \begin{pmatrix} \frac{4}{\sqrt{41}} \\ -\frac{5\sqrt{41}}{41} \end{pmatrix} \tag{16}$$

Wegen der Reihenfolge muss der Vektor wieder umgedreht werden.

$$\Rightarrow \vec{v}_3 = \begin{pmatrix} -\frac{4}{\sqrt{41}} \\ \frac{5\sqrt{41}}{41} \end{pmatrix} \tag{17}$$

Projektionen In den Abbildungen 2-4 sind die eindimensionalen Histogramme der Projetktionen auf die jeweiligen geraden zu finden. Die Projektion der Punkte auf die Gerade g_i der Population wird mit Hilfe der Formel

$$x = \vec{v}_i^{\top} \cdot \vec{x} \tag{18}$$

berechnet.

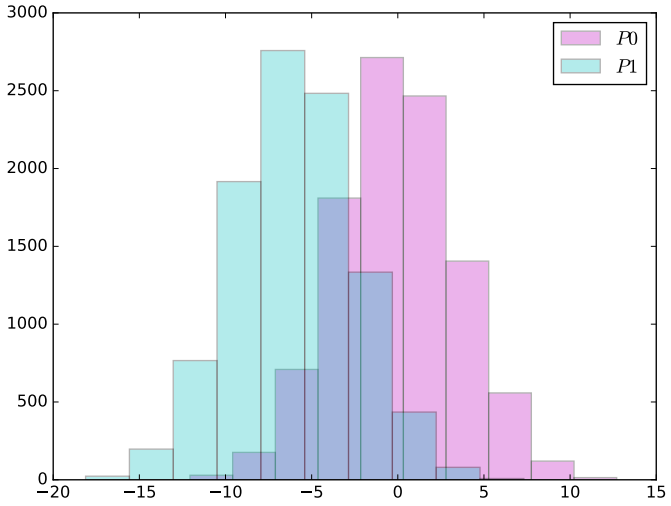


Abbildung 2: Histogramm der Projektion von den Populationen $P0$ und $P1$ auf die Projektionsgeraden $g_1(x)$.

2.3 c)

Betrachtet man nun $P0$ als Signal und $P1$ als Untergrund kann die Effizienz und Reinheit des Signals als Funktion eines Schnitts λ_{cut} aufgetragen werden. Die jeweiligen Plotts sind in den Abbildungen 5 -7 dargestellt.

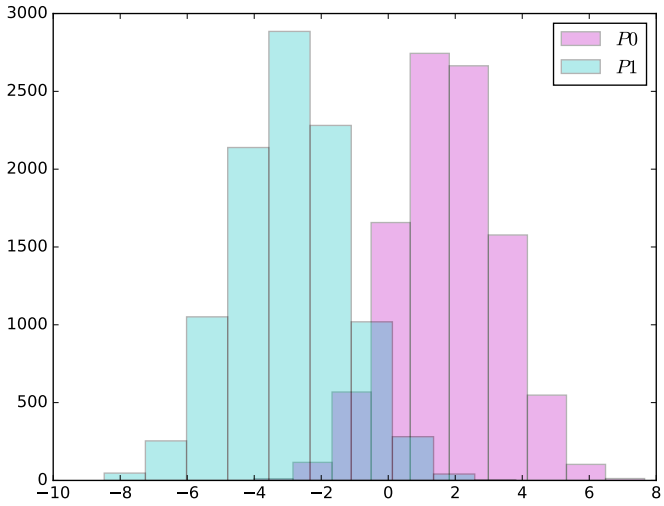


Abbildung 3: Histogramm der Projektion von den Populationen $P0$ und $P1$ auf die Projektionsgeraden $g_2(x)$.

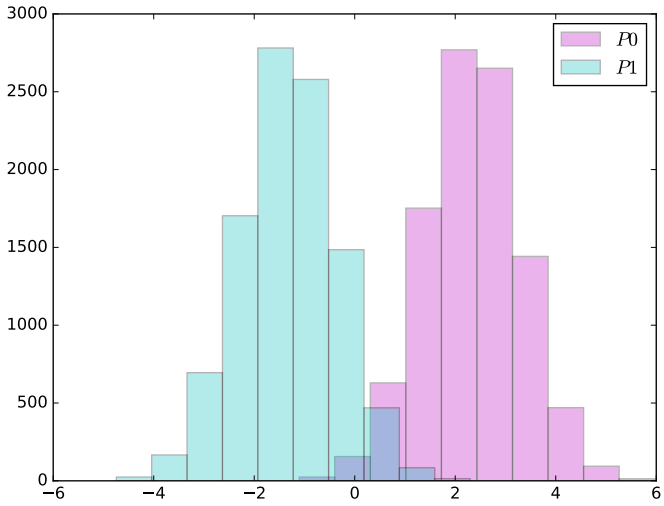


Abbildung 4: Histogramm der Projektion von den Populationen $P0$ und $P1$ auf die Projektionsgeraden $g_3(x)$.

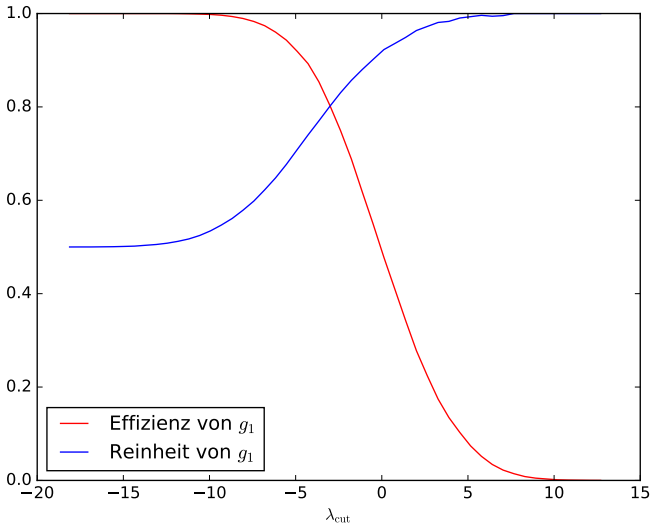


Abbildung 5: Effizienz und Reinheit des Schnittes ausgehend von der Gerade g_1 in Abhängigkeit von λ_{cut} .

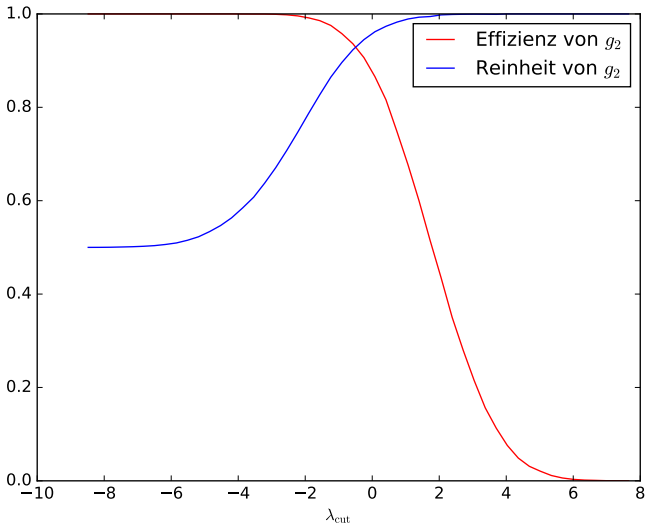


Abbildung 6: Effizienz und Reinheit des Schnittes ausgehend von der Gerade g_2 in Abhängigkeit von λ_{cut} .

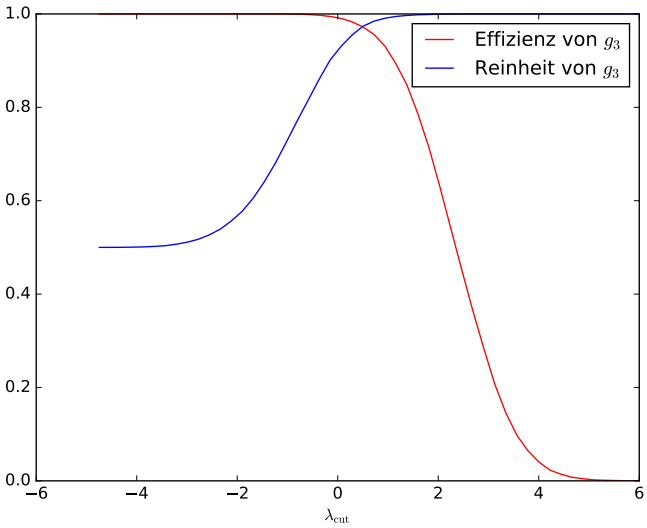


Abbildung 7: Effizienz und Reinheit des Schnittes ausgehend von der Gerade g_3 in Abhängigkeit von λ_{cut} .

3 Aufgabe3

A3 SMD

$\vec{x}_1 \in \left\{\begin{pmatrix}2\\2\\1\end{pmatrix}, \begin{pmatrix}2\\3\\2\end{pmatrix}, \begin{pmatrix}2\\1\\2\end{pmatrix}, \begin{pmatrix}1\\2\\0\end{pmatrix}, \begin{pmatrix}3\\2\\0\end{pmatrix}\right\}$

$\vec{x}_2 \in \left\{\begin{pmatrix}2.5\\2.5\\0\end{pmatrix}, \begin{pmatrix}2.5\\1.5\\0\end{pmatrix}, \begin{pmatrix}4\\2\\0\end{pmatrix}, \begin{pmatrix}5.5\\2.5\\0\end{pmatrix}, \begin{pmatrix}5.5\\1.5\\0\end{pmatrix}\right\}$

a) Mittelwerte

$\vec{\mu}_1 = \frac{1}{5}\begin{pmatrix}\sum_i^5 x_i\\ \sum_i^5 y_i\\ \sum_i^5 z_i\end{pmatrix} = \frac{1}{5}\begin{pmatrix}10\\10\\5\end{pmatrix} = \begin{pmatrix}2\\2\\1\end{pmatrix}$

$\vec{\mu}_2 = \frac{1}{5}\begin{pmatrix}20\\10\\0\end{pmatrix} = \begin{pmatrix}4\\2\\0\end{pmatrix}$

Streumatrizen — # Vektoren = 5

$S_i = \sum_1^{n_i} (\vec{x}_{ij} - \vec{\mu}_i)(\vec{x}_{ij} - \vec{\mu}_i)^T$

$S_1 = \sum_i^5 (\vec{x}_{1j} - \vec{\mu}_1)(\vec{x}_{1j} - \vec{\mu}_1)^T$

$\vec{x}_{11} - \vec{\mu}_1 = \begin{pmatrix}0\\0\\0\end{pmatrix}$; $\vec{x}_{12} - \vec{\mu}_1 = \begin{pmatrix}0\\1\\1\end{pmatrix}$

$\vec{x}_{13} - \vec{\mu}_1 = \begin{pmatrix}0\\-1\\1\end{pmatrix}$; $\vec{x}_{14} - \vec{\mu}_1 = \begin{pmatrix}-1\\0\\-1\end{pmatrix}$

$\vec{x}_{15} - \vec{\mu}_1 = \begin{pmatrix}1\\0\\-1\end{pmatrix}$

$\begin{pmatrix}0\\0\\0\end{pmatrix}(0,0,0)^T = \begin{pmatrix}0&0&0\\0&0&0\\0&0&0\end{pmatrix}$; $\begin{pmatrix}0\\1\\1\end{pmatrix}(0,1,1)^T = \begin{pmatrix}0&0&0\\0&1&1\\0&1&1\end{pmatrix}$

$\begin{pmatrix}0\\-1\\1\end{pmatrix}(0,-1,1)^T = \begin{pmatrix}0&0&0\\0&1&-1\\0&-1&1\end{pmatrix}$; $\begin{pmatrix}-1\\0\\-1\end{pmatrix}(-1,0,-1)^T = \begin{pmatrix}1&0&1\\0&0&0\\1&0&1\end{pmatrix}$

$\begin{pmatrix}1\\0\\-1\end{pmatrix}(1,0,-1)^T = \begin{pmatrix}1&0&-1\\0&0&0\\-1&0&1\end{pmatrix}$

$S_1 = \begin{pmatrix}2&0&0\\0&2&0\\0&0&4\end{pmatrix}$

Abbildung 8

zweite Popul.

S_2 : $\vec{x}_{21} - \vec{\mu}_2 = \begin{pmatrix} -1.5 \\ +0.5 \\ 0 \end{pmatrix}$

$\vec{x}_{22} - \vec{\mu}_2 = \begin{pmatrix} -1.5 \\ -0.5 \\ 0 \end{pmatrix}$; $\vec{x}_{23} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$

$\vec{x}_{24} - \vec{\mu}_2 = \begin{pmatrix} 1.5 \\ 0.5 \\ 0 \end{pmatrix}$; $\vec{x}_{25} - \vec{\mu}_2 = \begin{pmatrix} 1.5 \\ -0.5 \\ 0 \end{pmatrix}$

$\begin{pmatrix} -1.5 \\ +0.5 \\ 0 \end{pmatrix} (-1.5, -0.5, 0)^T = \begin{pmatrix} 2.25 & -0.75 & 0 \\ -0.75 & 0.25 & 0 \\ 0 & 0 & 0 \end{pmatrix}$

$\begin{pmatrix} -1.5 \\ -0.5 \\ 0 \end{pmatrix} (-1.5, -0.5, 0)^T = \begin{pmatrix} 2.25 & 0.75 & 0 \\ 0.75 & 0.25 & 0 \\ 0 & 0 & 0 \end{pmatrix}$

$\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} (0,0,0)^T = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$

$\begin{pmatrix} 1.5 \\ 0.5 \\ 0 \end{pmatrix} (1.5, 0.5, 0)^T = \begin{pmatrix} 2.25 & 0.75 & 0 \\ 0.75 & 0.25 & 0 \\ 0 & 0 & 0 \end{pmatrix}$

$\begin{pmatrix} 1.5 \\ -0.5 \\ 0 \end{pmatrix} (1.5, -0.5, 0)^T = \begin{pmatrix} 2.25 & -0.75 & 0 \\ -0.75 & 0.25 & 0 \\ 0 & 0 & 0 \end{pmatrix}$

$S_2 = \begin{pmatrix} 9 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$

$S_W = S_1 + S_2 = \begin{pmatrix} 11 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 4 \end{pmatrix}$

$S_B = (\vec{\mu}_1 - \vec{\mu}_2)(\vec{\mu}_1 - \vec{\mu}_2)^T$

$= \begin{pmatrix} -2 \\ 0 \\ 1 \end{pmatrix} (-2\ 0\ 1)^T = \begin{pmatrix} 4 & 0 & -2 \\ 0 & 0 & 0 \\ -2 & 0 & 1 \end{pmatrix}$

Abbildung 9

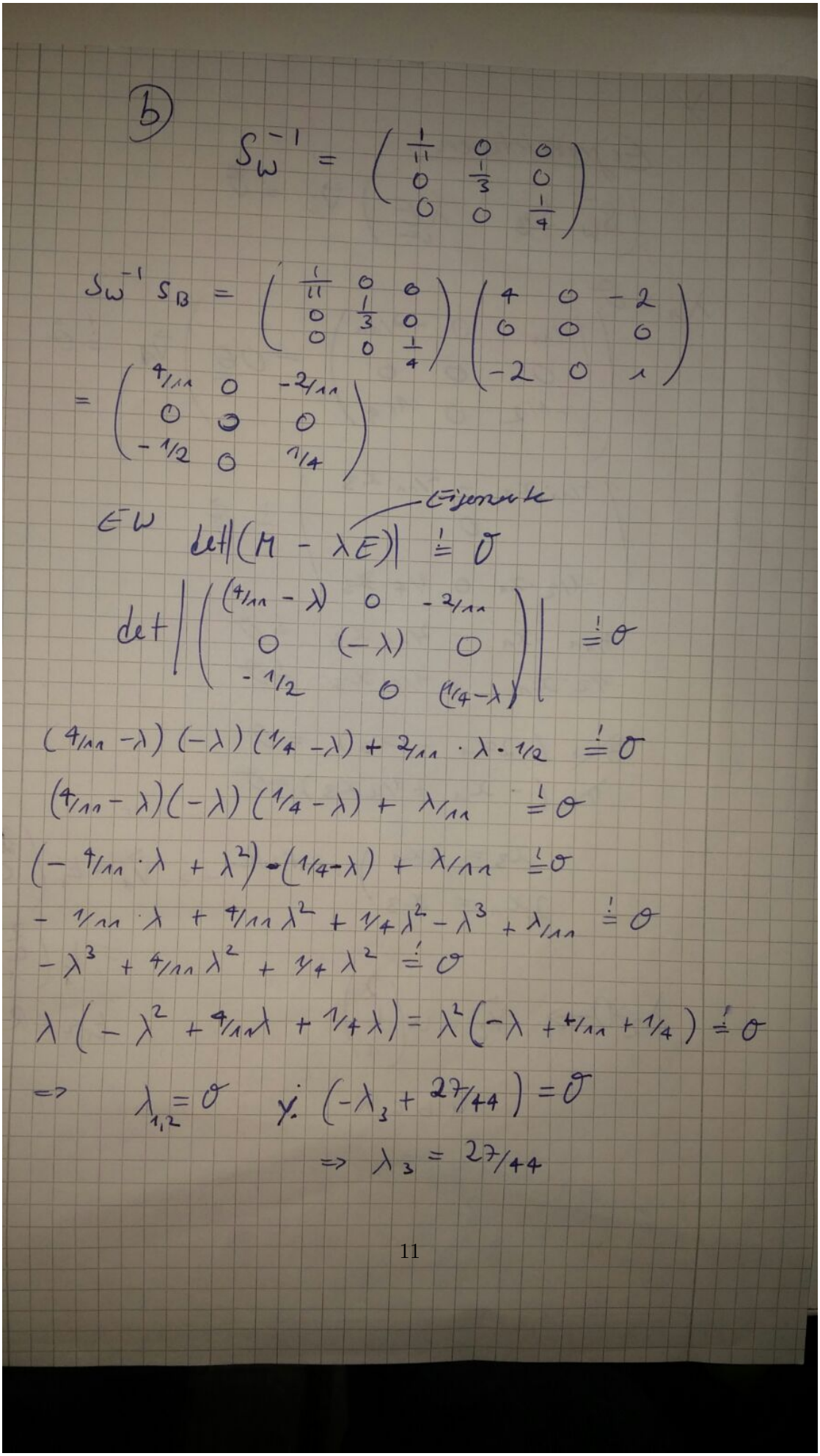


Abbildung 10

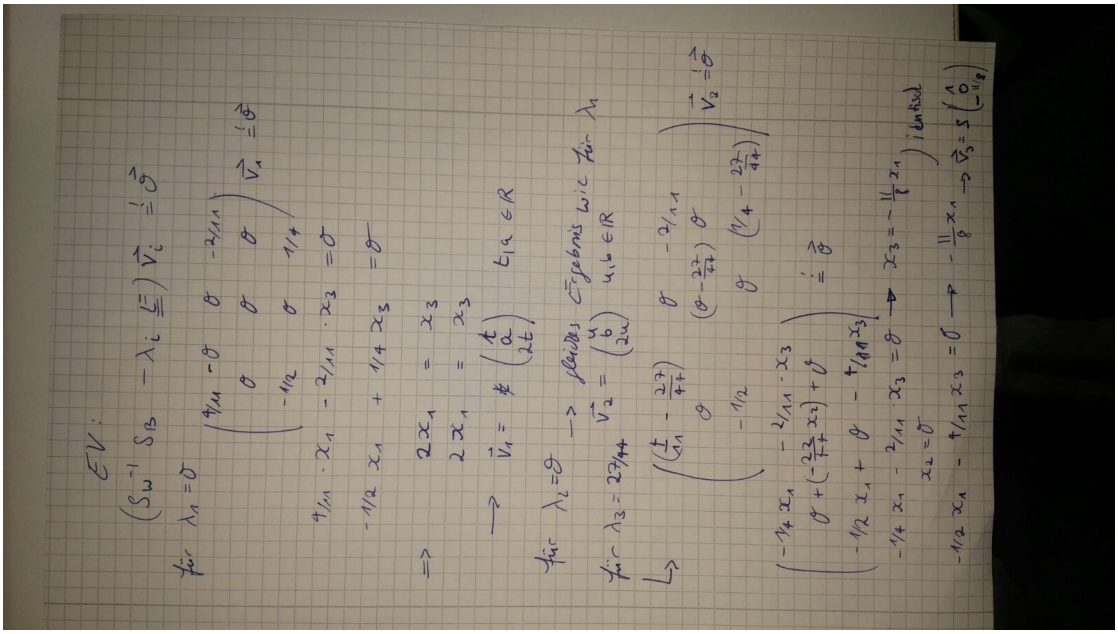


Abbildung 11

.

Abbildung 12

.

$x_{21} = \Phi^{-1}\left(-2.5 \frac{2}{11}\right) \approx -1.47$

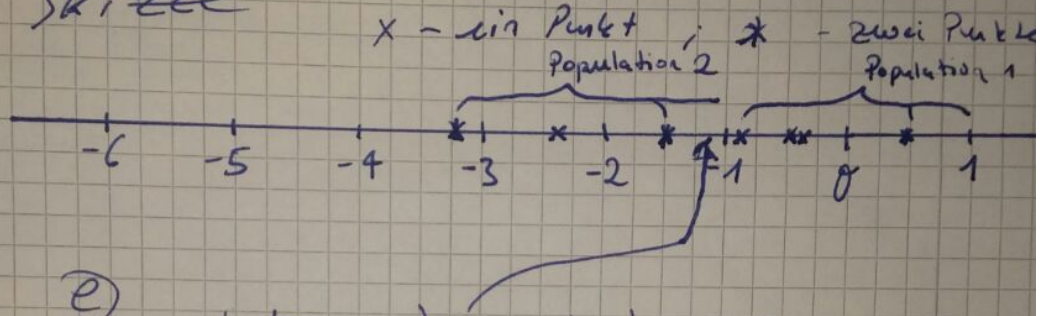
$x_{22} = \Phi^{-1}\left(-2.5 \cdot \frac{2}{11}\right) \approx -1.47$

$x_{23} = \Phi^{-1}\left(-\frac{16}{11}\right) \approx -2.35$

$x_{24} = \Phi^{-1}\left(-5.5 \cdot \frac{2}{11}\right) \simeq -3.23$

$x_{25} = \Phi^{-1}\left(-5.5 \cdot \frac{2}{11}\right) \simeq -3.23$

Skizze



e) Jetzt λ_{cut} in der Mitte zwischen den jeweils am nächsten zueinander gelegenen Werten der Populationen

$$\lambda_{cut} = -\frac{1.47 + 0.88}{2} \simeq -1.175$$

$\text{Reinheit} = \frac{5+0}{5+0} = 1$

$\text{Effizienz} = \frac{5}{0+5} = 1$

$\Big)$ -> beides sehr gut.

λ_{cut} so gewählt weil Effizienz und Reinheit maximal sind.

Abbildung 13

4 Aufgabe4

4.1 a)

1.) Tokenisierung: Segmentierung eines Textes in Einheiten, z.b. einzelne Wörter oder Satzteile. Das Ziel ist die Entfernung von unwichtigen Tokens, z.B. Füllwörtern oder Satzzeichen.
2.) Fehlerhafte Daten entfernen: Wenn sich Ausreißer unter den Daten befinden, z.B. Gewicht von 400kg bei Menschen.
3.) Default Werte verenden: es werden Default Werte anstelle der Fehlerhaften Daten verwendet.
4.) Ableiten aus Anderen Daten: aus Daten können korrekte Werte abgeleitet werden (Anrede aus Vornamen)

4.2 b)

Es ist günstig Attribute auf einen einheitlichen Wertebereich zu normieren, für eine leichtere Weiterverarbeitung. (Z.B: Firmenzusatz: e.Kfr, e.Kfm zusammengefasst zu e.K)

4.3 c)

Lücken in den Datensätzen müssen sinnvoll ersetzt oder evtl. gelöscht werden.

4.4 d)

Beim Zusammenführen von Datensätzen muss beachtet werden, dass die Teile zueinander passen und kombiniert werden können.