SMD-Abgabe

4. Übungsblatt

Lars Kolk
lars.kolk@tu-dortmund.de

Julia Sobolewski
julia.sobolewski@tu-dortmund.de

Jannine Salewski
jannine.salewski@tu-dortmund.de

Abgabe: 15.11.2018

TU Dortmund – Fakultät Physik

1 Aufgabe 10

Gegeben ist die Population P_0 mit den Werten:

$$\mu_x = 0 \tag{1}$$
$$\mu_y = 3 \tag{2}$$
$$\sigma_x = 3{,}5 \tag{3}$$
$$\sigma_y = 2{,}6 \tag{4}$$
$$\rho = 0,9. \tag{5}$$

Ebenso ist eine Population P_1 gegeben. Bei dieser sind die Werte in x gegeben durch:

$$\mu_x = 6 \tag{6}$$
$$\sigma_x = 3{,}5. \tag{7}$$

Für die Werte in y sind folgende Zusammenhänge gegeben:

$$\mathrm{E}(y|x) = \mu_{y|x} = a + bx \text{ mit } a = -0{,}5 \text{ und } b = 0{,}6 \tag{8}$$
$$\mathrm{Var}(y|x) = \sigma^2_{y|x} = 1 \tag{9}$$

1.1 Teilaufgabe a)

1.1.1 Zeigen der 2D-Gaußverteilung

Die Formel für die bedingte Wahrscheinlichkeit der bivariaten Normalverteilung lautet:

$$f(y|x) = \frac{1}{\sqrt{2\pi}\sigma_y\sqrt{1-\rho^2}} \cdot \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{\tilde{y}}{\sigma_y} - \rho\frac{\tilde{x}}{\sigma_x}\right]^2\right). \tag{10}$$

wobei $\tilde{x} = x - \mu_x$ und $\tilde{y} = y - \mu_y$

Für die Wahrscheinlichkeitsdichtefunktion einer bivariaten Normalverteilung gilt dagegen:

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right]\right), \tag{11}$$

Mit $f(x,y) = g(x)f(y|x)$ folgt mit

$$g(x) = \frac{\sqrt{2(1-\rho^2)}}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\sqrt{\pi}\sigma_y \cdot \exp\left(-\frac{1}{2}\left(\frac{x-\mu_x}{\sigma_x}\right)^2\right) \tag{12}$$

der Zusammenhang

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \cdot \exp\left(-\frac{1}{2}\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - \frac{1}{2(1-\rho^2)}\left(\frac{y-\mu_y}{\sigma_y} - \rho\frac{x-\mu_x}{\sigma_x}\right)^2\right). \tag{13}$$

Durch Umformungen erhält man die Gleichung (11), da:

$$-\frac{(x-\mu_x)^2}{\sigma_x^2}\frac{\rho^2}{2(1-\rho^2)}-\frac{(x-\mu_x)^2}{2\sigma_x^2}=-\frac{(x-\mu_x)^2}{\sigma_x^2}\frac{1}{2(1-\rho^2)} \tag{14}$$

1.1.2 Bestimmen der Werte

Für die Berechnung der Werte werden folgende Gleichungen genutzt:

$$\sigma_{y|x}^2=(1-\rho^2)\sigma_y^2 \tag{15}$$

$$\mathrm{E}(y|x)=\rho\frac{\sigma_y}{\sigma_x}(x-\mu_x)+\mu_y \tag{16}$$

$$\tag{17}$$

Bei einem Koeffizientenvergleich der Gleichung (16) mit (8) ergeben sich folgende Zusammenhänge:

$$b=\frac{\sigma_y}{\sigma_x}\cdot\rho \tag{18}$$

$$a=-\rho\frac{\sigma_y}{\sigma_x}\mu_x+\mu_y \tag{19}$$

Aus (18) und (15) ergibt sich der Zusammenhang

$$p=\sqrt{\frac{1}{1+\frac{\sigma_{y|x}^2}{b^2\sigma_x^2}}}\approx 0,75 \tag{20}$$

womit sich $\sigma_y=\sqrt{\frac{\sigma_{y|x}}{1-p^2}}\approx 1,51$ und $\mu_y=\alpha+\frac{\rho\mu_x}{\sigma_x}\cdot\frac{\sigma_{y|x}}{\sqrt{1-p^2}}\approx 1,44$ ergeben .

1.2 Teilaufgabe b)

Es ergibt sich folgender Scatter-Plot:

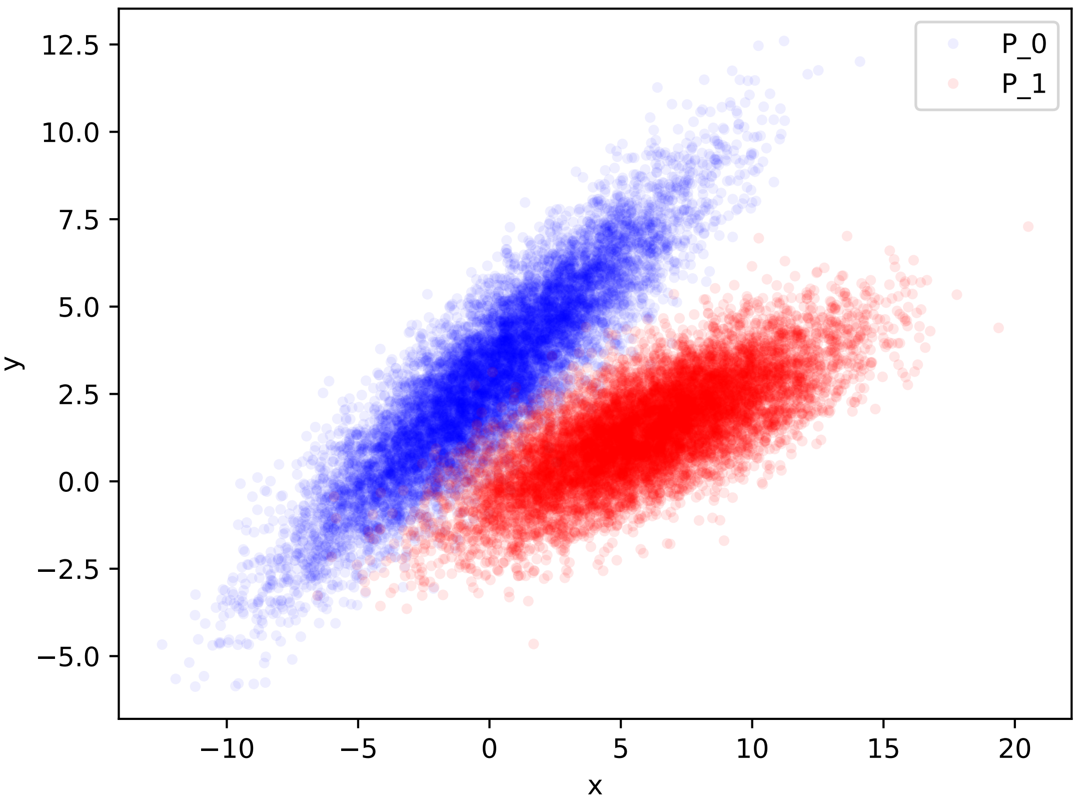


Abbildung 1: Scatter-Plot

1.3 Teilaufgabe c)

es ergeben sich folgende Ergebnisse:

Ausgabe des Programms: Population 0

Mittelwerte von P0: [0.00781587 3.00831114]
Varianz von x0: 12.14051138946726
Varianz von y0: 6.654329986905046
Kovarianz cov(x, y): 8.095175577329087
Korrelation rho: 0.9006490919236851

Ausgabe des Programms: Population 1

Mittelwerte von P0: [6.00106294 1.41078796]
Varianz von x0: 12.258004437206726
Varianz von y0: 2.2870875071918366
Kovarianz cov(x, y): 3.9966098123833347
Korrelation rho: 0.7548149136498276

Ausgabe des Programms: Population 0 + Population 1

Mittelwerte von P0: [2.99233451 2.20887921]
Varianz von x0: 21.37175112487808
Varianz von y0: 5.207492960450249
Kovarianz cov(x, y): 3.8264610565821275
Korrelation rho: 0.3627128132883111

SMD - Blatt 4

Aufgabe 11

Population 0: $\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}2\\1\end{pmatrix}, \begin{pmatrix}1,5\\2\end{pmatrix}, \begin{pmatrix}2\\2\end{pmatrix}, \begin{pmatrix}2\\3\end{pmatrix}, \begin{pmatrix}3\\3\end{pmatrix}$

Population 1: $\begin{pmatrix}1,5\\1\end{pmatrix}, \begin{pmatrix}2,5\\1\end{pmatrix}, \begin{pmatrix}3,5\\1\end{pmatrix}, \begin{pmatrix}2,5\\2\end{pmatrix}, \begin{pmatrix}3,5\\2\end{pmatrix}, \begin{pmatrix}4,5\\2\end{pmatrix}$

a) $\vec{\mu} = \frac{1}{N}\begin{pmatrix}\sum_{i=1}^{N} x_i\\ \sum_{i=1}^{N} y_i\end{pmatrix}$

$\Rightarrow \vec{\mu}_0 = \begin{pmatrix}23/12\\ 2\end{pmatrix}, \quad \vec{\mu}_1 = \begin{pmatrix}3\\ 3/2\end{pmatrix}$

$S_i = \sum_{j=1}^{n_i} (\vec{x}_j - \vec{\mu}_i)(\vec{x}_j - \vec{\mu}_i)^T$

$S_0 = \begin{pmatrix}121/144 & 11/12\\ 11/12 & 1\end{pmatrix} + \begin{pmatrix}1/144 & -1/12\\ -1/12 & 1\end{pmatrix} + \begin{pmatrix}25/144 & 0\\ 0 & 0\end{pmatrix} + \begin{pmatrix}1/144 & 0\\ 0 & 0\end{pmatrix} + \begin{pmatrix}1/144 & 1/12\\ 1/12 & 1\end{pmatrix} + \begin{pmatrix}169/144 & 13/12\\ 13/12 & 1\end{pmatrix}$

$= \begin{pmatrix}53/24 & 2\\ 2 & 4\end{pmatrix}$

$S_1 = \begin{pmatrix}9/4 & 3/4\\ 3/4 & 1/4\end{pmatrix} + \begin{pmatrix}1/4 & 1/4\\ 1/4 & 1/4\end{pmatrix} + \begin{pmatrix}1/4 & -1/4\\ -1/4 & 1/4\end{pmatrix} + \begin{pmatrix}1/4 & -1/4\\ -1/4 & 1/4\end{pmatrix} + \begin{pmatrix}1/4 & 1/4\\ 1/4 & 1/4\end{pmatrix} + \begin{pmatrix}9/4 & 3/4\\ 3/4 & 1/4\end{pmatrix}$

$= \begin{pmatrix}11/2 & 3/2\\ 3/2 & 3/2\end{pmatrix}$

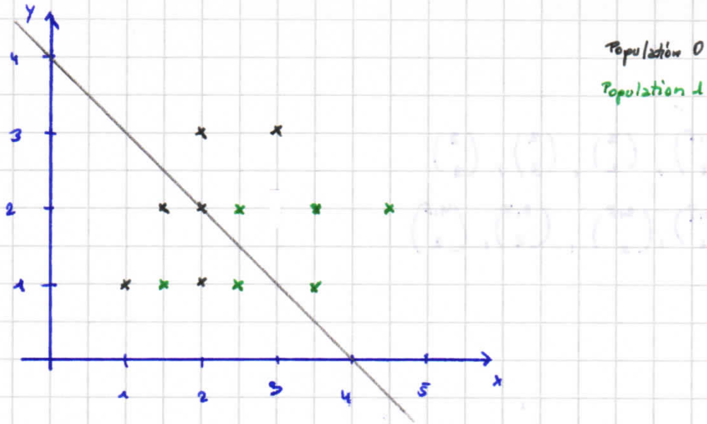
$S_W = S_0 + S_1 = \begin{pmatrix}185/24 & 7/2\\ 7/2 & 11/2\end{pmatrix} \Rightarrow S_W^{-1} = \begin{pmatrix}0,182 & -0,116\\ -0,116 & 0,256\end{pmatrix}$

$S_B = (\vec{\mu}_0 - \vec{\mu}_1)(\vec{\mu}_0 - \vec{\mu}_1)^T = \begin{pmatrix}169/144 & -13/24\\ -13/24 & 1/4\end{pmatrix}$

b) $\vec{\lambda} = S_W^{-1}(\vec{\mu}_0 - \vec{\mu}_1) = \begin{pmatrix}-\frac{370}{1447}\\ \frac{367}{1447}\end{pmatrix} = \begin{pmatrix}-0,256\\ 0,254\end{pmatrix}$

$\vec{\lambda} = \lambda \vec{e}_\lambda = 0,360 \begin{pmatrix}-0,710\\ 0,704\end{pmatrix}$

c)



d) Projektion $= \vec{\lambda}^T \vec{x}$

Population 0: $-0{,}005$; $-0{,}716$; $0{,}343$; $-0{,}012$; $0{,}693$; $-0{,}017$

Population 1: $-0{,}361$; $-1{,}071$; $-1{,}781$; $-0{,}367$; $-1{,}076$; $-1{,}786$

e) $\lambda_{cut} = -0{,}364$ $\longrightarrow$ gewählt, weil Verhältnis zw. Effizienz und Reinheit 1 ist.

$\rightarrow t_p = 5$ Reinheit: $\frac{t_p}{t_p + f_p} = 0{,}83$

$t_n = 1$

$f_p = 1$ Effizienz: $\frac{t_p}{t_p + f_n} = 0{,}83$

$f_n = 5$

2 Aufgabe 12

2.1 a)

Mittelwerte:

$$\vec{\mu}_0 = \begin{pmatrix} -0,027 \\ 2,980 \end{pmatrix}$$

$$\vec{\mu}_1 = \begin{pmatrix} 5,986 \\ 3,085 \end{pmatrix}$$

2.2 b)

Kovarianzmatrizen:

$$V_0 = \begin{pmatrix} 12,209 & 8,158 \\ 8,158 & 6,7223 \end{pmatrix}$$

$$V_1 = \begin{pmatrix} 12,352 & 7,411 \\ 7,411 & 5,477 \end{pmatrix}$$

Kombinierte Kovaranzmatrix:

$$V_{0,1} = \begin{pmatrix} 21,322 & 7,943 \\ 7,943 & 6,103 \end{pmatrix}$$

2.3 c)

Zu Berechnung der Fisher-Diskriminante, müssen zunächst die Streumatrizen berechnet werden:

$$S_0 = \begin{pmatrix} 122077,077 & 81575,940 \\ 81575,940 & 67221,910 \end{pmatrix}$$

$$S_1 = \begin{pmatrix} 123509.502 & 74100.151 \\ 74100.151 & 54767.673 \end{pmatrix}$$

Daraus wird die Gesamtstreuung S_W berechnet

$$S_W = S_0 + S_1 = \begin{pmatrix} 245586,579 & 155676,091 \\ 155676,091 & 121989,583 \end{pmatrix}$$

Die Fisher-Diskriminante lässt sich nun mit Hilfe der Formel

$$\vec{\lambda} = S_W^{-1}(\vec{\mu}_0 - \vec{\mu}_1) \tag{21}$$

berechnen:

$$\vec{\lambda} = \begin{pmatrix} -0,0001253 \\ 0,00015904 \end{pmatrix}$$

Diese lässt sich als Geradengleichung der Form $\vec{\lambda} = \lambda \cdot \vec{e}_\lambda$ darstellen:, mit

$$\lambda = 0,00020247$$

und

$$\vec{e}_\lambda = \begin{pmatrix} -0,619 \\ 0,785 \end{pmatrix} \tag{22}$$

Der Einheitsvektor (22) lässt sich für die Projektion in den nächsten Aufgabenteilen verwenden.

2.4 d)

Zu Projektion der einzelnen Punkte auf die Gerade $\vec{\lambda}$ wird folgende Formel verwendet:

$$P_\lambda(\vec{x}) = (\vec{x} \cdot \vec{e}_\lambda) \cdot \vec{e}_\lambda$$

wobei für die eindimensionale Verteilung nur der Betrag genommen wird, also nur $(\vec{x} \cdot \vec{e}_\lambda)$. Die eindimensionale Verteilung auf der Geraden ist in Abbildung 2 zu sehen.

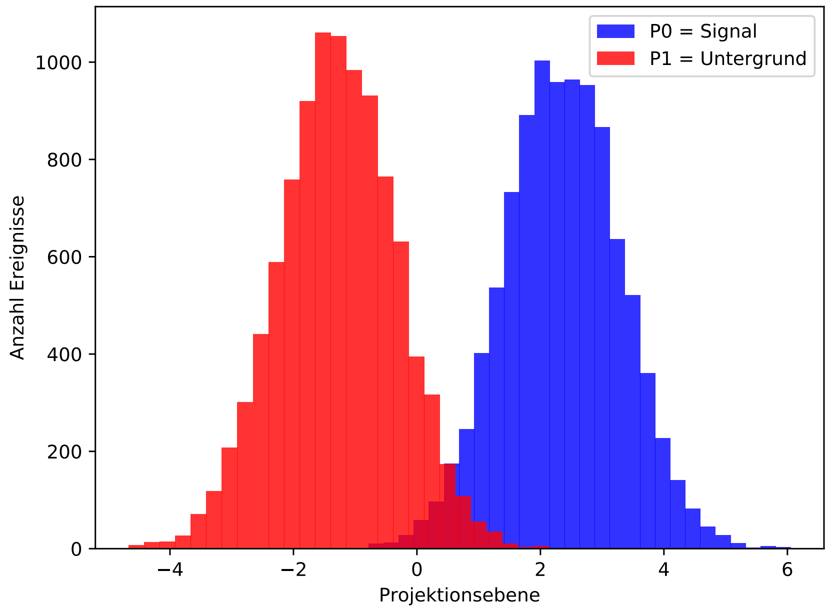


Abbildung 2: Projektion auf die Gerade.

2.5 e)

Die Effizienz wird mit der Formel:

$$\text{Effizienz} = \frac{t_p}{t_p + f_p}$$

berechnet und die Reinheit mir:

$$\text{Reinheit} = \frac{t_p}{t_p + f_n} \tag{23}$$

Die Effizienz und die Reinheit in Abhängigkeit der Cut-Stelle λ_{cut} ist in Abbildung 3 dargestellt.

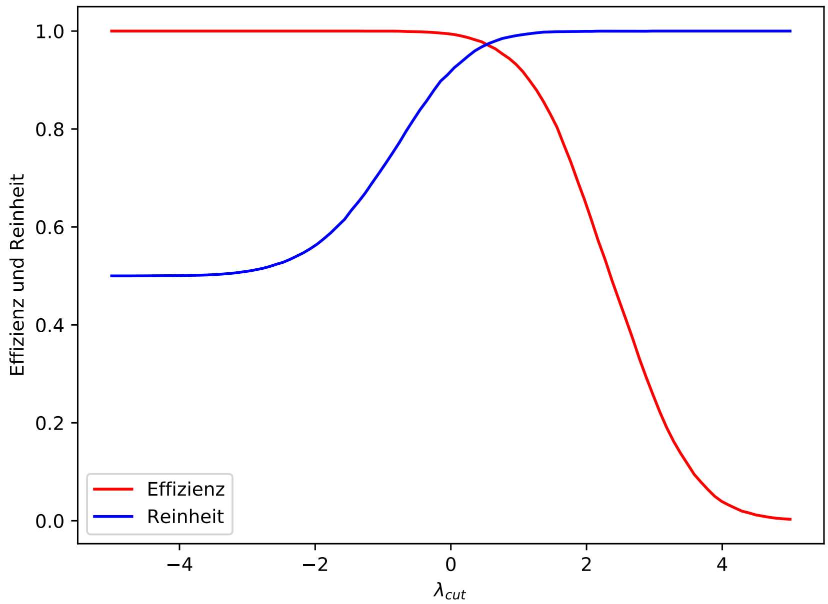


Abbildung 3: Effizienz und Reinheit in Abhängigkeit der Cut-Stelle.

2.6 f)

Das Verhältnis zwischen Signal und Untergrund in Abhängigkeit von λ_{cut} ist in Abbildung 4 zu sehen. Der Maximalwert liegt hier bei

$$\lambda_{\text{cut,Verhltnis}} \approx 2,17$$

2.7 g)

Die Signifikanz $\frac{S}{\sqrt{S+B}}$ in Abhängigkeit von λ_{cut} ist in Abbildung 5 dargestellt. Der Maximalwert liegt hier bei:

$$\lambda_{\text{cut,Signifikanz}} \approx 0,45$$

2.8 f)

Das ganze soll nun mit einem anderen Signal erneut untersucht werden:

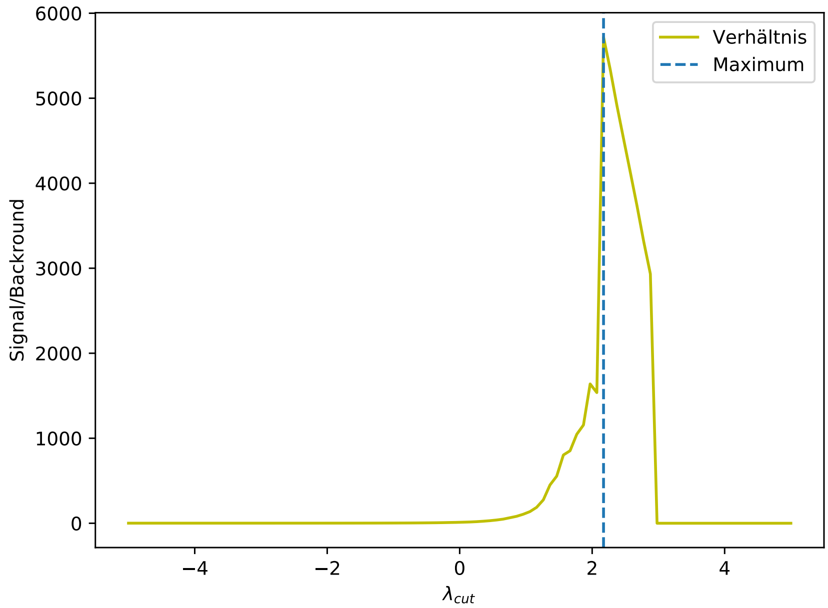


Abbildung 4: Verhältnis S/B in Abhängigkeit der Cut-Stelle.

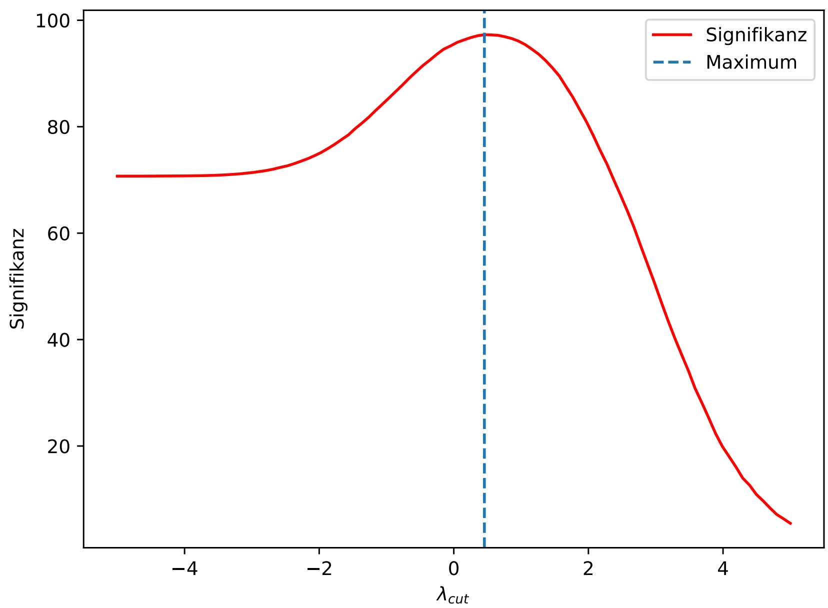


Abbildung 5: Signifikanz in Abhängigkeit der Cut-Stelle.

Mittelwert:

$$\vec{\mu}_2 = \begin{pmatrix} -0,0958 \\ 2,8788 \end{pmatrix}$$

Kovarianzmatrix:

$$V_2 = \begin{pmatrix} 12,236 & 8,160 \\ 8,160 & 6,758 \end{pmatrix}$$

Kombinierte Kovaranzmatrix:

$$V_{2,1} = \begin{pmatrix} 15,398 & 7,582 \\ 7,582 & 5,597 \end{pmatrix}$$

Streumatrix:

$$S_2 = \begin{pmatrix} 12223,886 & 81252,338 \\ 81252,338 & 6751,132 \end{pmatrix}$$

Gesamtstreuung:

$$S_{W2} = \begin{pmatrix} 135733,388 & 82252,489 \\ 82252,489 & 61519,105 \end{pmatrix}$$

Fisher-Diskriminante:

$$\vec{\lambda}_2 = \begin{pmatrix} -0,000254 \\ 0,000298 \end{pmatrix} = 0,374 \cdot 10^{-3} \begin{pmatrix} -0,603 \\ 0,797 \end{pmatrix} = \lambda_2 \cdot \vec{e}_{\lambda 2}$$

Die eindimensionale Verteilung ist in Abbildung 6 dargestellt.

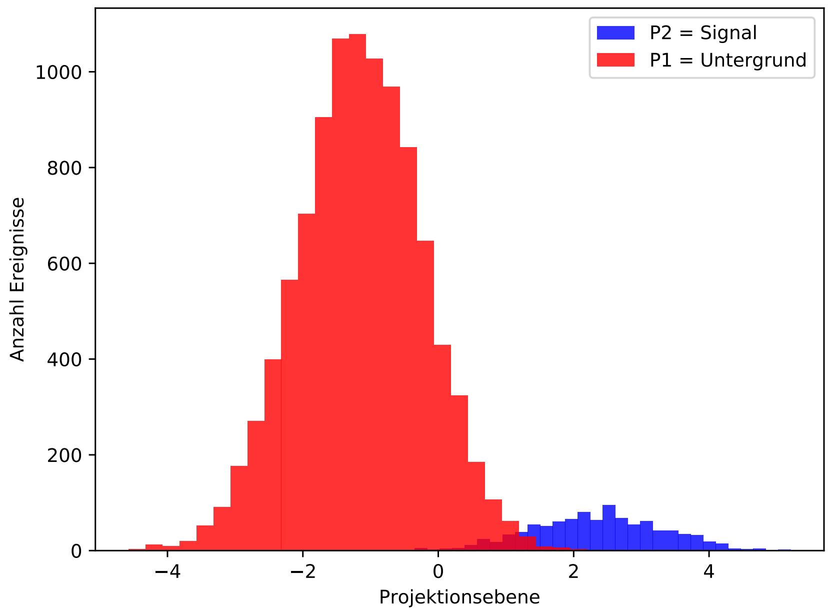


Abbildung 6: Eindimensionale Verteilung mit P2.

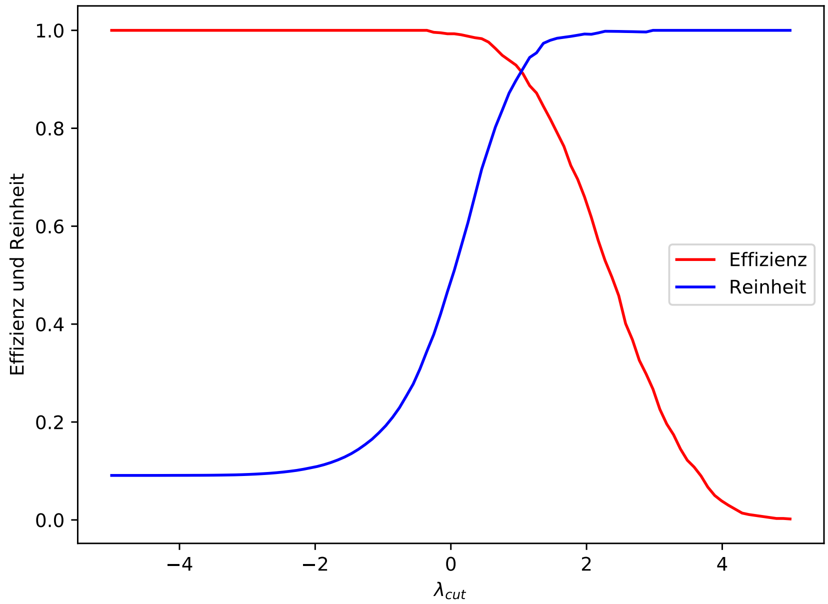


Abbildung 7: Effizienz und Reinheit der zweiten Verteilung.

Die Effizienz und die Reinheit sind in Abbildung 7 zu sehen.
Das Verhätnis S/B in Abhängigkeit von der Cut-Stelle ist in Abbildung 8 dargestellt. Das Maximum liegt bei:

$$\lambda_{\text{cut,Verhaeltnis2}} \approx 2,27$$

Die Signifikanzin Abhängigkeit von der Cut-Stelle ist in Abbildung 9 dargestellt. Das Maximum liegt bei:

$$\lambda_{\text{cut,Signifikanz2}} \approx 1,06$$

Die Trennung funktioniert für kleinere oder gleichgroße Untergründe, im Bezug auf das Signal, besser. Dies ist an den Maxima der Signifikanzkurve zu erkennen, da das Maximum der ersten Verteilung deutlich höher liegt, als die der zweiten.

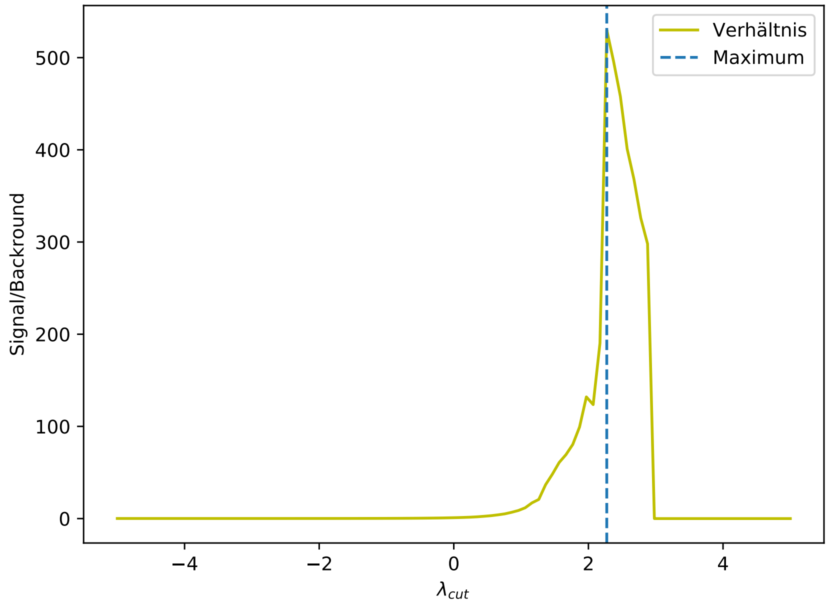


Abbildung 8: Verhältnis in Abhängigkeit der Cut-Stelle der zweiten Verteilung.

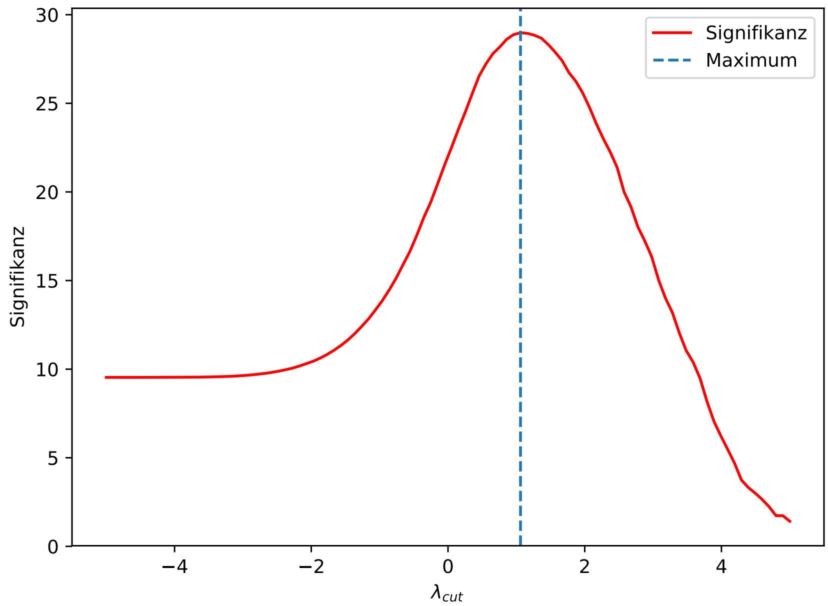


Abbildung 9: Signifikanz in Abhängigkeit der Cut-Stelle der zweiten Verteilung.