technische universität
dortmund

Bachelorarbeit

Untersuchung von Genmutationsdaten mittels Latent Dirichlet Allocation

Malik Atamne
Dezember 2017

Gutachter:
Prof. Dr. Katharina Morik
Sibylle Hess

Technische Universität Dortmund
Fakultät für Informatik
Lehrstuhl für Künstliche Intelligenz (8)
http://www-ai.cs.uni-dortmund.de

Inhaltsverzeichnis

Kapitel 1

Einleitung

1.1 Motivation und Hintergrund

Heutzutage erzeugt jede Person eine große Menge an Daten, sei es durch das Schießen von Fotos, Aufnehmen von Videos, Schreiben von Nachrichten oder Erstellen von Dokumenten. Es wird dementsprechend immer schwieriger, schnell auf bestimme Daten zuzugreifen, nach bestimmten Daten zu suchen und die Daten zu interpretieren.

Wird beispielsweise versucht ein Dokument zu interpretieren, indem die Themen, die das Dokument behandelt, herausgefunden werden sollen, führte früher kein Weg daran vorbei, sich das Dokument selbst anzuschauen. Doch mittlerweile wurden Topic Modelling Algorithmen vorgestellt, die dieses Problem automatisiert lösen können.

Topic Modelling ist ein Verfahren um abstrakte Themen aus einem Datensatz automatisiert zu entnehmen. Dieses Verfahren wird häufig benutzt, um herauszufinden, welche Themen ein Textdokument behandelt, um somit Informationen über das Textdokument und den Zusammenhang zwischen mehreren Textdokumenten zu erhalten.

Durch das Anwenden von Topic Modelling auf Textdokumente können die Strukturen dieser schnell enthüllt werden. Somit wird die Suche nach bestimmten Textdokumenten stark vereinfacht. Ein weiterer Vorteil von Topic Modelling ist, dass die einzelnen Textdokumente mit Hilfe der beobachteten Strukturen miteinander verglichen werden können. Ein Anwendungsfall dafür könnte beispielsweise sein, einem Kunden ein Buch vorzuschlagen, welches die selben Themen behandelt, wie die Bücher die der Kunde schon mal gelesen hat.

Es lassen sich jedoch nicht nur übliche Textdokumente mittels Topic Modelling untersuchen. Jede andere Art von Datensatz, die in eine Form gebracht wird, die einem Dokument ähnlich ist, kann mit Topic Modelling untersucht werden. So können auch Genmutationsdaten von Patienten, die an Neuroblastoma erkrankt sind, mittels dieser Algorithmen analysiert werden. Demnach wird versucht, die zugrundeliegende Struktur der Genmutati-

onsdaten herauszufinden, um Patienten abhängig von der Struktur miteinander vergleichen zu können.

1.2 Aufbau der Arbeit

In Kapitel 2 werden vorerst alle benötigten Grundlagen besprochen. Es werden wichtige Begriffe für das Verständnis der Genmutationsdaten kurz erklärt. Darauf folgen Erläuterungen einiger Wahrscheinlichkeitsmodelle, die für die in dieser Arbeit angewendeten Topic Models benötigt werden. Im letzten Abschnitt des Kapitels werden die Topic Models ausführlich dargelegt.

Im Anschluss werden in Kapitel 3 die Genmutationsdaten vorgestellt und auf die Präparierung der Daten, um diese mit den Topic Models untersuchen zu können, eingegangen.

In Kapitel 4 wird das Vorgehen der Analyse erklärt und die Implementierungen der Topic Models dargestellt. Daraufhin werden die Ergebnisse vorgestellt.

Anschließend werden in Kapitel 5 die Ergebnisse interpretiert und in Kapitel 6 ein abschließendes Fazit formuliert.

Kapitel 2

Grundlagen

2.1 Genetik

Da es sich bei dem zu untersuchenden Datensätzen um Genmutationsdaten von Patienten handelt, müssen vorher noch einige Fachbegriffe bezüglich der Genetik geklärt werden.

Die gesamten Erbinformationen eines Lebewesens werden als **Genom** bezeichnet. Das Genom des Menschen besteht aus insgesamt 46 **Chromosomen** (23 von dem Vater und 23 von der Mutter), die wiederum aus **Desoxyribonukleinsäure-Strängen** (DNA-Stränge) bestehen. Ein DNA-Strang ist als Doppelhelix aufgebaut und besteht aus vielen Nukleotiden. Ein **Nukleotid** besteht aus einem Phosphorsäurerest, einem Monosaccharid (Zucker) und einer organischen **Base**. Insgesamt gibt es vier Basen. Adenin, Guanin, Cytosin und Thymin. Dabei sind Adenin und Thymin sowie Guanin und Cytosin zueinander komplementär. Dementsprechend bilden Adenin und Thymin sowie Guanin und Cytosin Basenpaare.

Bestimmte Sequenzen von Basen in einem DNA-Strang werden als **Gene** bezeichnet. Diese enthalten Exons und Introns, wobei die Exons besonders wichtig sind, da sie den codierenden Bereich eines Gens beinhalten. Mit dem codierenden Bereich werden **Ribonukleinsäure-Stränge** (RNA-Stränge) hergestellt, in denen die Information für die Herstellungen bestimmter Proteine codiert sind. Die Introns werden vor der Herstellung des Proteins von der RNA herausgespleißt. Jedoch kann es dazu kommen, dass bei der Aussetzung von energiereichen Strahlen (z.B. UV-Strahlung oder Röntgenstrahlung) oder bei der Einnahme von schädlichen Stoffen Gene mutieren. Wenn solch eine **Mutation** auftritt, handelt es sich dabei entweder um eine Veränderung einer Base in einem Gen ($G \rightarrow C$), eine Löschung einer oder mehrerer Basen ($GA \rightarrow A$) oder eine Einfügung von Basen ($A \rightarrow CA$). Welche Konsequenzen die Mutation haben wird, hängt stark von der Stelle ab, an der sie aufgetreten ist. Einige Positionen sind für die Herstellung der Proteine irrelevant und tragen dadurch auch oft keine Folgen mit sich. Besonders schlimm kann es jedoch werden, wenn sich die Mutation im Codogenen Strang befindet, denn dies kann dazu führen, dass

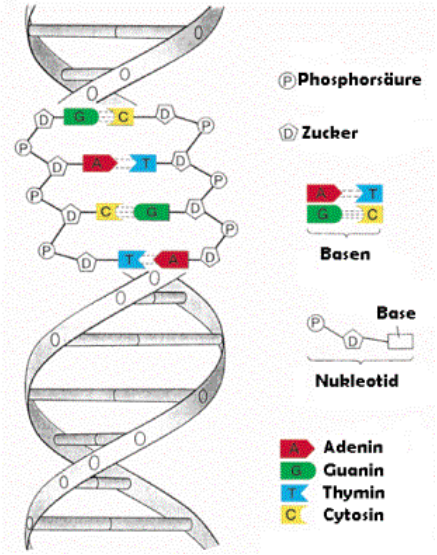


Abbildung 2.1: Die Abbildung zeigt einen DNA-Strang und die einzelnen Komponenten, aus denen ein DNA-Strang zusammengebaut ist. Jedes Nukleotid besteht au seinem Phosphorsäurerest, einem Monosaccharid und zwei Basen. Die Nukleotiden können sich nur in den Basen unterscheiden. So besteht das erste Nukleotid aus den Basen Guanin und Cytosin, das zweite aus Adenin und Thymin, das dritte aus Cytosin und Guanin und das vierte aus Thymin und Adenin. Die Nukleotiden können nur in einer der vier genannten Zusammenstellungen in einem DNA-Strang auftreten.[1]

ein Protein nicht mehr hergestellt werden kann oder dass ein anderes Protein hergestellt wird. Je nachdem welche Auswirkungen die Mutation auf die Herstellung des Proteins hat, kann die Zelle, in der die Mutation aufgetreten ist, zur Krebszelle werden.

Bei den Genmutationsdaten handelt es sich um einen Datensatz, der Informationen über mutierte Gene von Patienten, die am **Neuroblastom** erkrankt sind, beinhalten. Beim Neuroblastom handelt es sich um eine Krebserkrankung, die gehäuft im Kindesalter auftritt. 15% aller krebstoten Kinder sind am Neuroblastom gestorben [10]. In einer Studie von 1971 [5] hat sich herausgestellt, dass nur 30% der Kinder, die mit Neuroblastom diagnostiziert wurden, überlebt haben. Dabei spielte das Alter eine große Rolle. 82% der Kinder, die noch unter einem Jahr alt waren, haben überlebt. Jedoch haben nur 32% der Kinder überlebt, die zwischen 1-2 Jahre alt waren. Von den Kindern, die über 2 Jahre alt waren, haben nur noch 10% überlebt. Dennoch ist es meistens nicht die ursprüngliche Entstehung des Neuroblastoms, das den Tod eines Kindes verursacht, sondern das nach einer Remission entstehende Rezidivgewebe. Eine Remission beschreibt einen vollständigen oder vorübergehenden Nachlass einer Krankheit. Beim Neuroblastom wird das Gewebe als Rezidivgewebe bezeichnet, das sich nach der Rückbildung der ursprünglichen Neuroblastom-

[1] https://www.tgg-leer.de/projekte/genetik/dna2/dna2.html

Erkrankung bildet. Das Rezidivgewebe entsteht allerdings nicht bei jedem Patienten. Es wird angenommen, dass unterschiedliche Mutationen der Grund für das Nichtentstehen des Rezidivgewebes ist. Deshalb werden Untersuchungen durchgeführt, um herauszufinden, welche Mutationen bei der Entstehung des Rezidivgewebes eine Rolle spielen. Es hat sich bereits herausgestellt, dass die Anzahl der Mutationen in den Zellen des Rezidivgewebes im Vergleich zu dem ursprünglichen Tumorgewebe sehr viel höher ist. Außerdem sind die meisten im Tumorgewebe gefundenen Mutationen, ebenfalls im Rezidivgewebe enthalten

2.2 Wahrscheinlichkeitsverteilungen

Die Topic Models, die in dieser Arbeit für die Untersuchung der Genmutationsdaten verwendet werden, sind Wahrscheinlichkeitsmodelle, die unterschiedliche Wahrscheinlichkeitsverteilungen verwenden, um bestimmte Variablen zufällig zu wählen. Bevor die Wahrscheinlichkeitsmodelle vorgestellt werden, werden alle für die Wahrscheinlichkeitsmodelle benötigten Wahrscheinlichkeitsverteilungen erläutert.

2.2.1 Poisson-Verteilung

Die Poisson-Verteilung ist eine diskrete Wahrscheinlichkeitsverteilung, die die Wahrscheinlichkeit des Auftretens eines Ereignisses in Abhängigkeit von dem Mittelwert aller Ereignisse, die in einem festen Zeitraum auftreten können, darstellt. Die Verteilung ist definiert durch:

$$P(X = x) = \frac{\lambda^x}{x!} * e^{-\lambda}; \quad x \geq 0\ , x \in \mathbb{N}$$

Die Variable λ steht für den Mittelwert der Ereignisse und x für ein bestimmtes Ereignis.

Beispielsweise ist es mit der Wahrscheinlichkeitsfunktion möglich herauszufinden, wie hoch die Wahrscheinlichkeit dafür ist, dass an einem Tag genau 980 Autos eine Brücke überqueren, statt den durchschnittlichen 1000. In diesem Fall entspricht $\lambda = 1000$ und $x = 980$. Eingesetzt in die Funktion ergibt dies ungefähr $0,0104$, was umgerechnet $1,04\%$ entspricht.

In Abbildung 2.2 ist zu erkennen, dass die Wahrscheinlichkeit von $P(X = x)$ mit $x = \lambda$ immer am größten ist. Je weiter ein Ereignis von dem Mittelwert entfernt ist, desto unwahrscheinlicher ist das Ereignis.

2.2.2 Multinomialverteilung

Die Multinomialverteilung ist eine diskrete Wahrscheinlichkeitsverteilung, die eine Verallgemeinerung der Binomialverteilung ist. Wenn sich die Ereignisse in zwei Klassen aufteilen lassen, dann können die Wahrscheinlichkeiten der Ereignisse als eine Binomialverteilung

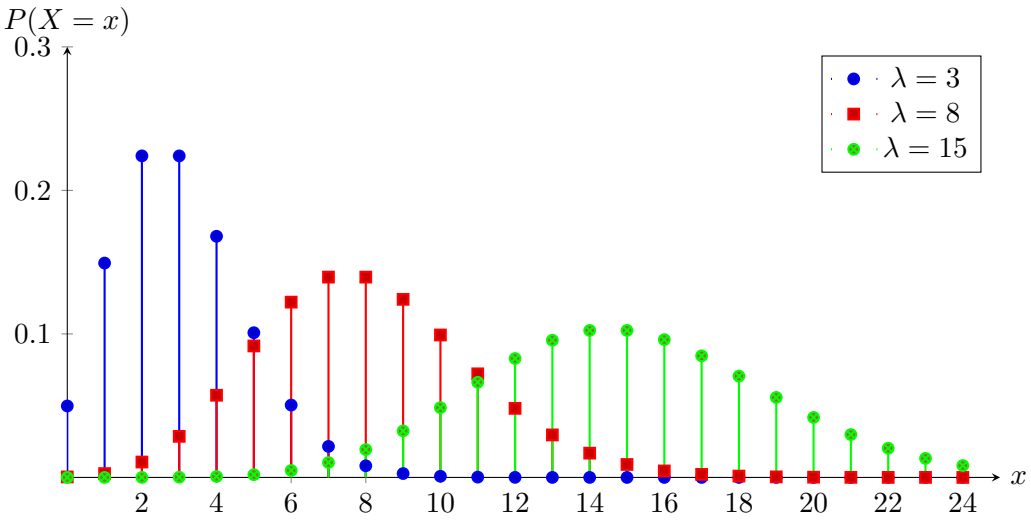


Abbildung 2.2: Graphische Darstellung von drei unterschiedlichen Poisson-Verteilungen. Es ist zu erkennen, dass die jeweiligen Poisson-Verteilungen ihre größte Wahrscheinlichkeit an dem Wert für λ haben, da dieser Wert das Ereignis darstellt, welches im Durchschnitt am häufigsten auftritt.

dargestellt werden. Die Einschränkung, dass die Ereignisse sich in zwei Klassen aufteilen lassen müssen, wird bei der Multinomialverteilung aufgehoben. Die Wahrscheinlichkeitsfunktion einer Multinomialverteilung sieht folgendermaßen aus:

$$P(X_1 = x_1, \ldots, X_k = x_k) = \frac{n!}{x_1! * \cdots * x_k!} * p_1^{x_1} * \cdots * p_k^{x_k}; \quad x \in \mathbb{N} \tag{2.1}$$

Das x steht wieder für das Auftreten eines bestimmten Ereignisses, n für die Anzahl der entnommenen Stichproben und p für die Wahrscheinlichkeiten der jeweiligen Ereignisse.

Als Beispiel für ein Experiment, welches mit einer Binomialverteilung modelliert werden kann ist der Münzwurf. Die Ereignisse sind in diesem Beispiel *Kopf* und Zahl, die in die Klassen *Kopf* und *Nicht Kopf* eingeteilt werden können. Auch das Experiment des Würfelwurfs, bei dem der Betrag der Ereignismenge sechs beträgt, kann in zwei Klassen eingeteilt werden, nämlich in zum Beispiel *Sechs* und *Nicht Sechs*. Um jedoch jedes Ereignis in der Ereignismenge eines Würfelwurfs als eine eigene Klasse ansehen zu können, wird das Würfelwurf Experiment mit einer Multinomialverteilung modelliert werden. Dabei steht x_k für die Häufigkeit der gewürfelten Seite mit der Zahl k, n für die Anzahl der Würfe und p_k für die Wahrscheinlichkeit, dass der Würfel auf der Seite mit der Zahl k landet. Beispielsweise kann mit der Formel (2.1) herausgefunden werden , wie hoch die Wahrscheinlichkeit ist, dass nach zwölf maligen würfeln eines fairen sechsseitigen Würfels, jede Zahl genau 2 mal auftritt.

$$\begin{aligned} P(X_1 = 2, X_2 = 2, X_3 = 2, X_4 = 2, X_5 = 2, X_6 = 2) &= \frac{12!}{2! * 2! * 2! * 2! * 2! * 2!} * \frac{1}{6}^{12} \\ &= 0.00343 \end{aligned}$$

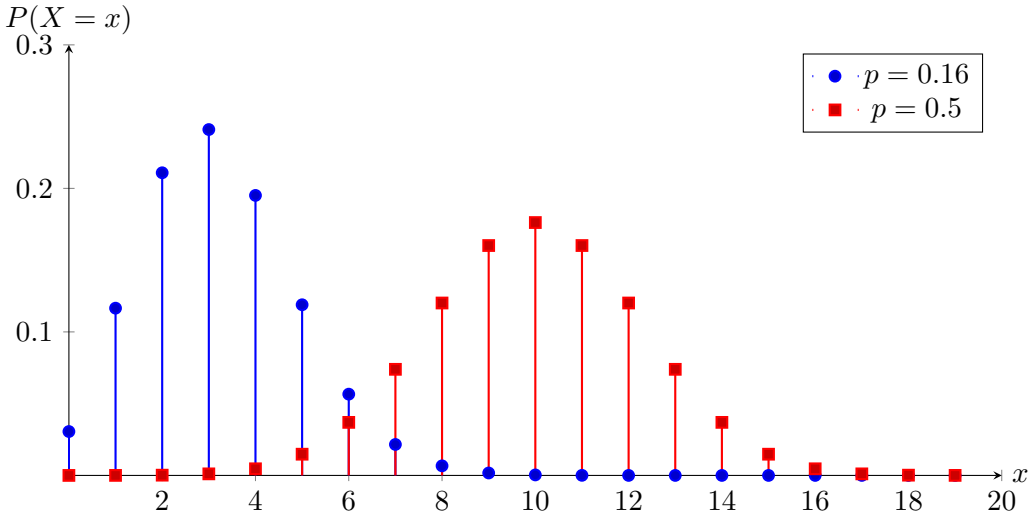


Abbildung 2.3: Binomialverteilung: In rot ist das Experiment des Münzwurfs und in blau das des Würfelwurfs dargestellt. Beide Experimente wurden 20 mal durchgeführt. Dabei wurde bei dem Münzwurf auf die Häufigkeit des Auftretens von Kopf und bei dem Würfelwurf auf die Häufigkeit des Auftretens der Zahl Sechs geprüft.

2.2.3 Bernoulli-Verteilung

Die Bernoulli-Verteilung ist eine diskrete Wahrscheinlichkeitsverteilung, die ein Spezialfall der Binomialverteilung ist. Jedes Experiment, das mit einer Binomialverteilung modelliert werden kann, kann auch mit einer Bernoulli-Verteilung modelliert werden, falls sich die Anzahl der durchgeführten Versuche n auf 1 beschränkt. Eine Variable nimmt den Wert 1 mit der Wahrscheinlichkeit p an, falls das durchgeführte Zufallsexperiment zum Erfolg führt (Beim Münzwurf wurde auf Kopf getippt und es ist Kopf gefallen). Dementsprechend nimmt die Variable den Wert 0 mit der Wahrscheinlichkeit p-1 an, falls das Zufallsexperiment zum Misserfolg führt.

2.2.4 Dirichlet-Verteilung

Die Dirichlet-Verteilung ist eine stetige Wahrscheinlichkeitsverteilung, mit der die Wahrscheinlichkeiten von Multinomialverteilungen modelliert werden können. Die Dichtefunktion einer Dirichlet-Verteilung sieht wie folgt aus:

$$P(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \dots \theta_k^{\alpha_k - 1}$$

wobei $\theta \in [0,1]^k$, $\sum_{i=1}^{k} \theta_i = 1$, α ein k-dimensionaler Vektor mit Komponenten $a_i > 0$ für alle i, $0 \leq i \leq k$ und $\Gamma(x)$ die Gammafunktion bezeichnen. Die Form und die Art der Dirichlet-Verteilung ist dabei von dem Vektor α abhängig.

Wird

$$\alpha_i = c; \quad c \geq 0, \ \forall i : 1 \leq i \leq k \tag{2.2}$$

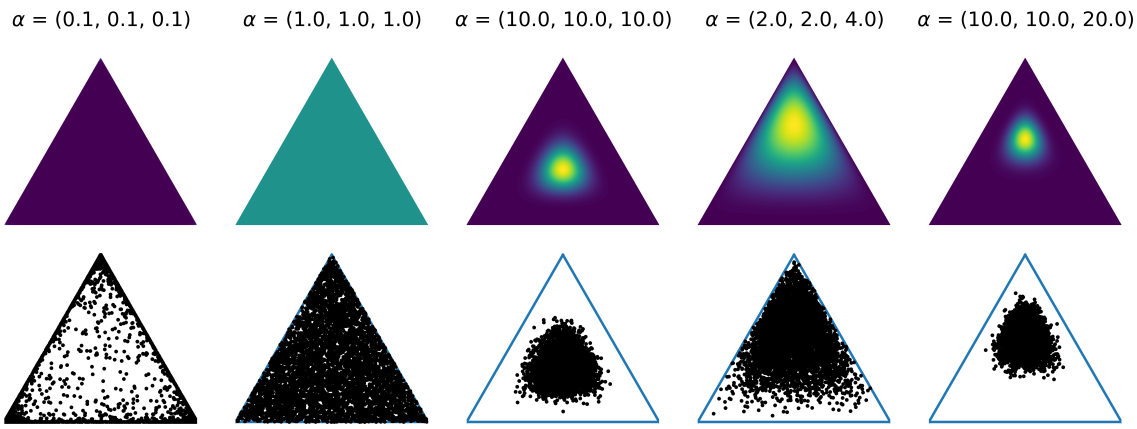


Abbildung 2.4: Die Abbildung zeigt Dirichlet-Verteilungen mit 5 verschieden Werten für α. In der oberen Zeile sind die Dirichlet-Verteilungen als Dreiecke mit unterschiedlichen Farbintensitäten dargestellt. Die Farbintensität beschreibt die Wahrscheinlichkeit für das Ziehen einer Multinomialverteilung in einem Bereich. Je heller die Farbe ist, desto wahrscheinlicher ist das Ziehen der Multinomialverteilung in diesem Bereich.
In der unteren Zeile sind die gezogenen Multinomialverteilung dargestellt. Es wurde aus jeder Dirichlet-Verteilung in der oberen Zeile 5000 Stichproben entnommen und in der unteren Zeile dargestellt.
(Zu beachten ist, dass die Ecken des Dreiecks in der Dirichlet-Verteilung mit dem Parametervektor $\alpha = (0.1, 0.1, 0.1)$ heller sein sollten als sie in der Abbildung gezeigt sind. Dadurch dass die Wahrscheinlichkeit für das Ziehen einer Multinomialverteilung in den Ecken des Dreiecks am größten ist, hätten die Ecken heller eingefärbt sein müssen, jedoch ist der Eckpunkt so klein, dass es auf der Abbildung nicht mehr zu erkennen ist.

gewählt, dann bezeichnet man die daraus entstehende Dirichlet-Verteilung als symmetrisch.

Je näher α in einer symmetrischen Dirichlet-Verteilung an 1 gewählt wird, desto gleichverteilter werden die von der Dirichlet-Verteilung gezogenen Multinomialverteilungen. Um eine symmetrische Dirichlet-Verteilung zu erzeugen, die jeder Multinomialverteilung dieselbe Wahrscheinlichkeit zuweist, muss dementsprechend in Gleichung (2.2) $c = 1$ gewählt werden. Die Variation der Stichproben ist dabei sehr groß, da es durch die gleichverteilte Dirichlet-Verteilung keine Tendenz für das Ziehen von bestimmten Multinomialverteilungen gibt und somit jede Multinomialverteilung gleichwahrscheinlich ist. Falls c kleiner als 1 gewählt wird, folgt daraus eine symmetrische Dirichlet-Verteilung, die hohe Wahrscheinlichkeiten für das Ziehen von Multinomialverteilungen hat, die nur wenigen Ereignissen hohe Wahrscheinlichkeiten zuweist.

Eine asymmetrische Dirichlet-Verteilung entsteht, wenn die Komponenten von α sich mindestens in einem Wert unterscheiden. Eine asymmetrische Dirichlet-Verteilung mit $\alpha = (1000, 1000, 2000)$ kann beispielsweise die Multinomialverteilungen von dreiseitigen

Würfeln beschreiben, die in einer Fabrik hergestellt werden. Die Dirichlet-Verteilung sagt in diesem Fall aus, dass die in der Fabrik hergestellten Würfel mit sehr großer Wahrscheinlichkeit doppelt so häufig auf die Seite 3 fallen werden, als auf die anderen Seiten. Durch die hohe Summe der Komponenten von α, unterscheiden sich die Multinomialverteilungen der Würfel nur minimal, dementsprechend werden nur sehr wenige Würfel hergestellt, die nicht der oben genannten Eigenschaft entsprechen.

Produziert die Fabrik jedoch Würfel, deren Multinomialverteilungen aus einer asymmetrischen Dirichlet-Verteilung mit $\alpha = (1, 1, 2)$ entnommen wurden, resultieren bei der Produktion Würfel, die nur eine Tendenz dazu haben, doppelt so häufig auf die Seite 3 zu fallen als auf die anderen Seiten. Dementsprechend werden öfter Würfel produziert, die häufiger auf die Seite 3 fallen, als auf die anderen Seiten, anstelle von Würfel, die zum Beispiel häufiger auf die Seite 1 fallen. Jedoch tendiert die Fabrik nur dazu und kann folglich auch viele Würfel produziert, die dieser Eigenschaft nicht entsprechen.

In Abbildung 2.4 stellen in der oberen Reihe hellere Farben eine größere Wahrscheinlichkeit für das Ziehen einer Multinomialverteilung in dem entsprechendem Bereich dar. Es wurden von jeder Dirichlet-Verteilung jeweils 5000 Stichproben entnommen und in der unteren Reihe für die jeweiligen Dirichlet-Verteilungen dargestellt. Die ersten drei Dirichlet-Verteilungen sind symmetrisch, die letzten beiden asymmetrisch. Jede Ecke des Dreiecks stellt eine Seite des Würfels da. Dementsprechend stellt ein Punkt an einer Ecke der Dirichlet-Verteilung einen Würfel dar, der mit 100-prozentiger Wahrscheinlichkeit auf eine bestimmten Seite fällt. Ein Punkt genau in der Mitte der Dirichlet-Verteilung stellt einen Würfel dar, der auf jede Seite gleichwahrscheinlich fallen kann. In der ersten Spalte ist eine symmetrische Dirichlet-Verteilung mit dem Parametervektor $\alpha = (0.1, 0.1, 0.1)$ zu sehen. Diese Dirichlet-Verteilung weist den Ecken des Dreiecks eine sehr große und alle anderen Punkten eine sehr kleinere Wahrscheinlichkeit zu. Dies ist gut in der unteren Zeile dieser Spalte zu erkennen, denn dort sind die meisten der 5000 Stichproben in den Ecken und nur sehr wenige in der Mitte. Produziert eine Fabrik Würfel, deren Multinomialverteilungen aus dieser Dirichlet-Verteilung gezogen worden sind, werden mit sehr großer Wahrscheinlichkeit Würfel produziert, die mit sehr hoher Wahrscheinlichkeit auf eine bestimmte Seite fallen und nur mit sehr geringer Wahrscheinlichkeit auf die anderen Seiten. In der zweiten Spalte ist eine symmetrische Dirichlet-Verteilung mit dem Parametervektor $\alpha = (1.0, 1.0, 1.0)$ zu sehen. Diese Dirichlet-Verteilung weist jeder Multinomialverteilung die gleiche Wahrscheinlichkeit zu. Dementsprechend ist das komplette Dreieck in der oberen Zeile dieser Spalte in einer Farbe gefärbt. Auch in der unteren Zeile dieser Spalte ist zu erkennen, dass keine Multinomialverteilung häufiger auftritt als die andere, dementsprechend sind die Multinomialverteilungen gleichverteilt. Lässt sich die Produktion der Firma durch diese Dirichlet-Verteilung modellieren, so entstehen Würfel, deren Eigenschaften nicht vorhergesagt werden können, da jede mögliche Multinomialverteilung gleichverteilt gezogen werden kann. Es entstehen demnach so viele faire Würfel, wie Würfel die nur auf

eine bestimmten Seite fallen können. In der dritten Spalte ist eine symmetrische Dirichlet-Verteilung mit dem Parametervektor $\alpha = (10.0, 10.0, 10.0)$ zu sehen. Es ist zu erkennen, dass die Multinomialverteilungen gehäuft in der Mitte des Dreiecks gezogen worden sind. Am Wahrscheinlichsten ist somit die Multinomialverteilung, die sich genau in der Mitte des Dreieck befindet, also die, die jedem Ereignis die selbe Wahrscheinlichkeit zuweist. Werden Würfel in einer Fabrik produziert, die mit dieser Dirichlet-Verteilung modelliert wird, so entstehen mit der größten Wahrscheinlichkeit faire Würfel, jedoch werden auch sehr viele Würfel produziert, die nicht fair sind. Dies ist in der unteren Abbildung dieser Spalte gut zu erkennen, da viele Punkte nicht genau in der Mitte des Dreiecks liegen. In der vierten Spalte ist eine asymmetrische Dirichlet-Verteilung mit dem Parametervektor $\alpha = (2.0, 2.0, 4.0)$ zu sehen. Diese Dirichlet-Verteilung tendiert dazu, große Wahrscheinlichkeiten näher an die obere Ecke des Dreiecks zuzuweisen, da die dritte Komponente des Parametervektors α großer ist als die beiden anderen. Auch in der unteren Zeile dieser Spalte ist zu erkennen, dass die Punkte immer weniger werden, je größer die Entfernung zu der oberen Ecke ist. Die meisten Würfel, die in einer mit dieser Dirichlet-Verteilung modellierten Fabrik produziert werden, fallen am häufigsten auf die dritte Seite. Jedoch ist es durchaus möglich, dass Würfel produziert werden, die diese Eigenschaft nicht erfüllen, da viele Punkte in der Abbildung der zweiten Zeile weit von der oberen Ecke des Dreiecks entfernt sind. In der letzten Spalte ist eine asymmetrische Dirichlet-Verteilung mit dem Parametervektor $\alpha = (10.0, 10.0, 20.0)$ zu sehen. Diese Dirichlet-Verteilung ist sehr ähnlich zu der in der vierten Spalte, jedoch ist der Bereich, aus dem die Multinomialverteilungen gezogen werden können viel zentrierter. Lässt sich die Produktion der Firma durch diese Dirichlet-Verteilung modellieren, so ist die Wahrscheinlichkeit viel größer, dass die produzierten Würfel häufiger auf die dritte Seite fallen auf die anderen Seiten. In der Abbildung der zweiten Zeile ist zu erkennen, dass sich nur wenige Punkte nicht in der oberen Hälfte des Dreiecks befinden und somit nur wenige Würfel die Eigenschaft nicht erfüllt.

In Abbildung 2.5 wurden aus symmetrischen Dirichlet-Verteilungen mit den Parametervektoren $\alpha = 100, \alpha = 10, \alpha = 1$ und $\alpha = 0.1$ Punkte gezogen und als Multinomialverteilungen dargestellt. Es wurden von jeder Dirichlet-Verteilung 5 Multinomialverteilungen gezogen und verdeutlicht. In der ersten Zeile sind die gezogenen Multinomialverteilungen einer Dirichlet-Verteilung gezeigt, die aus einem Parametervektor $\alpha = 100$ entstanden ist. Zu erkennen ist, dass die gezogenen Multinomialverteilungen sehr nah an eine gleichverteilte Multinomialverteilung kommen. In der zweiten Zeile wurden aus einer Dirichlet-Verteilung mit einem Parametervektor $\alpha = 10$ Stichproben entnommen und angezeigt. Es ist zu beobachten, dass die Multinomialverteilungen ähnlich zu den Multinomialverteilungen in der Zeile darüber sind, jedoch sind diese etwas weniger gleichverteilt. Die dritte Zeile zeigt Multinomialverteilungen, die aus einer Dirichlet-Verteilung mit dem Parametervektor $\alpha = 1$ gezogen worden sind. Auch hier ist festzustellen, dass die Multinomialverteilungen weniger gleichverteilt sind als die in der Zeile darüber. In der letzten Zeile sind Multinomialvertei-

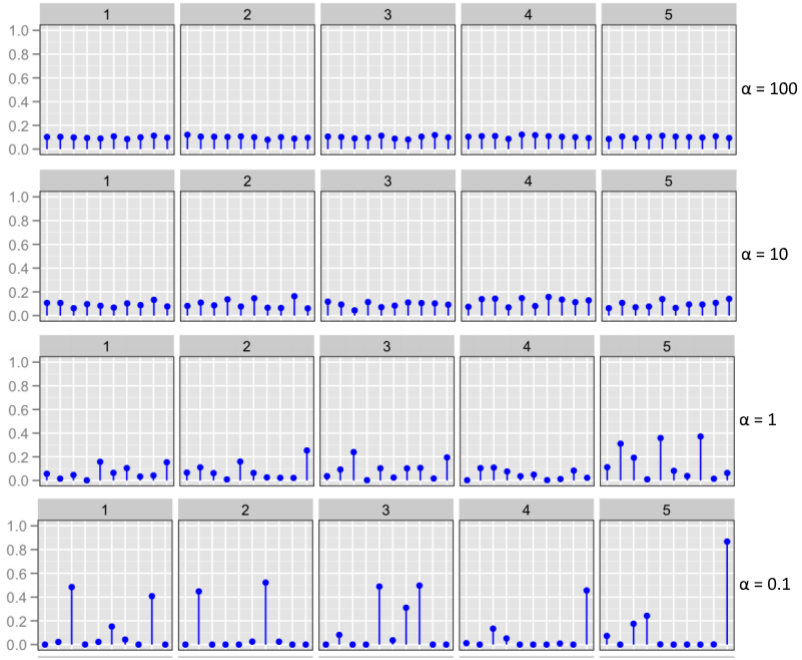


Abbildung 2.5: Gezogene Multinomialverteilungen von 4 verschiedenen symmetrischen Dirichlet-Verteilungen. Die Y-Achse beschreibt die Wahrscheinlichkeit des Auftretens der Ereignisse. In jeder Zeile wurden 5 Stichproben aus der jeweiligen Dirichlet-Verteilung gezogen. Die Multinomialverteilungen beschreiben das Auftreten von insgesamt 10 Ereignissen.

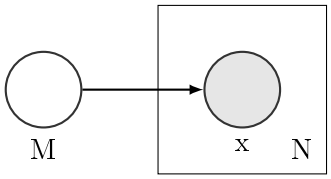
lungen aus einer Dirichlet-Verteilung mit dem Parametervektor $\alpha = 0.1$ gezogen worden. Hier kann beobachtet werden, dass nur wenige Ereignisse eine hohe Wahrscheinlichkeit besitzen. Die Multinomialverteilungen sind somit am wenigsten gleichverteilt. Folglich ist gut zu erkennen, dass mit größer werdendem α , die Multinomialverteilungen gleichverteilter werden.

2.3 Generatives Wahrscheinlichkeitsmodell

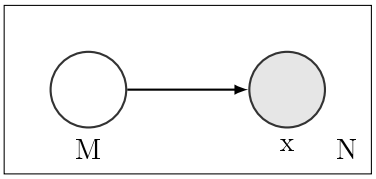
Ein generatives Wahrscheinlichkeitsmodell [12] beschreibt einen probabilistischen Prozess, der das Ziehen von Daten gemäß einer oder mehrerer Wahrscheinlichkeitsverteilungen reflektiert. Dabei kann das Ziehen der Daten von Größen abhängig sein, die dem Modell bekannt oder unbekannt sind.

Als einfaches Beispiel für ein generatives Wahrscheinlichkeitsmodell ist ein probabilistischer Prozess, der das Werfen einer Münze modelliert und somit Daten gemäß der Wahrscheinlichkeitsverteilung der Münze erzeugt. (Siehe Abbildung 2.6a)

Generative Wahrscheinlichkeitsmodelle können auch hierarchisch aufgebaut sein. So kann ein generatives Wahrscheinlichkeitsmodell folgendes Experiment modellieren. (Siehe Abbildung 2.6b)



(a) Beispiel Graphisches Modell für den N-fachen Wurf einer Münze mit unbekannter Wahrscheinlichkeitsverteilung.



(b) Beispiel Graphisches Modell für das N-fache Wählen und Werfen einer Münze mit unbekannter Wahrscheinlichkeitsverteilung.

Abbildung 2.6: Die Abbildungen zeigen graphische Modelle, die generative Wahrscheinlichkeitsmodelle beschreiben.

1. Wähle mit der Wahrscheinlichkeit von $0,5$ eine normale Münze A und mit der Wahrscheinlichkeit von $0,5$ eine Münze B, die auf beiden Seiten Kopf stehen hat, aus.

2. Wirf die gewählte Münze.

Die Wahrscheinlichkeit für das Werfen von Kopf oder Zahl entspricht nach dem generativen Modell:

$$P(x) = P(A) * P(x|A) + P(B) * P(x|B) = \begin{cases} \frac{1}{2} * \frac{1}{2} + \frac{1}{2} * 1 = 0.75; & \text{x = Kopf} \\ \frac{1}{2} * \frac{1}{2} + \frac{1}{2} * 0 = 0.25; & \text{x = Zahl} \end{cases}$$

Ein Wahrscheinlichkeitsmodell ist die Abstraktion vom physikalischen Prozess, der für die Erzeugung der Daten zuständig ist. In dem Modell wird dementsprechend nicht davon ausgegangen, dass es Münzen gibt, die mit einer Wahrscheinlichkeit von 0.75 auf die Kopfseite und mit 0.25 auf die Zahlseite fallen, sondern dass es einen Prozess gibt, der die gleichen Daten mit derselben Wahrscheinlichkeitsverteilung erzeugen kann. Der physikalische Prozess, also wie die Daten in der Realität erzeugt werden, kann dabei unbekannt sein.

Um die Abhängigkeiten der Variablen deutlicher zu machen, werden generative Wahrscheinlichkeitsmodelle oft mit Probabilistischen Graphischen Modellen veranschaulicht (Abbildung 2.6). Die Knoten stellen Variablen dar, wobei latente durch dunkle und nicht latente durch helle Knoten dargestellt werden. Die latenten Variablen sind Variablen, die dem Modell unbekannt sind. Kanten geben die Richtung der Abhängigkeit an und ein Kasten um einen oder mehrere Knoten beschreibt einen wiederholenden Vorgang.

Mit Hilfe statistischen Methoden können die unbekannten Größen eines generativen Wahrscheinlichkeitsmodells geschätzt werden, sodass das geschätzte Modell Daten erzeugen kann, die den Daten in der Realität entsprechen.

2.3.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) wurde von David Blei et al. [3] vorgestellt und ist ein generatives Wahrscheinlichkeitsmodell, welches für Textkorpora ausgelegt ist. LDA ist eine unüberwachte Methode, die es ermöglicht, Dokumente zu analysieren und zu vergleichen. Welche Themen behandelt werden und in welchem Verhältnis diese in einem Dokument zueinander stehen, soll mit LDA herausgefunden werden. Zu betonen ist jedoch, dass das LDA Modell nicht nur auf Texte angewendet werden kann. Unter anderem wird das LDA Modell in der Bioinformatik und zur Analyse von Bildern angewandt. Auf ersteres wird in späteren Kapitel noch genauer eingegangen. Im Folgenden wird das LDA Modell am Beispiel eines Textkorpus erklärt. Um Dokumente mit LDA untersuchen zu können, müssen einige Annahmen für Dokumente getroffen werden. Jedes Dokument muss sich für die Eingabe in das Modell in einer Bag of Words Struktur befinden. Dementsprechend wird angenommen, dass es irrelevant ist in welcher Reihenfolge die Wörter in einem Dokument vorkommen, um herauszufinden, welche Themen in einem Dokument behandelt werden. Des Weiteren wird angenommen, dass jedes Dokument mehrere Themen behandelt, sodass jedes Dokument durch eine Wahrscheinlichkeitsverteilung über Themen beschrieben werden kann. Außerdem soll jedes Thema durch eine Wahrscheinlichkeitsverteilung über das Vokabular des Textkorpus beschrieben werden können. Schließlich wird angenommen, dass für eine Anzahl an Themen K, ein Vokabular V und den Parametervektoren α, $\eta \in \mathbb{R}_{\geq 0}^{|K|}$, jedes Wort in einem Dokument aus dem folgenden probabilistischen Prozess entstanden ist.

1. Wähle die Wörterverteilung $\beta \in [1,0]^{K \times |V|}$ für den Textkorpus entsprechend einer Dirichlet-Verteilung(η)
2. Fuer jedes Dokument d in dem Textkorpus
   (a) Wähle die Anzahl der Wörter N_d von d entsprechend der Poisson-Verteilung(ε)
   (b) Wähle eine Themenverteilung $\theta_d \in [0,1]^K$ für ein Dokument entsprechend der Dirichlet-Verteilung(α)
   (c) Für jedes Wort w_{dn} der N_d Wörter in Dokument d
      i. Wähle ein Thema z_{dn} entsprechend der Multinomialverteilung(θ_d)
      ii. Wähle ein Wort w_{dn} entsprechend der Multinomialverteilung($\beta_{z_{dn}}$)

Jedes Dokument wird demzufolge mit Hilfe einer Themenverteilung θ und einer Wörterverteilung β erzeugt, wobei β im ersten Schritt des probabilistischen Prozesses für jeden Korpus nur ein einziges Mal bestimmt wird.

Die Anzahl der Themen, aus denen ein Dokument besteht, wird immer als bekannt angenommen und vorher festgelegt. Im zweiten Schritt muss für jedes Dokument die Anzahl

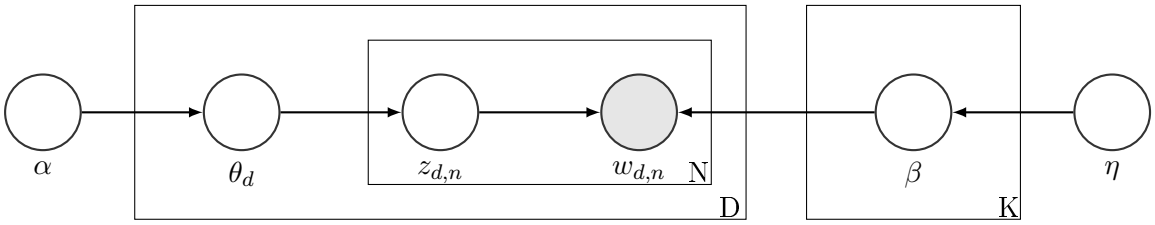


Abbildung 2.7: Probabilistisches Graphisches Modell von LDA

der Wörter mit dem Mittelwert ε und die Themenverteilung mit dem Parametervektor α bestimmt werden. Im letzten Teilschritt wird mit Hilfe der bestimmten Themenverteilung θ erst für jedes Wort im Dokument das Thema z_{dn} bestimmt und anschließend mit der Wörterverteilung β und dem bestimmten Thema z_{dn} das zu generierende Wort festgelegt. Durch den probabilistischen Prozess lässt sich entnehmen, dass die Wahrscheinlichkeit für das Wählen eines bestimmten Themas k für das Dokument d

$$p(z_{dn} = k|\theta_d) = \theta_{dk}$$

und die Wahrscheinlichkeit für die Generierung eines bestimmten Wortes v aus dem Thema k

$$p(w_{dn} = v|z_{dn} = k) = \beta_{kv}.$$

entspricht. In den folgenden Matrizen sind diese Beziehungen verdeutlicht.

$$\theta = \begin{array}{c} \\ z_0 \\ z_1 \\ \vdots \\ z_K \end{array} \begin{array}{c} \begin{array}{cccc} D_0 & D_1 & \dots & D_M \end{array} \\ \begin{pmatrix} \theta_{0,0} & \theta_{0,1} & \dots & \theta_{0,M} \\ \theta_{1,0} & \theta_{1,1} & \dots & \theta_{1,M} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{K,0} & \theta_{K,1} & \dots & \theta_{K,M} \end{pmatrix} \end{array} \qquad \beta = \begin{array}{c} \\ w_0 \\ w_1 \\ \vdots \\ w_V \end{array} \begin{array}{c} \begin{array}{cccc} z_0 & z_1 & \dots & z_K \end{array} \\ \begin{pmatrix} \beta_{0,0} & \beta_{0,1} & \dots & \beta_{0,K} \\ \beta_{1,0} & \beta_{1,1} & \dots & \beta_{1,K} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{V,0} & \beta_{V,1} & \dots & \beta_{V,K} \end{pmatrix} \end{array}$$

Wie in Abbildung 2.7 zu sehen ist, sind die Wörter (w_n) die einzigen Variablen die dem Modell bekannt sind. Aus welchen Themen die Wörter eines Dokuments entstanden sind (z_n), in welchem Verhältnis diese Themen in den Dokumenten zueinander stehen (θ_d), aus welchen Komponenten der Parametervektor für die Dirichlet-Verteilung besteht (α) und wie die Wörterverteilung aussieht (β), ist dem Modell nicht bekannt. Um aus den beobachteten Wörtern auf die latenten Variablen des Modells Rückschließen zu können, wird angenommen, dass die Wörter eines Dokuments mit dem oben genannten probabilistischen Modell erzeugt wurden. Dann wird versucht, mit Hilfe einer Schätzung von α und η herauszufinden, welche Themenverteilung θ und Wörterverteilung β die Wörter der Dokumente

am wahrscheinlichsten erzeugt haben können. Die Komponenten des Parametervektor α werden in der Regel kleiner als 1 gewählt. Dies führt zur Entstehung von Dokumenten, die nur wenige unterschiedliche Themen behandeln, was bei Dokumenten normalerweise der Fall ist.

Die zu approximierende a-posteriori Wahrscheinlichkeit der latenten Variablen ist gegeben durch

$$p(\theta, \beta, z|w, \alpha, \eta) = \frac{p(\theta, \beta, z, w|\alpha, \eta)}{p(w|\alpha, \eta)}.$$

Bei der Betrachtung der hierarchischen Struktur des probabilistischen Prozesses kann die folgende Angabe für die multivariate Verteilung über die Wörterverteilung β, Themenverteilung θ, eine Menge von Themen z und eine Menge von Wörtern w bei gegebenen Parametervektoren α und η gemacht werden:

$$p(\theta, \beta, z, w|\alpha, \eta) = \prod_{k=1}^{K} p(\beta_k|\eta) \prod_{d=1}^{D} p(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} p(z_{dn}|\theta_d) p(w_{dn}|\beta, z_{dn}) \right).$$

Um die marginalisierte Wahrscheinlichkeit von der Menge an Wörtern w in einem Dokument d zu bestimmen, muss über die Themenverteilungen θ integriert und über die Menge der Themen z summiert werden. Dies führt zu:

$$p(w|\alpha, \eta) = \int_{\beta} p(\beta|\eta) \int_{\theta} p(\theta|\alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n|\theta) p(w_n|z_n, \beta) \right) d\theta d\beta.$$

Die marginalisierte Wahrscheinlichkeit lässt sich jedoch in der Realität nicht berechnen und muss daher approximiert werden. Dafür können unter anderem Methoden wie *Gibbs Sampling* oder *variational inference* angewendet werden, wie in [11] und [3] vorgestellt.

Als Beispiel für eine Ausgabe des LDA Modells sind in Abbildung 2.8 vier Themen mit jeweils den wahrscheinlichsten Wörter, die in dem Thema auftreten, aufgelistet. Mit LDA ist es jedoch nicht möglich, die Themen automatisiert zu benennen, da dem Modell die Semantik der Wörter nicht bekannt ist. Es ist aber durchaus zu erkennen, dass die Wörter in den von dem LDA Modell ausgegebenen Themen mit einem Oberbegriff zusammengefasst werden können. Etwa kann die erste Spalte, die unter anderem Wörter wie *human*, *genome*, *dna* und *sequence* enthält, mit dem Oberbegriff Genetik zusammengefasst werden.

2.3.2 Supervised Latent Dirichlet Allocation

Supervised Latent Dirichlet Allocation (sLDA) wurde wie LDA ebenfalls von David Blei et al. [2] vorgestellt und unterscheidet sich insofern davon, dass mit sLDA Labels von Dokumenten bei einer Untersuchung mit betrachtet werden können und somit ein überwachtes Lernverfahren ist. Jedes Dokument muss dafür mit genau einem Label gekennzeichnet sein. Beispielsweise können mit sLDA Rezensionen von Amazon Produkten untersucht werden,

"Genetics"	**"Evolution"**	**"Disease"**	**"Computers"**
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Abbildung 2.8: Beispiel einer Ausgabe (β) von LDA nach der Untersuchung der englischen Zeitschrift *Science* [1]. Jede Spalte entspricht genau einem Thema, wobei die fettgedruckten Wörter die Oberbegriffe der nicht fettgedruckten Wörter bezeichnen. Die Wörter sind absteigend nach der Wahrscheinlichkeit, mit der diese in den jeweiligen Themen auftreten können, sortiert. Zu beachten ist, dass mit LDA nur die nicht fettgedruckten Wörter ausgegeben werden. Die Oberbegriffe wurden durch Interpretation der Wörter von Personen bestimmt.

wobei der geschriebene Kommentar dem Dokument und die Anzahl der Sterne dem Label entspricht. Dies ermöglicht es, mit Hilfe von vielen Beispielrezensionen das sLDA Modell zu trainieren, um somit Vorhersagen über die Anzahl der Sterne in noch nicht betrachteten Kommentaren zu machen.

Genauso wie LDA, nimmt sLDA an, dass die Dokumente mittels eines probabilistischen Prozesses erzeugt werden. Mit der Themenanzahl K, einem Vokabular V, den Parametervektoren α, $\mu \in \mathbb{R}_{\geq 0}^{|K|}$, dem Regressor $\bar{z} = (1/N_d)\sum_{n=1}^{N_d} z_n$, wobei N_d die Anzahl der Wörter in dem Dokument d entsprechen, die Regressionskoeffizienten η und der Varianz σ^2 werden die Wörter der Dokumente und die Zuweisung derer Labels mit dem folgenden probabilistischen Prozess generiert:

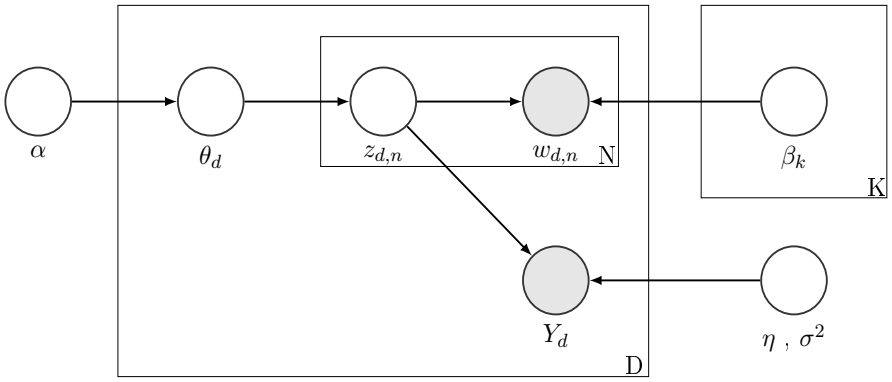


Abbildung 2.9: Probabilistisches Graphisches Modell von sLDA

1. Wähle die Wörterverteilung $\beta \in [1,0]^{K \times |V|}$ für den Textkorpus entsprechend der Dirichlet-Verteilung(μ)
2. Für jedes Dokument d in dem Textkorpus
   (a) Wähle die Anzahl der Wörter N_d entsprechend der Poisson-Verteilung(ε)
   (b) Wähle eine Themenverteilung $\theta_d \in [0,1]^K$ für ein Dokument entsprechend der Dirichlet-Verteilung(α)
   (c) Für jedes Wort w_{dn} der N_d Wörter in Dokument d
      i. Wähle ein Thema z_{dn} entsprechend der Multinomialverteilung(θ_d)
      ii. Wähle ein Wort w_{dn} entsprechend der Multinomialverteilung($\beta_{z_{dn}}$)
   (d) Wähle das Label y mit Hilfe eines Linearen Modells $N(\eta^\intercal \bar{z}, \sigma^2)$

Es ist zu erkennen, dass die Schritte 1. - 2.(c) des probabilistischen Prozesses dem Prozess der Erstellung von Dokumenten von LDA entsprechen. Der Unterschied zu LDA findet sich in Schritt 2.(d), denn da wird ein Label mit Hilfe eines Linearen Modells gewählt und dem Dokument zugewiesen.

Bei einem linearen Modell wird vorausgesetzt, dass die zu vorhersagende Variable linear von dem Regressor abhängig ist. Informell beschreibt der Regressor $\bar{z}$ die relative Häufigkeit der Themen in einem Dokument. Die Variable σ^2 bezeichnet die Varianz einer Normalverteilung, die zuständig für die Berechnung der Abweichung der vorherzusagenden Variable vom tatsächlichen Wert ist. Diese Variable wird vorwiegend Fehler genannt.

In Abbildung 2.9 ist zu erkennen, dass das probabilistische graphische Modell von sLDA sehr ähnlich zu dem des LDAs ist, bis darauf, dass das sLDA Modell die Labels Y_d der jeweiligen Dokumente mit betrachtet. Die Label der jeweiligen Dokumente Y_d ist dem Modell bekannt und ist von den jeweiligen Themenzuweisungen $z_{d,n}$ abhängig. Damit ist sichergestellt, dass das Label von dem Regressor abhängig ist.

Um die Labels der noch ungelabelten Dokumente vorhersagen zu können, muss wie bei LDA die a-posteriori Wahrscheinlichkeit der latenten Variablen approximiert werden.

Es wird vorerst davon ausgegangen, dass die Variablen α , $\beta_{1:K}$, η und σ^2 bekannt sind. Dementsprechend ist die zu approximierende a-posteriori Wahrscheinlichkeit der latenten Variablen gegeben durch

$$p(\theta, z_{1:N}|w_{1:N}, y, \alpha, \beta_{1:K}, \eta, \sigma^2) = \frac{p(\theta, z_{1:N}, w_{1:N}, y|\alpha, \beta_{1:K}, \eta, \sigma^2)}{p(w_{1:N}, y|\alpha, \beta_{1:K}, \eta, \sigma^2)} \tag{2.3}$$

wobei

$$p(\theta, z_{1:N}, w_{1:N}, y|\alpha, \beta_{1:K}, \eta, \sigma^2) = p(\theta|\alpha) \left(\prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta_{1:K}) \right) p(y|z_{1:N}, \eta, \sigma^2) \tag{2.4}$$

und

$$p(w_{1:N}, y|\alpha, \beta_{1:K}, \eta, \sigma^2) = \int p(\theta|\alpha) \sum_{z_1:N} \left(\prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta_{1:K}) \right) p(y|z_{1:N}, \eta, \sigma^2) d\theta. \tag{2.5}$$

Gleichung (2.4) entspricht der multivariaten Verteilung mit der ein Dokument nach dem probabilistischen Prozess erzeugt wird. Die marginalisierte Wahrscheinlichkeit ist gegeben durch Gleichung (2.5). Da Gleichung (2.3) wie bei LDA ebenfalls nicht berechenbar ist, muss diese approximiert werden. Mit der Approximation kann der Parametervektor α für die Dirichlet-Verteilung, die Parameter η und σ^2 für das lineare Modell und die Wörterverteilung der Themen $\beta_{1:K}$ geschätzt werden. Das Verfahren, welches für die Approximation der a-posteriori Wahrscheinlichkeit und die Schätzung der Variablen angewendet werden kann, nennt sich *variational expectation maximization*. Der *variational expectation maximization* Algorithmus besteht aus zwei Schritten, dem *E-Step* und dem *M-Step*, wobei im *E-Step* die a-posteriori Wahrscheinlichkeit approximiert wird und im *M-Step* die Variablen geschätzt werden, wie vorgestellt in [2]. Nachdem alle Variablen für das Modell bestimmt worden sind, können Vorhersagen über Labels von unbekannten Dokumenten mittels sLDA getroffen werden.

2.3.3 Labeled Latent Dirichlet Allocation

Labeled Latent Dirichlet Allocation (lLDA) wurde von Daniel Ramage et al. [9] vorgestellt und ist wie sLDA ein überwachtes Lernverfahren. Im Vergleich zu sLDA können bei lLDA die Daten jedoch mit mehr als nur einem Label gekennzeichnet sein. Dadurch können beispielsweise tweets von Twitter analysiert werden, indem der Text ohne die Hashtags im Tweet ein Dokument und die Hashtags die Labels des Dokuments darstellen. Dadurch ist es möglich einen direkten Zusammenhang zwischen Wörtern und Hashtags zu finden. Dafür muss jedoch festgelegt werden, dass die Labels den Themen entsprechen. Im Vergleich zu LDA und sLDA, in denen jedes Dokument alle Themen behandelt, wird bei lLDA angenommen, dass jedes Dokument nur die Themen behandelt, die den Labels entsprechen. Somit werden alle anderen Themen mit Wahrscheinlichkeit 0 von dem Dokument behandelt.

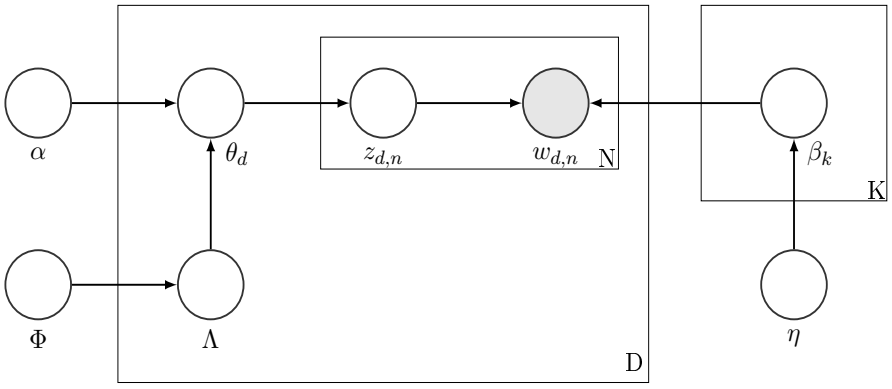


Abbildung 2.10: Probabilistisches Graphisches Modell von lLDA

Mit der Themenanzahl K, die der Anzahl der Labels entsprechen, einem Vokabular V, den Parametervektoren α, $\eta \in \mathbb{R}_{\geq 0}^{|K|}$, einem Erfolgsparametervektor $\Phi_k \in [0,1]^K$ und einer Projektionsmatrix $L^{(d)}$ für die Reduzierung der Dimension vom Parametervektor α, werden die Wörter der Dokumente und deren Zuweisung zu den Labeln mit dem folgenden probabilistischen Prozess erzeugt.

1. Wähle die Wörterverteilung $\beta \in [1,0]^{K \times |V|}$ für den Textkorpus entsprechend der Dirichlet-Verteilung(η)

2. Für jedes Dokument d in dem Textkorpus

   (a) Wähle die Anzahl an Wörter N_d entsprechend der Poisson-Verteilung(ε)

   (b) Für jedes Thema k

      i. Wähle einen binären Themen Anwesenheits/Abwesenheits-Indikatorvektor $\Lambda_k^{(d)} \in \{0,1\}$ entsprechend der Bernoulli-Verteilung(Φ_k).

   (c) Bestimme $\alpha^{(d)} = L^{(d)}\alpha$

   (d) Wähle eine Themenverteilung $\theta_d \in [0,1]^K$ für ein Dokument entsprechend der Dirichlet-Verteilung($\alpha^{(d)}$)

   (e) Für jedes Wort w_{dn} der N_d Wörter in Dokument d

      i. Wähle ein Thema z_{dn} entsprechend der Multinomialverteilung(θ_d)

      ii. Wähle ein Wort w_{dn} entsprechend der Multinomialverteilung($\beta_{z_{dn}}$)

Der probabilistische Prozess ist ähnlich zu dem des LDA Modells, dennoch unterscheidet der sich in einigen Schritten. Nachdem im ersten Schritt wie bei LDA die Wörterverteilung des Textkorpus entsprechend einer Dirichlet-Verteilung bestimmt wird, wird für jedes Dokument die Anzahl der Wörter entsprechend einer Poisson-Verteilung festgelegt. In Schritt 2.(b) wird für jedes Dokument ein binärer Themenindikatorvektor $\Lambda^{(d)}$ der Länge

K bestimmt, welcher an einer Stelle $\Lambda_k^{(d)}$ eine 1 hat, falls das k-te Thema in dem Dokument behandelt wird und eine 0 andernfalls. Dies wird mit der Bernoulli-Verteilung mit dem Erfolgsparameter Φ_k festgelegt. Um sicherzustellen, dass jedes Dokument auch nur die vorher festgelegten Themen behandelt, wird mit der Variable $L^{(d)} \in \{0,1\}^{M_d \times K}$ mit $M_d = |\lambda^{(d)}|$ und $\lambda^{(d)} = \{\ k \mid \Lambda_k^{(d)} = 1\ \}$ der Parametervektor α nach den Themen, die von dem Dokument d behandelt werden, gefiltert. Dies geschieht folgendermaßen:

1. Bestimme den Vektor $\lambda^{(d)}$
2. Für jedes $i \in \{1, \ldots, M_d\}$ und $j \in \{1, \ldots, K\}$, wähle $L_{ij}^{(d)}$ wie folgt:

$$L_{ij}^{(d)} = \begin{cases} 1 & \text{wenn } \lambda_i^{(d)} = j \\ 0 & \text{sonst} \end{cases} \tag{2.6}$$

In Schritt 2.(c) wird also nun mittels der Multiplikation der Matrix $L^{(d)}$ und dem Vektor α die Dimension des Vektors α reduziert, sodass $\alpha^d = (\alpha_{\lambda_1^{(d)}}, \ldots, \alpha_{\lambda_{M_d}^{(d)}})^\intercal$ nur noch die Komponenten besitzt, die die Themen, die von dem Dokument d behandelt werden, repräsentieren.

Um diese Berechnungen deutlicher zu machen, werden sie durch ein Beispiel mit K = 3 verdeutlicht. Für ein Dokument d wird der Themenindikatorvektor $\Lambda^{(d)} = (1,0,1)$ bestimmt. Daraus folgt $\lambda^{(d)} = \{1,3\}$. Mit der Formel (2.6) für die Berechnung von $L_{ij}^{(d)}$ ergibt sich die Matrix:

$$L^{(d)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Als letztes wird die Berechnung von $L^{(d)}\alpha$ durchgeführt, die zu $\alpha^{(d)} = (\alpha_1, \alpha_3)^\intercal$ führt und dementsprechend $\alpha^{(d)}$ zu einem Vektor mit weniger Dimensionen als α macht.

Die letzten beiden Schritte im probabilistischen Prozess sind die gleichen wie bei LDA, wobei nun der vorher berechnete Parametervektor $\alpha^{(d)}$ für die Bestimmung der Dirichlet-Verteilung und somit auch der Themenverteilung benutzt wird.

Wie bei den anderen beiden Methoden auch, müssen bei lLDA die latenten Variablen bestimmt werden. Da der probabilistische Prozess dem des LDA Modells, bis auf, dass der Parametervektor $\alpha^{(d)}$ auf die von dem Dokument behandelten Themen eingeschränkt ist, sehr ähnlich zu dem des lLDA Modells ist, kann für die Bestimmung der latenten Variablen ebenfalls die *Gibbs Sampling* Methode verwendet werden.

Kapitel 3

Wahrscheinlichkeitsmodell für Genmutationsdaten

Wie es in der Einleitung kurz erwähnt worden ist, werden in dieser Arbeit keine üblichen Textdokumente untersucht, sondern echte Genmutationsdaten von Patienten, die an Krebs erkrankt sind. Bei Genmutationsdaten handelt es sich um Daten, die die Gene eines Patienten beschreiben, die sich von den Genen des *Durchschnittsmenschen* unterscheiden. Die Genmutationsdaten müssen dementsprechend vor der Analyse präpariert werden, so dass die vorgestellten Methoden darauf anwendbar sind.

3.1 Textmodelle und genetische Daten

Da es sich sehr gut anbietet, genetische Daten als Textdokumente anzusehen, ist es nicht das erste Mal, dass Textmodelle auf genetischen Daten angewendet worden sind. Die Daten bestehen im Grunde genommen, wie Textdokumente, aus einer Menge von Zeichenketten, die mit einem Begriff zusammengefasst werden können. In Textdokumenten bezeichnen die Zeichenketten die Wörter und der Namen des Textdokuments den zusammenfassenden Begriff. In genetischen Daten entsprechen die Gene den Zeichenketten und die Patientennummer des Patienten, aus dem die genetischen Daten entnommen sind, den zusammenfassenden Begriff.

Beispielsweise wurden in dem Paper [13] genetische Daten von Salmonellen Bakterien mittels *LDA* untersucht, um die Unterschiedlichkeiten des Gens *fliC* in verschiedenen Variationen von Salmonellenbakterien zu beschreiben. In dem Paper [6] wurden mittels *C-Salt* genetischen Daten von Patienten untersucht, um Gemeinsamkeiten und Unterschiede zwischen Zellen aus dem Normal-, Tumor- und Rezidivgewebe zu finden. Diese Zusammenhänge können beispielsweise hilfreich für das Finden der Gene sein, die für das Wiederauftreten des Tumors verantwortlich sind.

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	13264N	13264R	13264T
1	10329	.	AC	A	19.2034	.	...|DDX11L1|...|Noncoding|...	GT:DP:...	0/1...	1/1...	0/0...
1	10354	.	C	A,AC	67.9597	.	...|WASH7P|...|Noncoding|...	GT:DP:...	1/2...	0/1...	1/2...
1	14522	.	G	A	29.4968	.	...|MIR6859-1.3|...|Noncoding|...	GT:DP:...	0/0...	0/1...	0/1...

Abbildung 3.1: Beispielausschnitt der genetischen Daten eines Patienten. Jede Zeile beschreibt eine Mutation eines Gens. Um die Abbildung kompakter zu machen, wurden einige Informationen, die für das Verständnis der Daten irrelevant sind, mit drei Punkten ersetzt.

3.2 Die Genmutationsdaten

Die zu untersuchenden Genmutationsdaten befinden sich in einem Format namens *varient call format* (VCF) [4] und wurden schon mal in [6] und [10] mit anderen als hier vorgestellten Methoden untersucht. Die ersten zu untersuchenden Genmutationsdaten beinhalten Informationen über 18 Patienten, von denen jeweils zwei Stichproben aus dem Normal- und Tumorgewebe entnommen worden sind. Von 16 Patienten wurde jeweils eine Stichprobe und von einem Patienten wurden 4 Stichproben aus dem Rezidivgewebe entnommen. Folglich ergeben sich insgesamt 56 Stichproben, die untersucht werden. Alle Patienten sind am Neuroblastom erkrankt. Der zweite Datensatz enthält die Genmutationsdaten von insgesamt 98 Patienten, von denen jeweils zwei Stichproben aus dem Normal- und Tumorgewebe entnommen worden sind. Alle Patienten sind an Krebs erkrankt.

In Abbildung 3.1 wird deutlicher gemacht, welche Informationen über die Gene des Patienten gespeichert worden sind. Nicht alle Informationen werden für die Untersuchung mit LDA, sLDA und lLDA benötigt, dementsprechend werden die wichtigsten im Folgenden kurz erläutert.

- #**CHROM**. Um welches Chromosom es sich handelt.
- **POS**. Die Startposition der Mutation.
- **REF**. Die Folge der Basen, die bei einem *Durchschnittsmenschen* an der gegebenen Position auftritt. Es wird davon ausgegangen, dass diese Basen nicht mutiert sind und somit das Gen, in dem die Basenfolge auftritt, völlig funktionsfähig ist.
- **ALT**. Die Folge der Basen die, bei einem Patienten an der gegebenen Position ausgelesen wurde. Diese Basenfolge entspricht nicht der Basenfolge des *Durchschnittsmenschen* an der gleichen Position. Dementsprechend kann es sein, dass das Gen, in dem die Basenfolge auftritt, nicht mehr die selbe Funktion erfüllt wie beim *Durchschnittsmenschen*. Falls in unterschiedlichen Stichproben an der selben Position eine andere Basenfolge gemessen wurde, werden die gemessenen Basen mit Kommata getrennt.
- **QUAL**. Die Qualität der Messung. Je höher die zugewiesene Zahl ist, desto unwahrscheinlicher ist es, dass die gemessenen Basen in der Realität einer anderen entsprechen.

- **INFO**. Wichtige Informationen über die Mutation. In dieser Spalte steht unter anderem in welchem Gen die Mutation aufgetreten ist, um was für eine Mutation es sich handelt und ob sie in einem wichtigen Bereich aufgetreten ist.

- **13264N/13264R/13264T**. In welcher Stichprobe die Mutation aufgetreten ist. Die Buchstaben hinter den Zahlen stehen dabei für eine Stichprobe aus dem Normalgewebe (N), dem Tumorgewebe (T) und dem Rezidivgewebe (R). Da sich in der *ALT* Spalte möglicherweise mehrere Mutationen befinden, wird in den jeweiligen Spalten unterschieden, welche Mutation in welcher Stichprobe aufgetreten ist.

In der ersten Zeile der Abbildung 3.1 ist eine Mutation im Chromosom mit der Nummer *1* an der Position *10329* aufgetreten. Bei dem *Durchschnittsmenschen* ist in diesem Chromosom an der genannten Position die Basenfolge AC codiert. Bei dem Patienten wurde an der selben Position jedoch nur die Base A gelesen. Dies entspricht der Löschung einer Base. Die Base C ist bei dem Patienten an dieser Position nicht vorhanden und wurde dementsprechend gelöscht. Die Qualität der Messung entspricht *19.2034* und ist somit eine sehr unsichere Messung. Die Mutation ist in einem nicht codierenden (*Noncoding*) Bereich des Gens *DDX11L1* aufgetreten. Da der Mensch von jedem Chromosom zwei Sätze besitzt, wird in der Spalte der Stichproben unterschieden, in welchem der beiden Sätzen die Mutation aufgetreten ist. Im Normalgewebe ist die Mutation nur im zweiten Satz aufgetreten. Im Rezidivgewebe ist das Gen in beiden Sätzen mutiert und im Tumorgewebe ist es überhaupt nicht mutiert. In der zweiten Zeile ist zu beobachten, dass bei dem Patienten die Basen A und AC codiert sind, obwohl beim *Durchschnittsmenschen* die Base C codiert ist. Die Qualität der Messung beträgt *67.9597*, was auf eine nicht sehr genaue Messung deutet. Die Mutation ist in einem nicht codierenden (*Noncoding*) Bereich des Gens *WASH7P* aufgetreten. Im Normalgewebe ist im ersten Chromosomensatz die Base C zur Base A mutiert. Im zweiten Chromosomensatz wurde aus der Base C die Basen AC, demnach wurde die Base A an der gegebenen Position eingefügt. Im Rezidivgewebe ist nur im zweiten Chromosomensatz die Base C zur Base A mutiert. Im Rezidivgewebe sind die selben Mutationen wie im Normalgewebe aufgetreten. In der dritten Zeile ist bei dem Patienten statt der Base G die Base A gemessen worden. Die Qualität der Messung entspricht *29.4568* und ist folglich keine sichere Messung. Die Mutation ist im nicht codierenden Bereich (*Noncoding*) des Gens *MIR6859-1.3* aufgetreten, wobei diese nur im zweiten Chromosomensatz des Rezidiv- und Tumorgewebes aufgetreten ist.

Je nachdem welche Informationen in diesen Spalten enthalten sind, sollen die Daten zur Untersuchung genutzt werden oder nicht. Demzufolge müssen die Genmutationsdaten vor der Analyse präpariert werden, sodass nur die relevanten Informationen betrachtet werden.

3.3 Präparieren der Daten

Um die Genmutationsdaten mit den genannten Methoden untersuchen zu können, muss vorerst festgelegt werden, welche Genmutationen relevant sind und somit in den zu untersuchenden Datensatz hinzugefügt werden. Die Datensätze können mit zwei unterschiedlichen Verfahren bestimmt werden. In dem ersten Verfahren werden alle Gene, die sich bei den Patienten beim Vergleich mit dem *Durchschnittsmenschen* als mutiert herausstellen, dem Datensatz hinzugefügt. Im zweiten Verfahren werden nur die Gene dem zu untersuchenden Datensatz hinzugefügt. die in einem Gewebe eines Patienten mutiert sind, aber in einem anderen Gewebe des selben Patienten nicht. Dies wird für alle Patienten wiederholt und als Datensatz zusammengefasst. Dementsprechend enthält der Datensatz nur die Genmutationen, die sich bei der Betrachtung von zwei Gewebearten unterscheiden. Ist beispielsweise ein bestimmtes Gen an einer bestimmten Position eines Chromosoms eines Patienten im Normalgewebe und im Tumorgewebe mutiert, dann wird das Gen beim ersten Verfahren in den Datensatz hinzugefügt, jedoch beim zweiten Verfahren nicht. Von Interesse sind nicht nur die Unterschiede zwischen dem Tumorgewebe und dem Normalgewebe (TvN), sondern auch die zwischen dem Rezidivgewebe und dem Tumorgewebe (RvT) und die zwischen dem Rezidivgewebe und dem Normalgewebe (RvN).

Da die Modelle für die Eingabe annehmen, dass es sich um einen Textkorpus handelt, dessen Dokumente sich in einer Bag of Words Struktur befinden, stellt sich nun die Frage, was in dem zu untersuchenden Datensatz dem Textkorpus, dem Dokument und dem Wort entspricht.

Je nachdem mit welchem Verfahren die Datensätze präpariert werden und welches Modell auf den Datensatz angewendet werden soll, wird die Eingabe für das Modell anders gewählt. Da das unüberwachte LDA Modell Labels nicht mit betrachten kann, muss jedes Gewebe unabhängig voneinander betrachtet werden. Wird der Neuroblastomdatensatz mit dem ersten Verfahren bestimmt, dann entsteht für jedes Gewebe ein Textkorpus, welches mit LDA untersucht werden soll. Jedes dieser Textkorpora enthält ein Dokument für jede Stichprobe aus dem Gewebe des Patienten und somit für das Normal- und Tumorgewebe insgesamt 18 Dokumente und für das Rezidivgewebe 20 Dokumente. Jedes Dokument enthält die in der entsprechenden Stichprobe mutierten Gene des Patienten. Wird der Datensatz mit dem zweiten Verfahren bestimmt, dann entstehen drei Textkorpora, die aus Dokumenten bestehen, die die Unterschiede zwischen TvN, RvT und RvN kennzeichnen. Dadurch entstehen Dokumente, die aus einer Menge von Genen und der Anzahl der Mutationen des Gens bestehen und daher einer Bag of Words Struktur entsprechen .

Bei den überwachten Modellen sieht es etwas anders aus. Da diese Modelle die Labels der Dokumente mit betrachten können, ist es kontraproduktiv für jedes Gewebe einen Textkorpus zu erzeugen, da somit die Beziehungen zwischen den Gewebearten nicht herausgefunden werden können. Demnach wird aus dem kompletten Neuroblastomdatensatz nur

ein Textkorpus erstellt, welcher somit 56 Dokumente enthält und für den zweiten Datensatz ein Textkorpus erstellt, welcher 196 Dokumente enthält. Somit entsprechen die Dokumente den Stichproben und die mutierten Gene den Wörtern. Die unüberwachten Modelle werden nur auf dem Datensatz angewendet, welcher mit dem ersten Verfahren bestimmt worden ist, da beim zweiten Verfahren, den Dokumenten, die aus zwei Gewebearten entstanden sind, keine eindeutige Gewebeart als Label zugewiesen werden kann.

Es soll jedoch nicht jedes mutierte Gen in den zu untersuchenden Datensatz aufgenommen werden, da es sich je nach der Qualität der Messung auch um einen Fehler handeln könnte. Demzufolge werden nur die mutierten Gene aufgenommen, bei denen eine große Wahrscheinlichkeit besteht, dass die Messung nicht fehlerhaft war. Folglich müssen die Daten nach dem Qualitätsmaß gefiltert werden. Es hat sich herausgestellt, dass eine Messung ab einer Qualität von 200 als eine Fehlerfreie Messung gewertet werden kann und somit auch in den Datensatz aufgenommen werden kann [10].

Da es Abschnitte in der DNA gibt, die bei heutigem Forschungsstand bei der Transkription für den Aufbau der RNAs keine Rolle spielen, sollen diese Abschnitte nicht in den Datensatz aufgenommen werden, da eine Mutation in diesen Abschnitten dementsprechend keine Folgen haben kann. Bei einem wichtigen Abschnitt handelt es sich um den codogenen Strang. Jede Mutation, die nicht in einem codogenen Strang aufgetreten ist, wird dementsprechend nicht mit in den Datensatz aufgenommen.

Mutationen können in Kategorien eingeteilt werden, welche die Art der Mutation und somit auch wie bedeutend die Mutation ist beschreiben. Je nachdem um welche Art der Mutation es sich handelt, sollen diese in den Datensatz mit aufgenommen werden oder nicht. Ein Beispiel für eine bedeutende Mutation, die für den Datensatz relevant ist, ist die *missense vaiant*, welche einen Austausch einer Base in einem codogenen Strang beschreibt, wodurch nach der Mutation ein anderes Protein als vorher hergestellt wird. Andere bedeutende Mutationen sind unter anderem *stop gained*, *frameshift variant* und *splice region variant*.

Zusammenfassend kann also gesagt werden, dass jede Mutation, welche diese Eigenschaften

- Qualität der Messung ≥ 200
- Befindet sich im codogenen Strang
- Ist eine bedeutende Mutation (z.B. *missense variant*)

erfüllt, dem Datensatz hinzugefügt wird.

Wird ein Datensatz mit dem ersten Verfahren bestimmt, dann werden Gene, die im kompletten Datensatz weniger als 5 Mal mutiert sind, nicht in den zu untersuchenden Datensatz aufgenommen. Da diese Gene nur selten mutiert sind und somit auch nur bei

sehr wenigen Patienten aufgetreten sind, wird davon ausgegangen, dass das Gen keine Verbindung zu dem Neuroblastom hat.

Kapitel 4

Experimente

Bisher war immer die Rede von *Themen*, wenn es um die Untersuchung von Dokumenten mittels LDA ging, jedoch wurde dieser Begriff nur so gewählt, um eine Intuition für das, was mit LDA herausgefunden wird, zu erlangen. Da im Folgenden Experimente auf Genmutationsdaten durchgeführt werden, ist es nicht mehr so einfach zu verstehen, was mit Themen im Bezug auf Genmutationsdaten gemeint ist. Tatsächlich werden mit LDA nämlich latente Subtypen herausgefunden. Werden die Subtypen eines mit LDA untersuchtem Dokuments interpretiert, wird schnell festgestellt, dass die Subtypen oft ziemlich genau dem entsprechen, was wir unter einem Thema verstehen. Doch im Bezug auf die Genmutationsdaten ist es nicht so einfach herauszufinden, welche Bedeutung die Subtypen haben und ob sie in Genmutationsdaten überhaupt vorhanden sind.

Wenn es möglich ist, die latenten Subtypen der Genmutationsdaten der Patienten, die am Neuroblastom erkrankt sind, herauszufinden, können Biologen diese Subtypen interpretierten. Dadurch kann abhängig von den Subtypen, aus denen ein Patient besteht, möglicherweise herausgefunden werden, ob sich bei einem Patienten Rezidivgewebe bilden wird oder nicht.

4.1 Vorgehen

Zu Beginn werden die Daten nach den in Abschnitt 3.3 genannten Kriterien präpariert. Die folgenden Methoden werden auf dem Neuroblastomdatensatz angewendet, der mit dem ersten Verfahren aus Abschnitt 3.3 bestimmt wurde.

Nach dem Präparieren müssen die Parameter für Modelle gewählt werden. Bei LDA und sLDA muss vorher festgelegt werden, wie viele Subtypen in dem Datensatz erwartet werden. Dies ist bei Genmutationsdaten besonders schwer, da die Anzahl oft durch das Wiederholen des Experiments und der Interpretation der Subtypen festgelegt wird. Wie schon erwähnt wurde, ist die Interpretation der Subtypen bei Genmutationsdaten nicht

so einfach wie bei Textdokumenten. Der Parametervektor α wird $1/K$ für jede Methode gewählt, so wie es in [2] auch gemacht wurde.

Zu Beginn soll untersucht werden, welche Themenanzahl am Wahrscheinlichsten zu den besten Ergebnissen führen wird. Dafür werden die Daten mittels LDA und 5, 10 und 15 Themen analysiert. Um herauszufinden, welche Anzahl an Themen für die genetischen Daten zu den besten Ergebnissen führt, wird betrachtet, wie wahrscheinlich die jeweiligen Themen in den Dokumenten vorkommen. Werden die Themen miteinander verglichen und es stellt sich heraus, dass ein Thema im Verhältnis zu den anderen Themen mit sehr geringer Wahrscheinlichkeit in den Dokumenten vorkommt, dann ist das Thema weniger aussagekräftig als die anderen Themen. Dementsprechend wird zuerst mit Hilfe der von LDA ausgegebenen Themenverteilung θ die Summe über alle Wahrscheinlichkeiten für das Auftreten des jeweiligen Themas in einem Dokument gebildet.

$$\Theta^{k_K} = \sum_{d=1}^{|D|} \theta_d^{k_K} \tag{4.1}$$

Dabei steht k für ein bestimmtes Thema, K für die Themenanzahl, mit der der Datensatz mittels LDA untersucht wurde und $|D|$ für die Anzahl der Dokumente im Datensatz.

Nach der Bestimmung der geeignetsten Themenanzahl, wird LDA auf den Datensatz mit dieser Themenanzahl angewendet und grafisch dargestellt. Anschließend wird überprüft, welche Themen bei der Untersuchung von 2 Stichproben sich am ähnlichsten sind, diese einander zugewiesen und die größten Unterschiede in den Wörterverteilungen berechnet. Die Themenpaarungen werden mit

$$\Psi_{k,k'}^{S,S'} = \sqrt{\sum_{i=1}^{|V|} (\beta_{i,k}^{S} - \beta_{i,k'}^{S'})^2} \tag{4.2}$$

bestimmt. Die Variablen k und k' entsprechen Themennummern, wobei k ein Thema aus der mit LDA analysierten Stichprobe S ist und k' ein Thema aus der mit LDA analysierten Stichprobe S' ist. V entspricht dem Vokabular des Datensatzes und β einer Wörterverteilung. Es entsteht die Matrix $\Psi \in [0, n]^{K \times K}$ mit $n \in \mathbb{N}$. Aus der Matrix Ψ werden nach und nach die Indizes i und j bestimmt, die den kleinsten Wert beinhalten und dementsprechend die $k = i$ und $k' = j$ einander zugewiesen. Nach der Bestimmung der Themenpaare wird mit

$$\psi_{k,k'} = \beta_{i,k}^{S} - \beta_{i,k'}^{S'} \; ; \qquad 1 \geq i \geq |V| \tag{4.3}$$

bestimmt, was die größten Unterschiede in der Wörterverteilung der Themen sind.

Anschließend werden die Daten mit sLDA untersucht, die Beziehungen zwischen den Stichproben innerhalb der jeweiligen Themen veranschaulicht und die Wörterverteilung dargestellt. Außerdem wird versucht, mit Hilfe von sLDA vorherzusagen, aus welchem Gewebe eine Stichprobe entnommen worden ist. Da nur insgesamt 56 Stichproben vorhanden

sind, wird mittels Kreuzvalidierung bestimmt, wie gut die Vorhersage ist. Danach wird der Datensatz mit lLDA untersucht.

Nachdem alle Methoden auf dem Datensatz angewendet worden sind, soll der Datensatz mit dem zweiten Verfahren aus Abschnitt 3.3 bestimmt und untersucht werden. Es sollen generelle Unterschiede zwischen den Gewebearten herausgefunden und anschließend der Datensatz mit LDA untersucht werden.

Abschließend wird der Datensatz, der die 98 Patienten, die an unterschiedlichen Krebsarten erkrankt sind, mit LDA und sLDA untersucht. Es soll mit sLDA wieder versucht werden herauszufinden, aus welchem Gewebe eine Stichprobe entnommen worden ist.

4.2 Umsetzung

Der Datensatz wird mit der Programmiersprache Julia und der Webapplikation Jupyter Notebook[2] präpariert. Für die Anwendung von LDA wird das Programm RapidMiner Studio[3] verwendet. Die Anwendung von sLDA wird mit einer Implementierung[4] in C++ realisiert und für lLDA gibt es bereits eine Implementierung[5] in Python, die für diese Arbeit benutzt wird.

Alle Berechnungen werden auf folgender Hardware durchgeführt:

Prozessor:	AMD Ryzen 7 1700X @ 3,8GHz × 8
Arbeitsspeicher:	16 GB
Grafik:	AMD Radeon R390
Betriebssystem:	Windows 10

4.3 Ergebnisse

4.3.1 Teil 1

In Teil 1 sind die Ergebnisse der Anwendung der in Kapitel 2 vorgestellten Methoden auf den Neuroblastom Datensatz. Dieser Datensatz enthält alle Genmutationen, die die in Abschnitt 3.3 genannten Eigenschaften erfüllen.

Latent Dirichlet Allocation

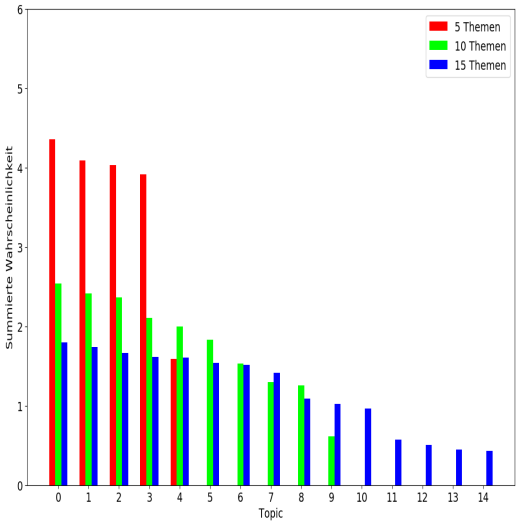
In Abbildung 4.1 wurden mit der Formel (4.1) die summierten Wahrscheinlichkeiten für jedes Thema berechnet und dargestellt. In jeder dieser Abbildungen wurde in rot mit der von LDA mit 5 Themen ausgegebene Themenverteilung θ die summierte Themenverteilung

[2] https://jupyter.org
[3] https://rapidminer.com/
[4] http://www.cs.cmu.edu/ chongw/slda/
[5] https://github.com/akullpp/SLDA

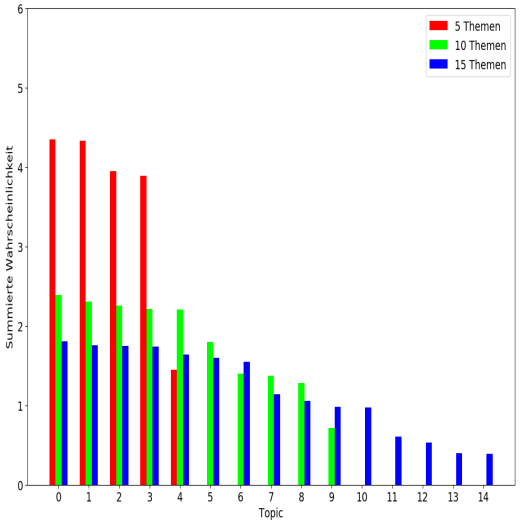
Θ^{k_5} veranschaulicht. In grün ist $\Theta^{k_{10}}$ mit 10 Themen und in blau ist $\Theta^{k_{15}}$ mit 15 Themen illustriert. Da die Labels der Genmutationsdaten bei LDA nicht mit betrachtet werden können, musste das Verfahren für jedes Label wiederholt werden. In Abbildung 4.1a wurde Θ mit der Themenverteilung θ des Normalgewebes bestimmt und dargestellt. In Abbildung 4.1b wurde das gleiche für die Themenverteilung des Tumorgewebes und in 4.1c für die Themenverteilung des Rezidivgewebes gemacht.

In den folgenden Experimenten wird die Themenanzahl auf 5 festgelegt. Die Ergebnisse der Ausführung von LDA auf die Gewebearten ist in der Abbildung 4.2 zu finden. Von dem Patient mit der Patientennummer 2 wurde keine Stichprobe aus dem Rezidivgewebe entnommen, deshalb sind die Wahrscheinlichkeit des Auftretens eines Themas in 4.2e an der Stelle 2 für jedes Thema 0. Da von dem Patient mit der Patientennummer 3 insgesamt vier Stichproben aus dem Rezidivgewebe entnommen worden sind, wurde aus Übersichtlichkeitsgründen in Abbildung 4.2e der durchschnittliche Wert der Themen in den 4 Stichproben bestimmt und dargestellt.

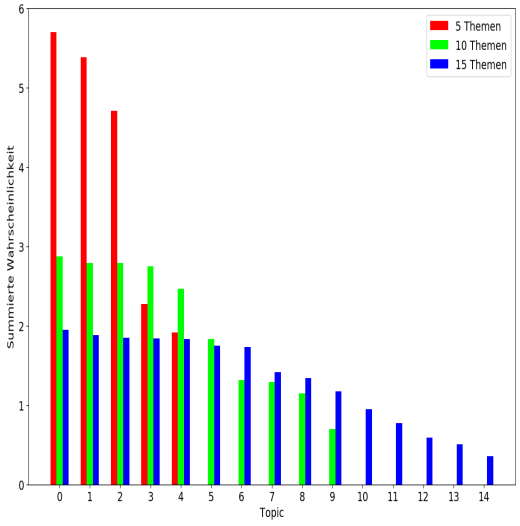
In Abbildung 4.3a, 4.4a und 4.5a ist die Berechnung von Ψ mit der Formel (4.2) und in den Abbildungen 4.3b, 4.4b und 4.5b die Themenpaarungen zu finden. Die Abbildungen 4.6, 4.7 und 4.8 zeigen die Berechnungen von ψ mit der Formel (4.3). Für jede Themenpaarung wurden 20 Wörter dargestellt, deren Wahrscheinlichkeiten sich am meisten unterscheiden. Die ersten 10 Wörter sind am wahrscheinlichsten im Tumorgewebe und die letzten 10 Wörter am wahrscheinlichsten im Rezidivgewebe aufgetreten.



(a) $\Theta^{k_5}, \Theta^{k_{10}}, \Theta^{k_{15}}$ für das Normalgewebe

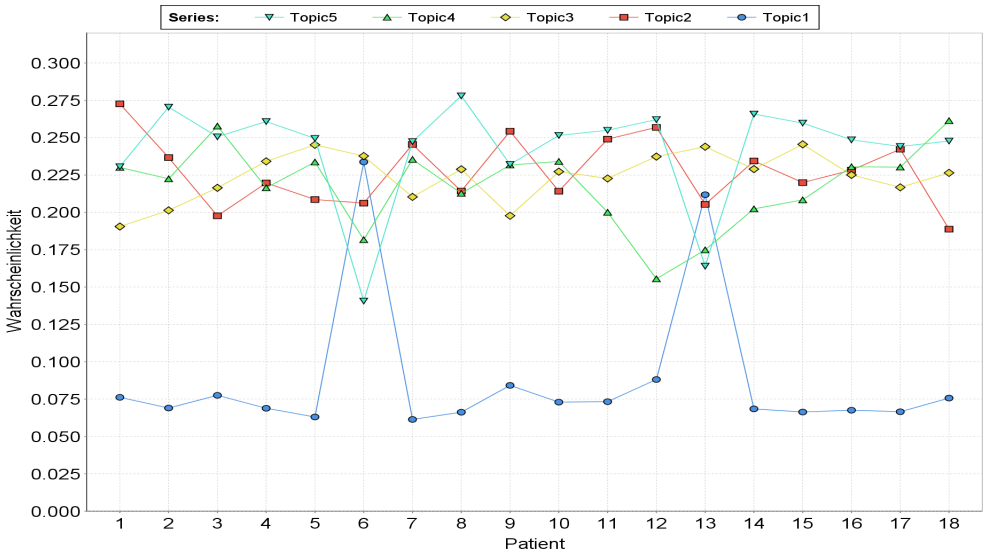


(b) $\Theta^{k_5}, \Theta^{k_{10}}, \Theta^{k_{15}}$ für das Tumorgewebe



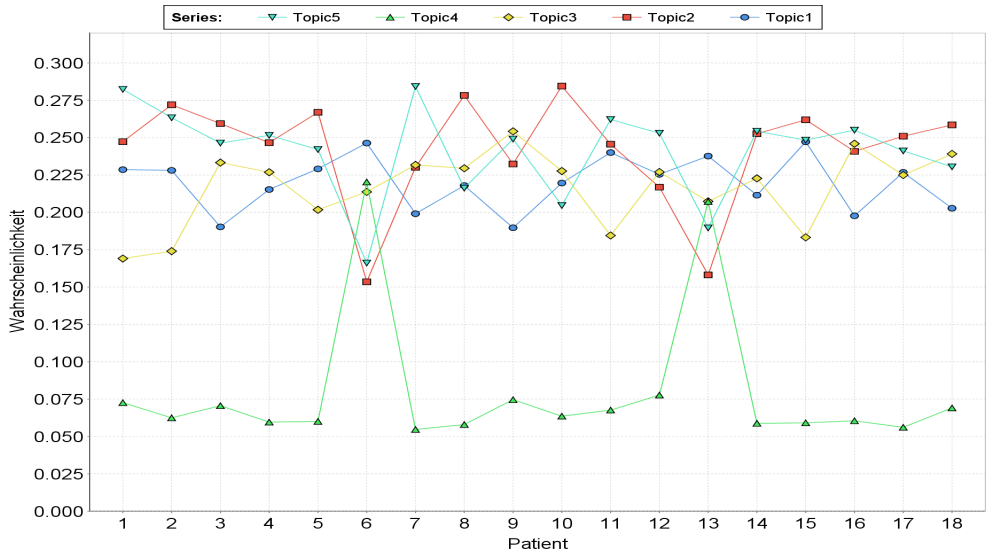
(c) $\Theta^{k_5}, \Theta^{k_{10}}, \Theta^{k_{15}}$ für das Rezidivgewebe

Abbildung 4.1: Die Y-Achse beschreibt die Summe der Wahrscheinlichkeiten der Themen über alle Dokumente, die X-Achse ist mit der Themennummer indiziert. Der Datensatz wurde mit LDA und 5, 10 und 15 Themen untersucht und die Wahrscheinlichkeit des Auftretens eines Themas über alle Dokumente summiert. In Abbildung **(a)** wurden nur die mutierten Gene im Normalgewebe analysiert. In Abbildung **(b)** wurden die mutierten Gene im Tumorgewebe und in Abbildung **(c)** die im Rezidivgewebe untersucht.



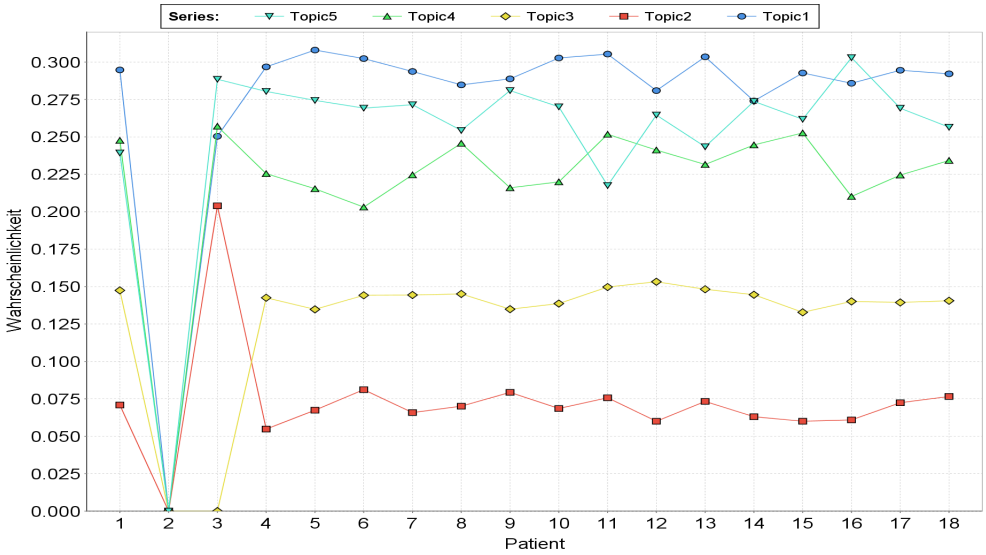
(a) Die Themenverteilung nach der Untersuchung des Normalgewebes von 18 Patienten mit LDA.

(b) Die Woerterverteilung nach der Untersuchung des Normalgewebes von 18 Patienten mit LDA.



(c) Die Themenverteilung nach der Untersuchung des Tumorgewebes von 18 Patienten mit LDA.

(d) Die Woerterverteilung nach der Untersuchung des Tumorgewebes von 18 Patienten mit LDA.



(e) Die Themenverteilung nach der Untersuchung des Rezidivgewebes von 18 Patienten mit LDA.

(f) Die Wörterverteilung nach der Untersuchung des Rezidivgewebes von 18 Patienten mit LDA.

Abbildung 4.2: Die von LDA ausgegebene Themenverteilung und Wörterverteilung nach der Untersuchung der im Normal-, Tumor- und Rezidivgewebe mutierten Gene der Patienten. In den Abbildungen 4.2a, 4.2c und 4.2e sind die Themenverteilungen abgebildet. Die Y-Achse beschreibt die Wahrscheinlichkeit, dass das jeweilige Thema in einem Dokument vorkommt. Die X-Achse ist mit den Patientennummern gekennzeichnet. Die Abbildungen 4.2b, 4.2d und 4.2f zeigen die Wörterverteilung der Themen in dem jeweiligen Gewebe. Aus welchem Thema die Wörter entnommen worden sind, ist an der Farbe zu erkennen. Hat ein Wort in der Wörterverteilung die selbe Farbe wie die in der Themenverteilung, dann gehört das Wort zu diesem Thema. Je größer das Wort ist, desto wahrscheinlicher tritt das Wort in dem Thema auf. Es werden die 10 wahrscheinlichsten Wörter pro Thema dargestellt.

	Topic1	Topic2	Topic3	Topic4	Topic5
1	0.04676244	0.05531436	0.0487803	0.01413076	0.05326571
2	0.03919226	0.05090548	0.04324846	0.05018185	0.03420381
3	0.02725401	0.04633594	0.04008443	0.04877327	0.04412148
4	0.04613469	0.03732019	0.04475099	0.05592655	0.04709021
5	0.04210465	0.03367305	0.04741944	0.05303958	0.03807883

(a) Die Ergebnisse der Berechnung von $\Psi^{N,T}_{k,k'}$ für jede mögliche Paarung aus k und k' sind in dieser Tabelle zu sehen.

	Normalgewebe	Tumorgewebe
1	1	4
2	2	5
3	3	1
4	4	3
5	5	2

(b) Es wurden mit $\Psi^{N,T}_{k,k'}$ die ähnlichsten Themen bestimmt und in dieser Tabelle zusammengefasst.

Abbildung 4.3: Die Spalten in Abbildung 4.3a bezeichnen die Ähnlichkeiten von allen Themen k' aus T zu einem Thema k aus N. Die Zeilen bezeichnen die Ähnlichkeiten von allen Themen k aus N zu einem Thema k' aus T. In Abbildung 4.3b wurden die ähnlichsten Themen einander zugewiesen.

	Topic1	Topic2	Topic3	Topic4	Topic5
1	0.05054537	0.05806389	0.04871151	0.0471518	0.04796385
2	0.04464402	0.05362399	0.03619787	0.04098845	0.04062072
3	0.0381897	0.05467021	0.04474811	0.03629768	0.03945907
4	0.0370055	0.05189889	0.05413844	0.04517182	0.04198422
5	0.03990436	0.05536404	0.04790296	0.03297155	0.03867876

(a) Die Ergebnisse der Berechnung von $\Psi^{N,R}_{k,k'}$ für jede mögliche Paarung aus k und k' sind in dieser Tabelle zu sehen.

	Normalgewebe	Rezidivgewebe
1	1	2
2	2	3
3	3	5
4	4	1
5	5	4

(b) Es wurden mit $\Psi^{N,R}_{k,k'}$ die ähnlichsten Themen bestimmt und in dieser Tabelle zusammengefasst.

Abbildung 4.4: Die Spalten in Abbildung 4.4a bezeichnen die Ähnlichkeiten von allen Themen k' aus R zu einem Thema k aus N. Die Zeilen bezeichnen die Ähnlichkeiten von allen Themen k aus N zu einem Thema k' aus R. In Abbildung 4.4b wurden die ähnlichsten Themen einander zugewiesen.

	Topic1	Topic2	Topic3	Topic4	Topic5
1	0.03756943	0.05401693	0.04126904	0.03531497	0.04177126
2	0.03325513	0.05579752	0.05235927	0.042953	0.04503937
3	0.04821013	0.05551043	0.04716411	0.04560495	0.03254726
4	0.05004735	0.05944504	0.05081144	0.04854153	0.04996621
5	0.04378043	0.05228817	0.04465825	0.03404577	0.04206871

(a) Die Ergebnisse der Berechnung von $\Psi^{T,R}_{k,k'}$ für jede mögliche Paarung aus k und k' sind in dieser Tabelle zu sehen.

	Tumorgewebe	Rezidivgewebe
1	1	3
2	2	1
3	3	5
4	4	2
5	5	4

(b) Es wurden mit $\Psi^{T,R}_{k,k'}$ die ähnlichsten Themen bestimmt und in dieser Tabelle zusammengefasst.

Abbildung 4.5: Die Spalten in Abbildung 4.5a bezeichnen die Ähnlichkeiten von allen Themen k' aus R zu einem Thema k aus T. Die Zeilen bezeichnen die Ähnlichkeiten von allen Themen k aus T zu einem Thema k' aus R. In Abbildung 4.5b wurden die ähnlichsten Themen einander zugewiesen.

	Gen	Ähnlichkeit
1	ZNF83	0.00283951
2	DST	0.00259433
3	MUC19	0.00226581
4	ELN	0.00221202
5	PABPC3	0.00209574
6	IRF3	0.00196244
7	TTN	0.00182173
8	YY1AP1	0.00148526
9	SIGLEC6	0.0014535
10	FLNB	0.00145141
11	LRP8	-0.00126348
12	TBCK	-0.00134464
13	PCDH15	-0.00146657
14	KIAA0226L	-0.00152076
15	MYO9B	-0.0015574
16	CREM	-0.00160102
17	DYSF	-0.00164139
18	OPA1	-0.00235419
19	PDE4DIP	-0.00424502
20	NEB	-0.00436416

(a) $\psi_{1,4}$

	Gen	Ähnlichkeit
1	TTN	0.01599057
2	ZNF568	0.00446121
3	PCDH15	0.00320371
4	SERPINA1	0.00319022
5	BRDT	0.00318132
6	KIAA1683	0.00315733
7	HLA-A	0.00313329
8	SFI1	0.00296617
9	DAG1	0.00280317
10	LTBP4	0.00256398
11	ATP7B	-0.00247754
12	RP1L1	-0.00257287
13	PSG4	-0.00258004
14	RIF1	-0.0027622
15	SP110	-0.00312049
16	DMKN	-0.00319707
17	GAA	-0.00325983
18	SORBS1	-0.00587432
19	MICA	-0.00897042
20	PKD1L2	-0.01156697

(b) $\psi_{2,5}$

	Gen	Ähnlichkeit
1	MUC16	0.00932425
2	NEB	0.00711492
3	SYNE2	0.00475522
4	ZNF595	0.00430159
5	KIAA0586	0.00396808
6	AKAP13	0.00316566
7	EIF4G1	0.0028838
8	TPO	0.00279097
9	IQCE	0.00263982
10	C19orf48	0.002589
11	CCDC129	-0.00238066
12	SYNRG	-0.00254171
13	CDH23	-0.00258955
14	LTBP4	-0.00261593
15	DAG1	-0.00289773
16	WNK1	-0.00294203
17	ZNF83	-0.00319666
18	ELN	-0.00328555
19	KIAA1683	-0.00356572
20	ZNF568	-0.00449784

(c) $\psi_{3,1}$

	Gen	Ähnlichkeit
1	PDE4DIP	0.01906325
2	HLA-C	0.01119508
3	MICA	0.00992468
4	MUC12	0.00608544
5	HLA-B	0.00484158
6	FLG	0.00410546
7	SP110	0.00343509
8	TPTE	0.00340992
9	CYP2A7	0.00337052
10	PDXDC1	0.00331753
11	NLRP1	-0.00330742
12	SERPINA1	-0.0033535
13	KIAA0586	-0.00383382
14	MKI67	-0.00409579
15	ZNF595	-0.00448447
16	SYNE2	-0.00510405
17	LILRB2	-0.00528076
18	EMR1	-0.00708586
19	MUC16	-0.00973022
20	TTN	-0.01782485

(d) $\psi_{4,3}$

	Gen	Ähnlichkeit
1	PKD1L2	0.01105932
2	EMR1	0.00639453
3	SORBS1	0.00586518
4	LILRB2	0.00477442
5	GAA	0.00328046
6	DMKN	0.00324639
7	C17orf80	0.00307408
8	LGALS8	0.00275497
9	PSG4	0.00259793
10	CLEC4M	0.00238921
11	MST1L	-0.00288355
12	CYP2A7	-0.00298497
13	PDXDC1	-0.00299273
14	FLG	-0.00376617
15	HLA-A	-0.00422529
16	HLA-B	-0.00447062
17	MUC12	-0.00525466
18	NEB	-0.00528395
19	HLA-C	-0.01023594
20	PDE4DIP	-0.01587487

(e) $\psi_{5,2}$

Abbildung 4.6: Es sind die Wörter dargestellt, die sich in den jeweiligen Themenpaarungen des Normal- und Tumorgewebes, die in Abbildung 4.3b zu sehen sind, am meisten unterscheiden.

	Gen	Ähnlichkeit
1	TTN	0.01568558
2	CACNA1G	0.01248065
3	CACNA1C	0.00737043
4	NEB	0.00674162
5	DYSF	0.00621728
6	DNMT3B	0.00435882
7	MOV10L1	0.00427124
8	CACNB2	0.00426404
9	RAD17	0.00378691
10	FSIP2	0.00359523
11	LILRA1	-0.00535043
12	CLCNKA	-0.00565905
13	DDX11	-0.00592066
14	CTCFL	-0.00832001
15	HLA-DRB5	-0.00978869
16	OBSCN	-0.01000847
17	MUC5B	-0.01098861
18	LILRB1	-0.0111607
19	HLA-C	-0.01149075
20	HLA-DRB1	-0.02222227

(a) $\psi_{1,2}$

	Gen	Ähnlichkeit
1	PCDH15	0.00640409
2	NRAP	0.00499548
3	MUC4	0.00492026
4	HRNR	0.0045437
5	DISC1	0.00345443
6	BRDT	0.00330989
7	PDLIM5	0.00318093
8	KIAA1683	0.00315056
9	HLA-A	0.00314076
10	SFI1	0.00300438
11	ZNF568	-0.00311328
12	FCGR3A	-0.00332115
13	CCDC169	-0.00350959
14	POLR1B	-0.00382531
15	NAV2	-0.00425776
16	ZNF595	-0.00430548
17	ZNF302	-0.00507595
18	GPR56	-0.00508329
19	ATXN3	-0.00598241
20	ANKLE1	-0.00698513

(b) $\psi_{2,3}$

	Gen	Ähnlichkeit
1	NEB	0.00711311
2	UMODL1	0.00570308
3	ZAN	0.00483448
4	SYNE2	0.00476382
5	SYNE1	0.00437452
6	NAV2	0.00378024
7	TACC2	0.00375009
8	KIAA0586	0.00325283
9	ZNF595	0.00295968
10	EIF4G1	0.00287958
11	CHIA	-0.00528433
12	ZNF419	-0.0054186
13	LILRB2	-0.00582339
14	MUC4	-0.00622586
15	HLA-DQB1	-0.00636037
16	EMR1	-0.00658831
17	HLA-A	-0.00685895
18	MICA	-0.00837752
19	PLEC	-0.01035033
20	TTN	-0.01096014

(c) $\psi_{3,5}$

	Gen	Ähnlichkeit
1	PLEC	0.01185207
2	MICA	0.00987221
3	HLA-DQB1	0.0085645
4	HLA-C	0.0055746
5	EYS	0.00414638
6	FLG	0.00412106
7	LILRB1	0.00378687
8	HLA-DRB5	0.00364842
9	MUC5B	0.00353485
10	CYP2A7	0.00341922
11	FN1	-0.00313512
12	CTAGE5	-0.00316795
13	ZAN	-0.00345884
14	MKI67	-0.00350926
15	HRNR	-0.00380512
16	SYNE2	-0.00391553
17	NRAP	-0.00402465
18	UMODL1	-0.00463158
19	OBSCN	-0.00477147
20	NEB	-0.00691529

(d) $\psi_{4,1}$

	Gen	Ähnlichkeit
1	OBSCN	0.00898774
2	EMR1	0.00642168
3	LILRB2	0.00634024
4	ZNF419	0.0059647
5	PDE4DIP	0.00588231
6	CHIA	0.00559014
7	HLA-A	0.00526127
8	ANKLE1	0.00367002
9	IGFN1	0.00362162
10	C17orf80	0.00349812
11	PROM1	-0.00267145
12	ZNF133	-0.00267292
13	GRK4	-0.00267552
14	KIAA0586	-0.00273696
15	ALPK1	-0.00273876
16	FGFR4	-0.00275021
17	SFI1	-0.00288233
18	ZFYVE16	-0.00289708
19	CYP2A7	-0.00314726
20	CACNA1G	-0.00372571

(e) $\psi_{5,4}$

Abbildung 4.7: Es sind die Wörter dargestellt, die sich in den jeweiligen Themenpaarungen des Normal- und Rezidivgewebes, die in Abbildung 4.3b zu sehen sind, am meisten unterscheiden.

	Gen	Ähnlichkeit
1	UMODL1	0.00566631
2	ZAN	0.00424026
3	SYNE1	0.00420665
4	TACC2	0.00389371
5	KIAA1683	0.00361568
6	ELN	0.0032948
7	ZNF83	0.00297593
8	WNK1	0.0029488
9	MYLK	0.00291455
10	LAMA3	0.0028879
11	POLR1B	-0.00383829
12	ZNF595	-0.00429358
13	MUC4	-0.00435763
14	ZNF302	-0.0050753
15	SERPINA1	-0.00588091
16	MUC20	-0.00592895
17	ATXN3	-0.00598264
18	ANKLE1	-0.00698803
19	HLA-DRB1	-0.01012562
20	TTN	-0.01335999

(a) $\psi_{1,3}$

	Gen	Ähnlichkeit
1	HLA-A	0.00736012
2	ZNF419	0.00613022
3	CHIA	0.00555842
4	HLA-C	0.00461427
5	OBSCN	0.00400754
6	FLG	0.00377543
7	IGFN1	0.00348967
8	ANKLE1	0.00347552
9	TPO	0.00320009
10	ATXN3	0.00312265
11	FN1	-0.00314755
12	CTAGE5	-0.00315294
13	ZAN	-0.00333814
14	MKI67	-0.00340563
15	SYNE2	-0.00382169
16	HRNR	-0.00385545
17	DMD	-0.00390884
18	NRAP	-0.00407209
19	AHNAK2	-0.00411597
20	UMODL1	-0.00464307

(b) $\psi_{2,1}$

	Gen	Ähnlichkeit
1	TTN	0.00683612
2	AHNAK2	0.00637846
3	DMD	0.00527733
4	SYNE2	0.0051037
5	MKI67	0.00407895
6	LILRB1	0.0039092
7	MUC5B	0.00382265
8	HLA-DRB5	0.00374412
9	CTCFL	0.00372789
10	SERPINA1	0.00335614
11	PCDH15	-0.00265314
12	SORBS1	-0.00279077
13	APOBEC3H	-0.00286292
14	GAA	-0.00303605
15	IGFN1	-0.00330862
16	CHIA	-0.00521135
17	ZNF419	-0.0053974
18	MUC4	-0.00622853
19	HLA-A	-0.00632288
20	MICA	-0.00837177

(c) $\psi_{3,5}$

	Gen	Ähnlichkeit
1	TTN	0.01386385
2	CACNA1G	0.01347778
3	NEB	0.01110578
4	CACNA1C	0.00799742
5	DYSF	0.00785867
6	RAD17	0.00441195
7	CACNB2	0.00432209
8	PDE4DIP	0.00417269
9	MOV10L1	0.00411005
10	LIMCH1	0.00408471
11	HLA-DQB1	-0.00549246
12	CLCNKA	-0.00565659
13	DDX11	-0.00592828
14	CTCFL	-0.00844085
15	HLA-DRB5	-0.00979032
16	OBSCN	-0.01000601
17	LILRB1	-0.01115902
18	HLA-C	-0.01149007
19	MUC5B	-0.01161221
20	HLA-DRB1	-0.02200202

(d) $\psi_{4,2}$

	Gen	Ähnlichkeit
1	HLA-DRB1	0.01228593
2	MICA	0.00897138
3	MUC4	0.00880135
4	NRAP	0.00482531
5	HRNR	0.00444201
6	MUC20	0.00414841
7	CTAGE5	0.0038221
8	FN1	0.00371119
9	GAA	0.00328023
10	PCDH15	0.0031695
11	PLIN4	-0.00242842
12	NBPF1	-0.00246623
13	PARP4	-0.00247625
14	SLFN11	-0.00250059
15	KIAA0586	-0.00255706
16	C19orf48	-0.00266468
17	GRK4	-0.00266799
18	SFI1	-0.00287374
19	CYP2A7	-0.00314629
20	CACNA1G	-0.00372382

(e) $\psi_{5,4}$

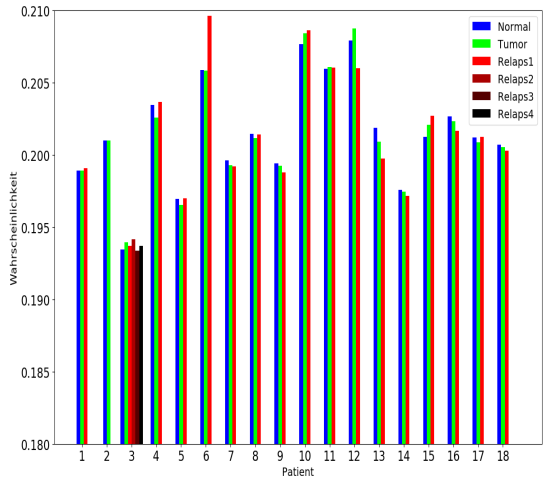
Abbildung 4.8: Es sind die Wörter dargestellt, die sich in den jeweiligen Themenpaarungen des Tumor- und Rezidivgewebes, die in Abbildung 4.3b zu sehen sind, am meisten unterscheiden.

Subervised Latent Dirichlet Allocation

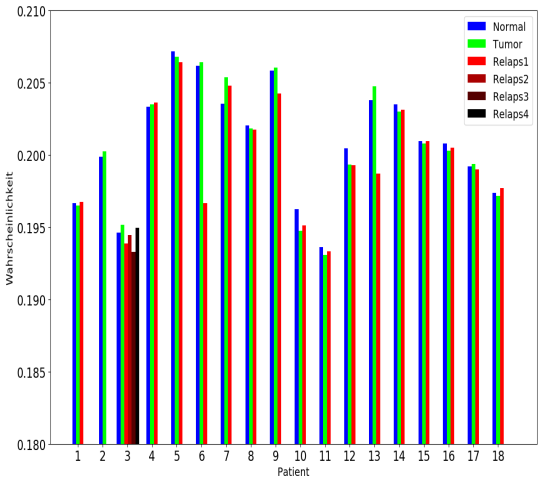
Die Abbildung 4.9 zeigt die von sLDA ausgegebenen Themenverteilungen für jedes Thema. Da von dem Patienten mit der Patientennummer 2 keine Stichprobe aus dem Rezidivgewebe entnommen worden ist, zeigen die Grafiken keinen Balken, die die Wahrscheinlichkeit für das Auftreten des jeweiligen Themas in dem Rezidivgewebe des Patienten darstellt. Bei dem Patienten mit der Patientennummer 3 werden vier Balken angezeigt, da von diesem Patienten vier Stichproben aus dem Rezidivgewebe entnommen worden sind.

In Abbildung 4.10 ist die Wörterverteilung der von sLDA bestimmten Themen zu finden.

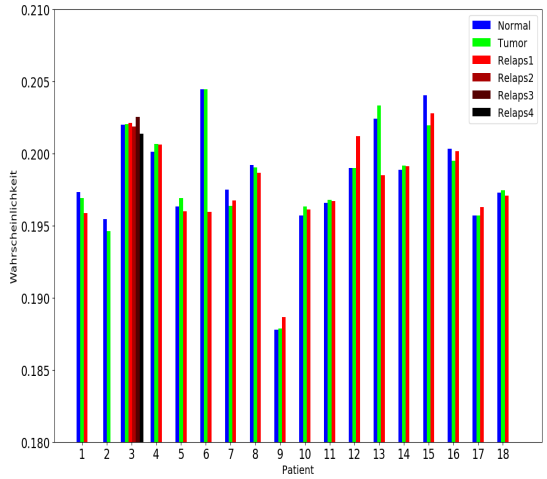
Die Durchschnittliche Genauigkeit der Vorhersage der Labels beträgt ungefähr 33%. Die Genauigkeit ist unabhängig davon, welcher Wert für die Themenanzahl und für den Parametervektor α gewählt wird.



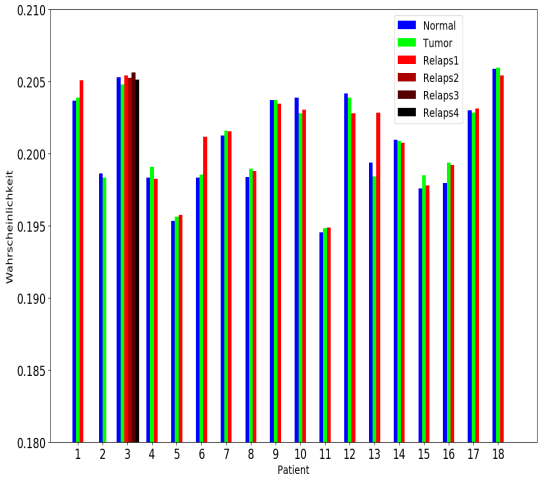
(a) Wahrscheinlichkeitsverteilung von Thema 1 über alle Stichproben



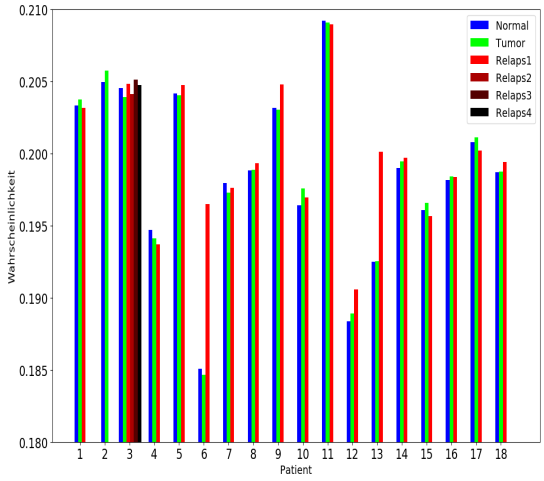
(b) Wahrscheinlichkeitsverteilung von Thema 2 über alle Stichproben



(c) Wahrscheinlichkeitsverteilung von Thema 3 über alle Stichproben



(d) Wahrscheinlichkeitsverteilung von Thema 4 über alle Stichproben



(e) Wahrscheinlichkeitsverteilung von Thema 5 über alle Stichproben

Abbildung 4.9: Die Y-Achse bezeichnet die Wahrscheinlichkeit für das Auftreten des jeweiligen Themas in einer Stichprobe. Die X-Achse kennzeichnet die Patientennummer.

Abbildung 4.10: Die Wordcloud zeigt die von sLDA ausgegeben Wörterverteilung. Pro Thema sind die 10 wahrscheinlichsten Wörter dargestellt. Je größer die Wahrscheinlichkeit ist, desto größer ist das Wort.

Labeled Latent Dirichlet Allocation

In der Abbildung 4.11 befindet sich die Themenverteilung, die bei der Untersuchung des Datensatzes mittels lLDA bestimmt wird. Da es insgesamt 20 Stichproben aus dem Rezidivgewebe und nur 18 aus dem Normal- und Tumorgewebe gibt, zeigt 4.11c die Themenverteilung für 2 Stichproben mehr als in 4.11a und 4.11b. Weil den Dokumenten, die jeweils einer bestimmten Stichprobe aus einem Gewebe entsprechen, nur zwei Labels zugewiesen worden sind, behandelt jedes Dokument auch nur genau 2 Themen. Dementsprechend ist die Wahrscheinlichkeit gleich Null, dass ein Dokument, das aus dem Normalgewebe entstanden ist, das Thema *Tumor* oder *Relaps* behandelt, und wird deswegen nicht in der Abbildung 4.11a abgebildet.

Die Wörterverteilung der 4 Themen ist in den Abbildungen 4.12 bis 4.15 als Wordcloud zu finden. Je dunkler ein Wort in der Wordcloud dargestellt ist, desto seltener kommt das Wort in den jeweils anderen Themen vor.

Patient	Common	Normal
1	0.999062876523	0.000937123477174
2	0.999107771777	0.000892228222587
3	0.995853518351	0.00414648164874
4	0.998701763811	0.00129823618945
5	0.998996899508	0.00100310049243
6	0.570906117257	0.429093882743
7	0.999323806519	0.000676193481495
8	0.999426908834	0.000573091165579
9	0.993316589332	0.00668341066792
10	0.999121188154	0.000878811846384
11	0.997286525905	0.0027134740948
12	0.992629316018	0.00737068398192
13	0.663764351702	0.336235648298
14	0.999169834404	0.00083016559619
15	0.999596557289	0.000403442711135
16	0.999293754138	0.000706245861841
17	0.999260998087	0.000739001912711
18	0.998406374502	0.00159362549801

(a) Themenverteilung der Stichproben aus dem Normalgewebe

Patient	Common	Tumor
1	0.998304933536	0.00169506646445
2	0.998724369271	0.00127563072853
3	0.989639861865	0.0103601381352
4	0.999057492931	0.000942507068803
5	0.999453178401	0.000546821599453
6	0.572469045885	0.427530954115
7	0.999580829957	0.000419170043314
8	0.999560323602	0.000439676398171
9	0.992046863642	0.00795313635781
10	0.999050547066	0.000949452934262
11	0.998729728585	0.00127027141466
12	0.997095880618	0.00290411938165
13	0.638108938007	0.361891061993
14	0.999213939473	0.000786060526661
15	0.999306390456	0.000693609544067
16	0.999466215916	0.000533784084338
17	0.999433230152	0.000566769847844
18	0.998373054261	0.00162694573916

(b) Themenverteilung der Stichproben aus dem Tumorgewebe

Patient	Common	Relaps
1	0.996647335991	0.0033526640087
3	0.664209832825	0.335790167175
3	0.662660057827	0.337339942173
3	0.667210906568	0.332789093432
3	0.655767968494	0.344232031506
4	0.998616009643	0.00138399035671
5	0.997923812963	0.00207618703737
6	0.929754523061	0.0702454769388
7	0.998508297595	0.00149170240537
8	0.99737210932	0.00262789067975
9	0.98846387371	0.0115361262902
10	0.998656094611	0.00134390538906
11	0.997422026879	0.00257797312146
12	0.998985667034	0.00101433296582
13	0.92877894557	0.0712210544305
14	0.996989528796	0.00301047120419
15	0.998391618032	0.00160838196774
16	0.996510379384	0.0034896206156
17	0.996967965901	0.0030320340994
18	0.995311186801	0.00468881319945

(c) Themenverteilung der Stichproben aus dem Rezidivgewebe

Abbildung 4.11: Der Datensatz wurde mit lLDA untersucht und die Themenverteilung für jede Stichprobe in Tabellen dargestellt.

Abbildung 4.12: Wordcloud der 70 wahrscheinlichsten Wörter in dem Thema *Common*.

Abbildung 4.13: Wordcloud der 70 wahrscheinlichsten Wörter in dem Thema *Normal*.

Abbildung 4.14: Wordcloud der 70 wahrscheinlichsten Wörter in dem Thema *Tumor*.

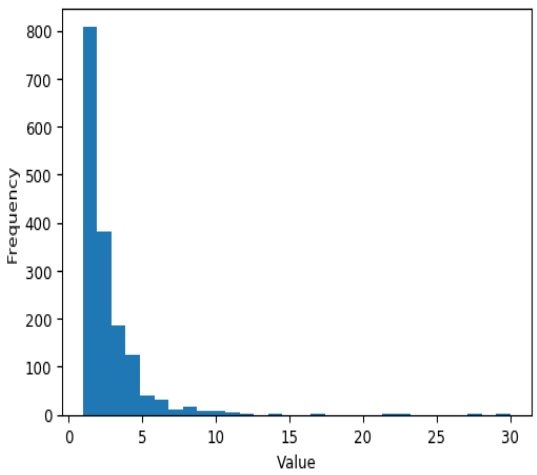
Abbildung 4.15: Wordcloud der 70 wahrscheinlichsten Wörter in dem Thema *Relaps*.

4.3.2 Teil 2

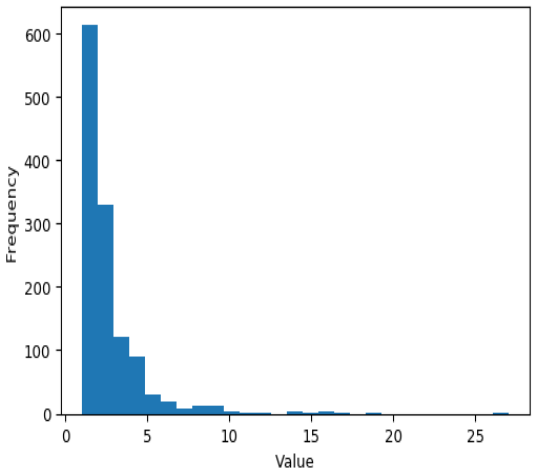
In Teil 2 wird LDA auf den Neuroblastom Datensatz angewendet. Der zu untersuchende Datensatz wurde mit dem zweiten Verfahren aus Abschnitt 3.3 bestimmt.

Latent Dirichlet Allocation

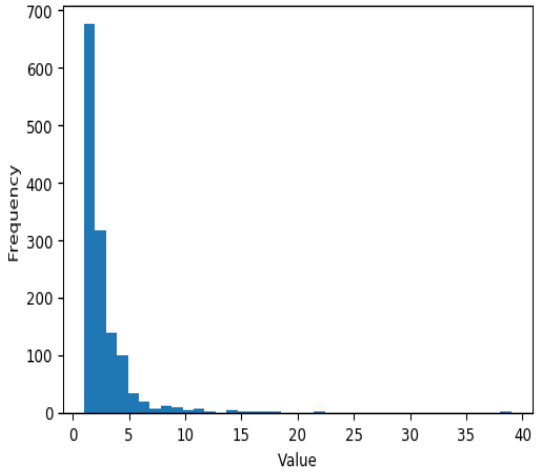
In Abbildung 4.16 sind die Unterschiede zwischen den Gewebearten, die in Abschnitt 3.3 genannt wurden, dargestellt. In der Abbildung 4.17 sind die Ergebnisse der Untersuchung des Datensatzes mit LDA zu finden.



(a) Tumorgewebe/Normalgewebe

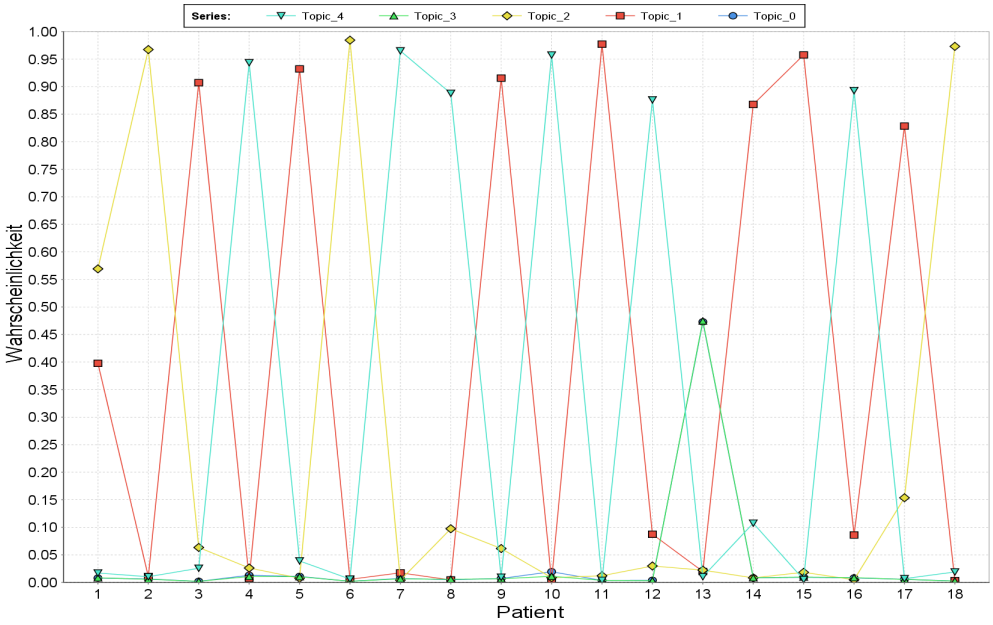


(b) Rezidivgewebe/Tumorgewebe



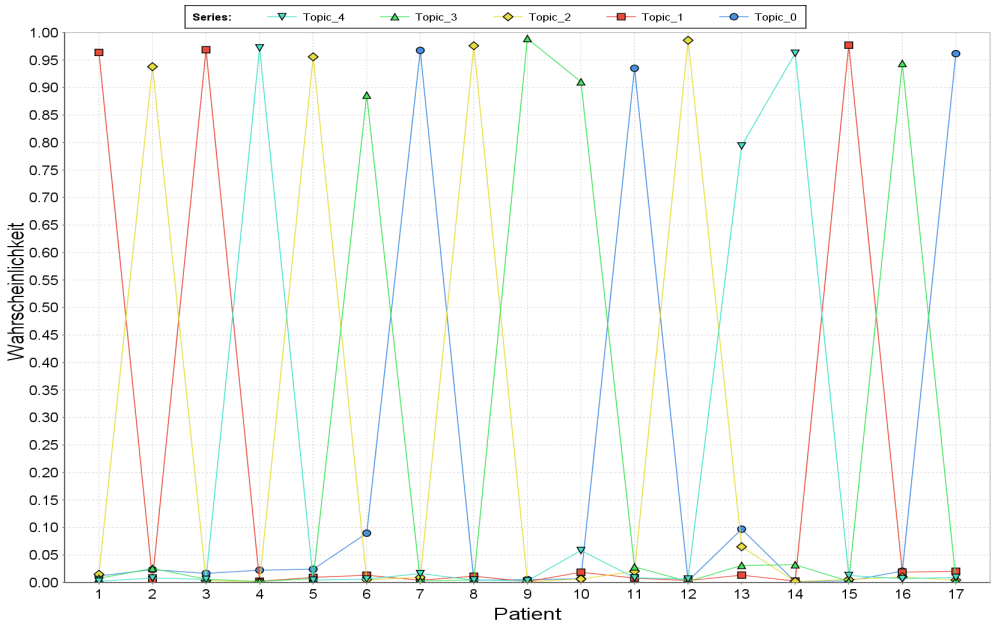
(c) Rezidivgewebe/Normalgewebe

Abbildung 4.16: Die X-Achse bezeichnet die Summe aller Mutationshäufigkeiten in dem jeweiligen Datensatz. Die Y-Achse bezeichnet ihre Häufigkeit.



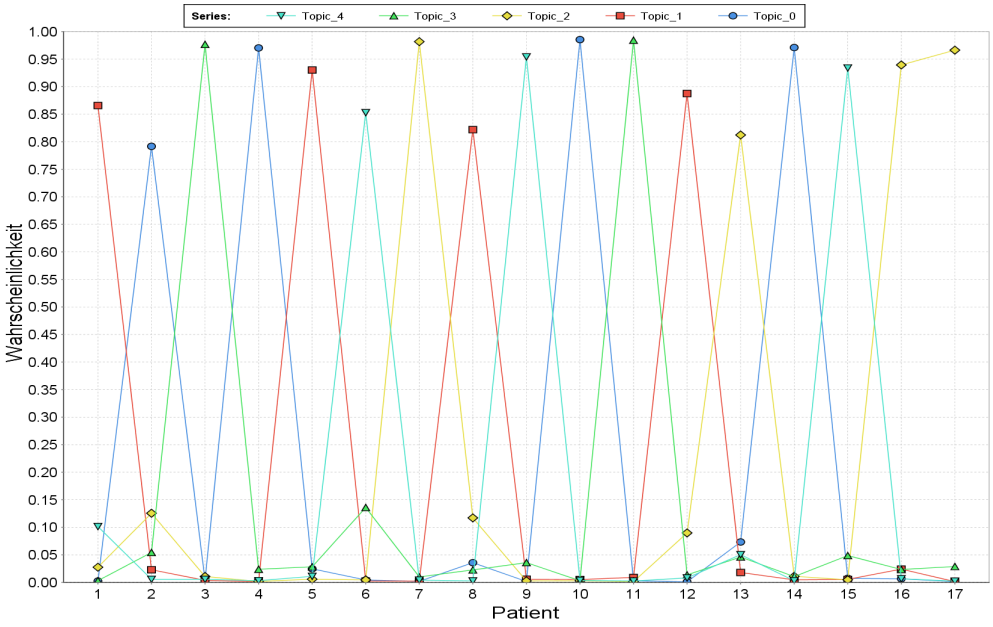
(a) Die Themenverteilung nach der Untersuchung des Datensatzes mit LDA. Es sind nur die Gene im Datensatz enthalten, die im Tumorgewebe mutiert sind, aber nicht im Normalgewebe.

(b) Die Wörterverteilung nach der Untersuchung des Datensatzes, welcher nur die Gene enthält, die im Rezidivgewebe mutiert sind, aber nicht im Tumorgewebe.



(c) Die Themenverteilung nach der Untersuchung des Datensatzes mit LDA. Es sind nur die Gene im Datensatz enthalten, die im Rezidivgewebe mutiert sind, aber nicht im Tumorgewebe.

(d) Die Wörterverteilung nach der Untersuchung des Datensatzes, welcher nur die Gene enthält, die im Rezidivgewebe mutiert sind, aber nicht im Tumorgewebe.



(e) Die Themenverteilung nach der Untersuchung des Datensatzes LDA. Es sind nur die Gene im Datensatz enthalten, die im Rezidivgewebe mutiert sind, aber nicht im Normalgewebe.

(f) Die Wörterverteilung nach der Untersuchung des Datensatzes, welcher nur die Gene enthält, die im Rezidivgewebe mutiert sind, aber nicht im Tumorgewebe.

Abbildung 4.17: Die von LDA ausgegebene Themen- und Wörterverteilung nach der Untersuchung des Datensatzes. In den Abbildungen 4.17a, 4.17c und 4.17e sind die Themenverteilungen abgebildet. Die Y-Achse beschreibt die Wahrscheinlichkeit, dass das jeweilige Thema in einem Dokument vorkommt. Die X-Achse ist mit den Patientennummern gekennzeichnet. Die Abbildungen 4.17b, 4.17d und 4.17f bilden die Wörterverteilungen der Themen in den 3 unterschiedlichen Datensätzen als Wordcloud ab. Aus welchem Thema die Wörter entnommen worden sind, ist an der Farbe zu erkennen. Hat ein Wort in den Wörterverteilungen die gleiche Farbe, wie die in der Themenverteilung, dann gehört das Wort zu diesem Thema, Je größer das Wort ist, desto wahrscheinlicher tritt das Wort in dem Thema auf. Es werden die 10 wahrscheinlichsten Wörter pro Thema dargestellt.

4.3.3 Teil 3

In Teil 3 wird der Datensatz, der 98 Patienten beinhaltet, die an unterschiedlichen Krebsarten erkrankt sind, untersucht. Für die Anwendung von LDA wurde der Datensatz mit dem zweiten Verfahren aus Abschnitt 3.3 bestimmt. Für sLDA musste der Datensatz mit dem ersten Verfahren aus Abschnitt 3.3 bestimmt werden.

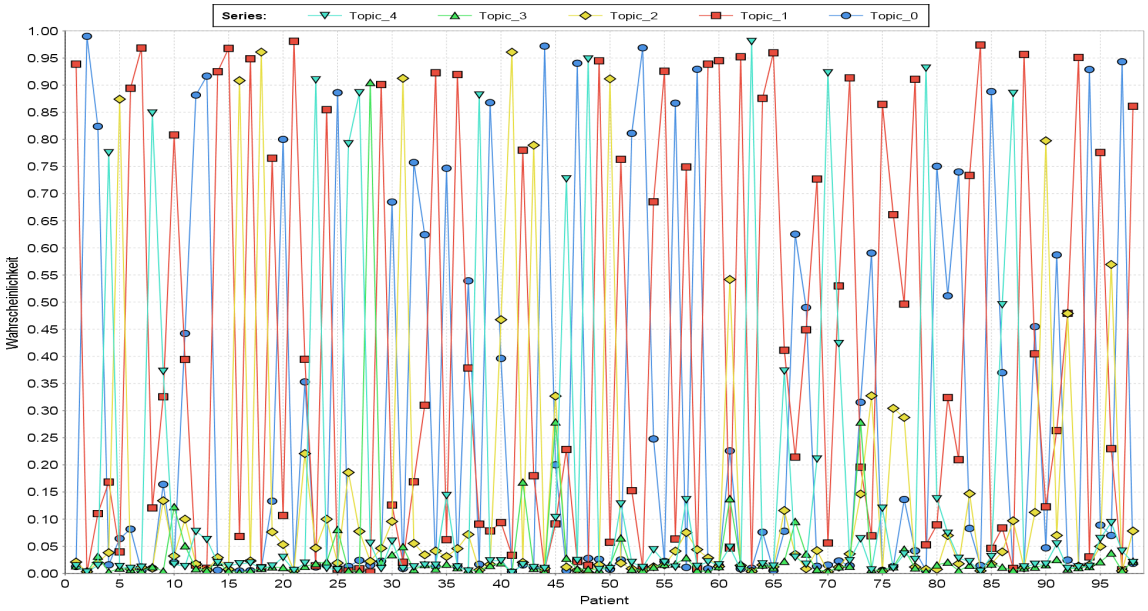
Latent Dirichlet Allocation

Der Datensatz wurde mit LDA untersucht und die sich ergebene Themenverteilung in Abbildung 4.18a und die Wörterverteilung in 4.18b abgebildet.

Supervised Latent Dirichlet Allocation

In Abbildung 4.19 ist die von sLDA ausgegebene Themenverteilung über alle Stichproben zu finden. In Abbildung 4.20 ist die dazugehörige Wörterverteilung als Wordcloud abgebildet.

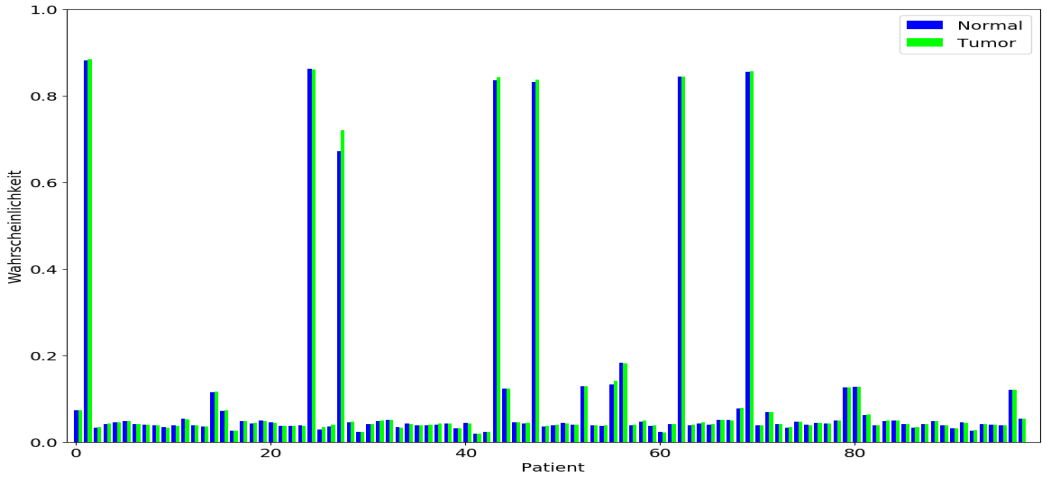
Die durchschnittliche Genauigkeit der Vorhersage der Labels beträgt ungefähr 50%. Die Genauigkeit ist unabhängig davon, welcher Wert für die Themenanzahl und für den Parametervektor α gewählt wird.



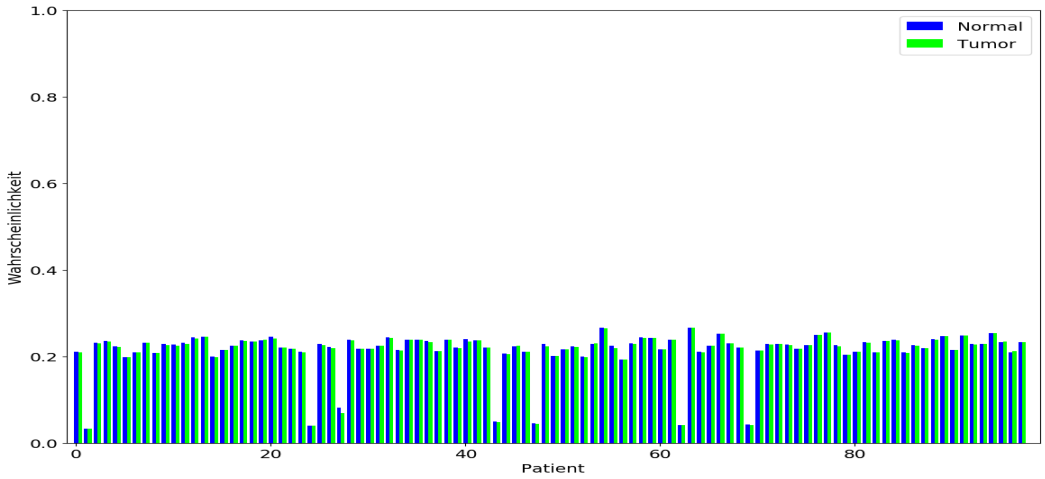
(a) Die Abbildung stellt die Themenverteilung der 5 Themen über die 98 Patienten dar, die sich bei der Untersuchung des Datensatzes mit LDA ergeben hat. Der Datensatz enthält nur die Gene, die im Tumorgewebe mutiert sind, aber nicht im Normalgewebe.

(b) Die Abbildung stellt die Wörterverteilung nach der Untersuchung des Datensatzes mit LDA.

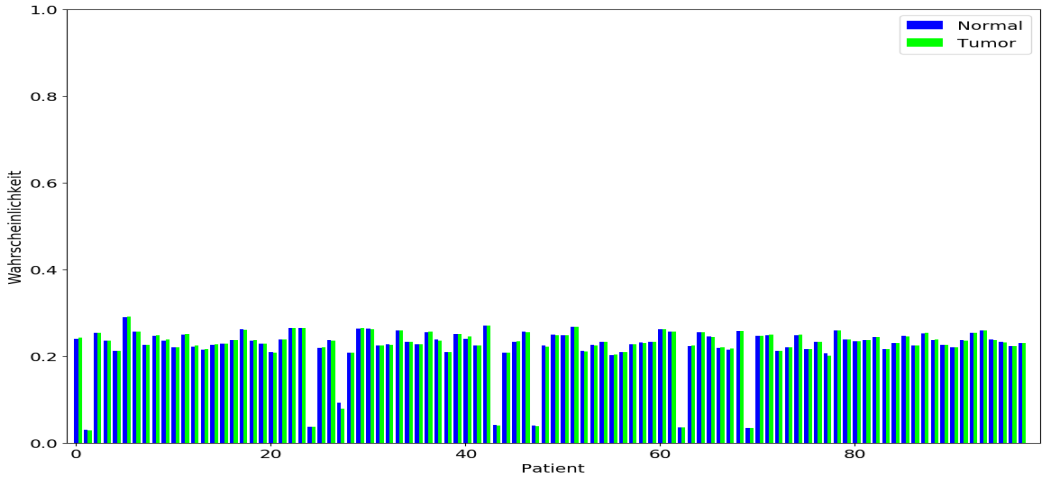
Abbildung 4.18: Die von LDA ausgegebene Themenverteilung und Wörterverteilung nach der Untersuchung des Datensatzes. In Abbildung 4.18a ist die Themenverteilung abgebildet. Die X-Achse bezeichnet Die Patientennummer und die Y-Achse die Wahrscheinlichkeit des Auftretens der jeweiligen Themen. In Abbildung 4.18b ist die Wörterverteilung als Wordcloud abgebildet. Aus welchem Thema die Wörter entnommen worden sind, ist an der Farbe zu erkennen. Hat ein Wort in den Wörterverteilung die gleiche Farbe wie die in der Themenverteilung, dann gehört das Wort zu diesem Thema, Je größer das Wort ist, desto wahrscheinlicher tritt das Wort in dem Thema auf. Es werden die 10 wahrscheinlichsten Wörter pro Thema dargestellt.



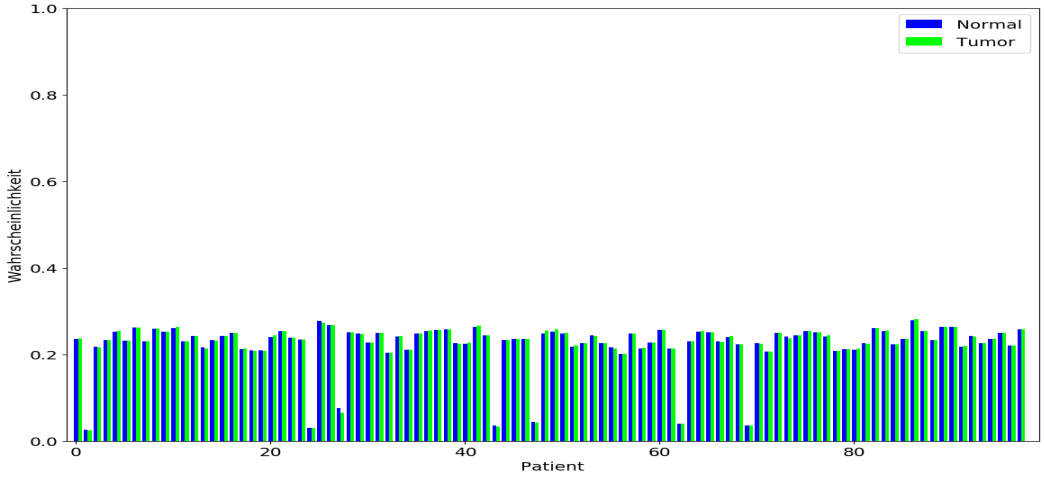
(a) Wahrscheinlichkeitsverteilung von *Thema 1* über alle 196 Stichproben



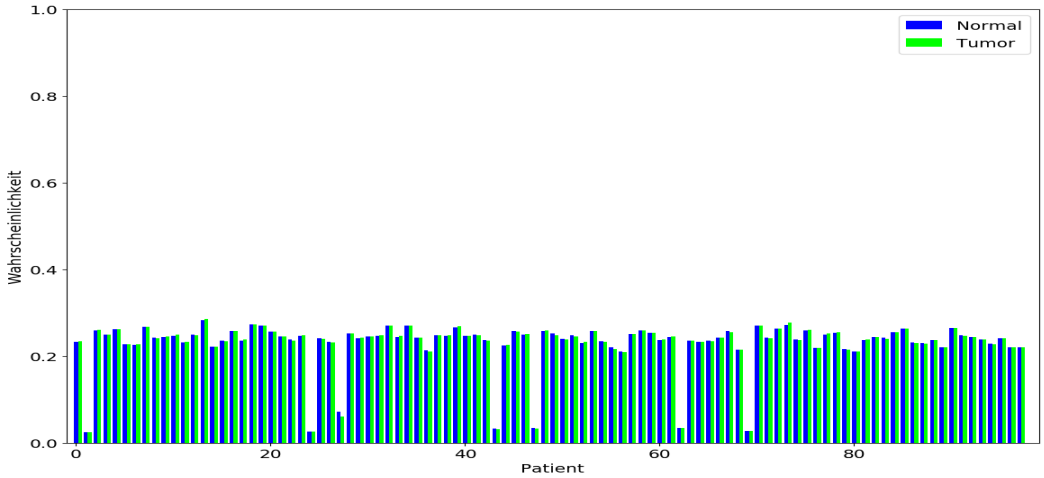
(b) Wahrscheinlichkeitsverteilung von *Thema 2* über alle 196 Stichproben



(c) Wahrscheinlichkeitsverteilung von *Thema 3* über alle 196 Stichproben



(d) Wahrscheinlichkeitsverteilung von *Thema 4* über alle 196 Stichproben



(e) Wahrscheinlichkeitsverteilung von *Thema 5* über alle 196 Stichproben

Abbildung 4.19: Die Y-Achse bezeichnet die Wahrscheinlichkeit für das Auftreten des jeweiligen Themas in einer Stichprobe. Die X-Achse kennzeichnet die Patientennummer.

Abbildung 4.20: Die Wordcloud zeigt die von sLDA ausgegebenen Wörterverteilung. Pro Thema sind die 10 wahrscheinlichsten Wörter dargestellt. Je größer die Wahrscheinlichkeit ist, desto größer ist das Wort.

Kapitel 5

Diskussion

In dieser Arbeit wurden 2 unterschiedliche Datensätze, die mit 2 verschiedenen Verfahren präpariert worden sind, mit LDA, sLDA und lLDA untersucht. Da bei den drei Modellen vorher festgelegt werden musste, wie viele Themen im Datensatz erwartet werden, wurde damit begonnen eine Abschätzung für eine Themenanzahl zu finden, die zu guten Ergebnissen führen soll. Dafür wurde LDA auf dem Neuroblastomdatensatz, welcher mit dem ersten Verfahren aus Abschnitt 3.3 präpariert wurde angewendet und mit der Formel 4.2 die Abbildungen 4.1 erstellt, um herauszufinden, wie groß die Anteile der Themen in den Dokumenten sind. Es stellt sich heraus, dass bei der Wahl von 5 Themen, die Themennummern 4 in den Abbildungen 4.1a, 4.1b und 4.1c und die Themennummer 3 in Abbildung 4.1c der Anteil plötzlich drastisch weniger wird. Dies ist ein Indiz dafür, dass die Themen beginnen spezifischer zu werden. Da es nur 18 Stichproben und dementsprechend 18 Dokumente gibt, besteht die Gefahr, dass bei der Wahl von einer großen Themenanzahl, die Themen nicht latente Subtypen sondern einzelne Patienten beschreiben. Angesichts der Tatsache, dass die summierten Wahrscheinlichkeiten der letzten Themen bei der Untersuchung des Datensatzes mit LDA und der erwarteten Themenanzahl von 10 und 15 sehr gering wurden, ist beschlossen worden, dass die Festlegung auf 5 Themen zu den besten Ergebnissen führen wird.

Bei der Betrachtung der Ergebnisse der Untersuchung desselben Datensatzes mit LDA (Abbildung 4.2), stellt sich heraus, dass in der Themenverteilung des Normalgewebes in 4.2a das 1. Thema und in der Themenverteilung des Tumorgewebes in 4.2c das 4. Thema sehr hohe Wahrscheinlichkeiten für die Patienten mit den Nummern 6 und 13 haben, aber nur sehr niedrige für alle anderen Patienten. Bei der Themenverteilung des Rezidivgewebes 4.2e ist kein Thema zu erkennen, welches ungefähr die gleichen Eigenschaften besitzt. Nach genauerer Betrachtung des Datensatzes hat sich herausgestellt, dass die Patienten mit den Nummern 6 und 13 signifikant mehr Genmutationen im Normal- und Tumorgewebe besitzen als die anderen Patienten. Im Rezidivgewebe sind bei den genannten Patienten ungefähr genau so viele Genmutationen aufgetreten, wie bei den anderen Patienten. Dies

könnte eine mögliche Erklärung für das Verhalten der beiden genannten Themen sein. Des Weiteren fällt auf, dass die Themenverteilungen des Normal- und Tumorgewebes sich sehr ähnlich sehen, wobei sich die Themenverteilung des Rezidivgewebes stark von diesen Unterscheidet. Dies weist darauf hin, dass sich die Genmutationen im Normal- und Tumorgewebe nur so wenig unterscheiden, dass diese Unterschiede nicht signifikant genug sind, um sie mit LDA erkennen zu können. Im Rezidivgewebe ändert sich die Themenverteilung erheblich, da jedes Thema ungefähr gleichwahrscheinlich in den jeweiligen Patienten auftritt. Eine Erklärung dafür ist vermutlich, dass die Patienten therapiert worden sind und der Neuroblastom demzufolge homogener geworden ist. Bei der Betrachtung der Wörterverteilungen in den Abbildungen 4.2b, 4.2d und 4.2f fällt auf, dass viele Gene (z.B *TTN*, *PDE4DIP* und die Gene, die mit *HLA* beginnen) in allen drei Wordclouds mit großer Wahrscheinlichkeit auftauchen. Diese Gene sind besonders groß und mutierten daher sehr häufig, jedoch ist es unwahrscheinlich, dass diese etwas mit dem Neuroblastom zu tun haben.

So wie es in den Themenverteilungen des Normal- und Tumorgewebes zu sehen ist, gibt es oft zwei Themen aus zwei unterschiedlichen Gewebearten, die bei jedem Patienten mit einer ungefähr gleichen Wahrscheinlichkeit auftreten. Von Interesse ist nun, wie sich die Gene in diesen zueinander ähnlichen Themen unterscheiden, da dies die Unterschiede zwischen zwei Gewebearten innerhalb der Themen verdeutlicht. In den Abbildungen 4.3 bis 4.4 ist jedoch zu erkennen, dass nicht jedem Thema aus einem Gewebe ein Thema aus einem anderen Gewebe eindeutig zugewiesen werden kann, da einige Themen sehr ähnlich zu mehreren Themen sind. Bei der Untersuchung der Wahrscheinlichkeit der Gene, die sich in einem Themenpaar unterscheiden (Abbildung 4.6 - 4.8), stellt sich heraus, dass sich einige Themenpaare trotzdem stark unterscheiden. Als Referenz für ein Themenpaar, welches sich sehr ähnlich ist, kann das 1. Thema aus dem Normalgewebe und das 4. Thema aus dem Tumorgewebe genommen werden (Abbildung 4.6a), da diese sich in den Themenverteilung der jeweiligen Gewebearten (Abbildung 4.2a & 4.2c) offensichtlich sehr ähnlich sind und der Tabelleneintrag in Abbildung 4.3 an dieser Stelle besonders niedrig im Verhältnis zu den anderen Einträgen ist. Das die Themenpaare sich häufig noch sehr stark in den Wahrscheinlichkeit der Gene unterscheiden, liegt möglicherweise daran, dass die von LDA ausgegebenen Themen nicht die gleichen Subtypen in den Gewebearten gefunden hat, sodass sich die Themen in den Wahrscheinlichkeiten der Gene sehr stark unterscheiden. Nur in dem oben genannten Fall, war es mit LDA möglich, zwei sich sehr ähnliche Themen in 2 unterschiedlichen Gewebearten zu finden.

Bei der Untersuchung der Ergebnisse, die sich bei der Anwendung von sLDA auf denselben Datensatz ergeben haben, stellt sich heraus, dass sich die Gewebearten in den Wahrscheinlichkeiten des Auftretens eines Themas nur insignifikant unterscheiden (Abbildung 4.9). Die Ausnahmen sind wieder die Patienten mit der Patientennummer 6 und 13, was wahrscheinlich wieder daran liegt, dass bei den genannten Patienten viel mehr Genmutationen im Normal- und Tumorgewebe aufgetreten sind als bei den anderen Patienten.

Da sich die Wahrscheinlichkeit des Auftretens eines Themas innerhalb der Stichproben eines Patienten bei jedem Patienten nur minimal unterscheidet, ist anzunehmen, dass die Untersuchung der Daten mit sLDA keine großen Erfolge bringt. Dies liegt eventuell daran, dass sich die Stichproben aus den verschiedenen Gewebearten eines Patienten in den Genmutationen nur minimal unterscheiden. In der Wörterverteilung in Abbildung 4.10 ist hauptsächlich die Häufigkeit der sehr großen Gene, die sehr häufig mutieren, zu erkennen. Da viele Gene in der Wordcloud sehr viel häufiger als ein Mal dargestellt sind und somit das Gen in vielen Themen sehr wahrscheinlich vorkommt, wird die Annahme, dass sLDA keine Erfolge auf dem Datensatz erbringt, unterstrichen. Das Ergebnis der Vorhersage der Labels ist mit einer Genauigkeit von 33% sehr schlecht, da es insgesamt 3 unterschiedliche Labels gibt. Dementsprechend ist die Vorhersage mit sLDA nicht besser als die zufällige Wahl eines Labels.

Bei der Betrachtung der Ergebnisse der Untersuchung des selben Datensatzes mit lLDA (Abbildung 4.11 - 4.15) zeigt sich, dass die Stichproben aus den 3 verschiedenen Gewebearten tatsächlich sehr ähnlich sind, da die Wahrscheinlichkeit des Auftretens des Themas *Common* bei den meisten Stichproben bei über 0.99 befindet. Die Ausnahmen sind die Stichproben aus dem Normal- und Tumorgewebe der Patienten mit der Patientennummer 6 und 13, und die Stichproben aus dem Rezidivgewebe des Patienten mit der Patientennummer 3. Bei Patient 6 und 13 liegt es wahrscheinlich wieder an der Häufigkeit der Genmutationen in dem Normal- und Tumorgewebe. Bei dem Patienten 3 liegt es möglicherweise daran, dass von diesem Patienten 4 sich sehr stark ähnelnde Stichproben aus dem Rezidivgewebe entnommen worden sind. Da es für diese 4 Stichproben aus dem Rezidivgewebe jeweils 3 weiter sehr ähnliche Stichproben aus dem Rezidivgewebe gibt, ist der Anteil des Themas *Relaps* bei diesen Stichproben besonders hoch. Aus der Wordcloud des Themas *Common* in Abbildung 4.12 ist zu entnehmen, dass mit lLDA erfolgreich die Gene, die sehr groß sind und somit sehr häufig in jedem Patienten mutieren, nur dem Thema *Common* zugeordnet worden sind, da die meisten dieser Gene sehr dunkel dargestellt sind. Bei der Betrachtung der Wordclouds der Themen *Normal* und *Tumor* in den Abbildungen 4.13 und 4.14 ist zu erkennen, das diese sich sehr ähnlich sind und dementsprechend nur wenige Gene dunkel dargestellt sind. Dies weist wieder darauf hin, dass sich die Gewebearten in den Genmutationen nur wenig unterscheiden. Bei der Wordcloud des Themas *Relaps* fällt auf, dass im Verhältnis zu den anderen Wordclouds viele Gene dabei sind, die sehr dunkel sind und somit nur in dieser Wordcloud auftauchen. Diese Gene könnten besonders interessant sein, da sie mit größerer Wahrscheinlichkeit im Rezidivgewebe mutieren und somit womöglich eine Ursache für das Entstehen des Rezidivgewebes sein können.

Da sich bei den Untersuchungen mit den 3 Methoden herausgestellt hat, dass sich die Gewebearten in den Genmutationen nur geringfügig unterscheiden, wurden die Daten mit einem anderen Verfahren präpariert, sodass der zu untersuchende Datensatz nur noch die Genmutationen enthält, die in einem Gewebe aufgetreten sind, aber in einem anderen nicht.

Dadurch werden die Gemeinsamkeiten zweier Gewebearten aus dem Datensatz entfernt und somit die Unterschiedlichkeit der Daten erhöht. Um die geringe Differenz der Gewebearten deutlicher zu machen, ist diese in der Abbildungen 4.16 dargestellt. Es ist zu erkennen, dass mit Abstand die meisten Gene nur maximal einmal mehr in einem Gewebe mutiert sind als im anderen Gewebe. Einige Gene sind zwar häufiger mutiert, jedoch nimmt die Anzahl drastisch ab. Diese Unterschiede scheinen zu gering zu sein, um besonders gute Ergebnisse mit den 3 Methoden zu erzielen. Besonders gut ist dies bei dem sLDA Modell aufgefallen.

Weil der Datensatz nun nur die Gene der Differenz zweier Gewebe beinhaltet, unterscheiden sich die Ergebnisse der Untersuchung mit LDA (Abbildung 4.17) sehr stark zu den Ergebnissen der Untersuchung des anderen Datensatzes. Besonders interessant ist die Themenverteilung der Gewebearten, da es den Anschein erweckt, dass die Patienten mit LDA in Klassen aufgeteilt werden können, wobei die Themen die Klassen entsprechen. Die Einteilung der Patienten in geeignete Klassen kann eventuell für die Behandlungen der Patienten abhängig von den Klassen, hilfreich sein. Bei genauerer Betrachtung stellt sich jedoch heraus, dass es kein Thema in den Themenverteilungen gibt, das dieselben Patienten in 2 unterschiedlichen Themenverteilungen beschreibt. Wird beispielsweise das Ergebnis der Themenverteilung in Abbildung 4.17a betrachtet, so behandeln die Patienten 1, 3 und 15 am wahrscheinlichsten das Thema 1. Allerdings gibt es kein Thema in den anderen Themenverteilungen, welches genau von diesen Patienten am wahrscheinlichsten behandelt wird. Ein besonders gutes Ergebnis wäre es gewesen, falls diese Eigenschaft erfüllt wäre, da in dem Fall Vorhersagen getroffen werden könnten. In den Wörterverteilungen der Ergebnisse ist zu beobachten, dass die Gene, die in den Wordclouds der vorherigen Ergebnisse sehr dominant waren, es in diesen Wordclouds nicht mehr sind. Dies ist dadurch zu erklären, dass diese Gene, falls sie in einem Gewebe mutiert sind, auch in den anderen Geweben mutieren. Dementsprechend sind diese Gene uninteressant und wurden nicht in den zu untersuchenden Datensatz hinzugefügt.

Da der Neuroblastomdatensatz aus nur 18 Patienten besteht, bestand die Gefahr, dass die Modelle keine guten Ergebnisse auf einem so kleinen Datensatz liefern können. Aus diesem Grund wurde in Teil 3 der Datensatz untersucht, welcher 98 Patienten mit jeweils 2 Stichproben beinhaltet. Es stellt sich heraus, dass die Ergebnisse (Abbildung 4.18) der Analyse des größeren Datensatzes mit LDA ähnlich zu den Ergebnissen des Neuroblastomdatensatzes sind. Die Themenverteilung in Abbildung 4.18a zeigt wieder, dass die meisten Patienten ein Thema eindeutig wahrscheinlicher behandeln als alle anderen Themen. Somit erscheint es so, als ob diese Patienten in Klassen abhängig von den Themen eingeteilt werden können. In der Wörterverteilung in Abbildung 4.18b ist zu beobachten, dass die Themen sich stark in den Genen unterscheiden. Einem Biologen ist es vielleicht möglich, aus diesen Genen herauszufinden, welcher Krebs bei dem Patienten, der dieses Thema am wahrscheinlichsten behandelt, entstehen wird.

Anschließen wurde sLDA auf dem großen Datensatz angewandt, um herauszufinden, ob sich die Ergebnisse bei der Betrachtung von einem größeren Datensatz verbessern. Da nur Stichproben aus dem Normal- und Tumorgewebe entnommen worden sind, ist es nicht möglich die Differenz der beiden zu einen Datensatz zusammenzufassen und mit sLDA zu untersuchen, da dieser Datensatz mit nur einem Label gekennzeichnet werden könnte. Dementsprechend ist sLDA auf dem Datensatz angewendet worden, der alle Genmutationen in beiden Gewebearten beinhaltet. Die sich ergebene Themenverteilung zeigen besonders interessante Verteilungen für die Patienten mit den Patientennummern 1, 24, 27, 43, 47, 62 und 69. Thema 1 scheint die Beschreibung dieser Menge der Patienten zu sein. Außerdem ist zu beobachten, dass die Themenverteilung eines Patienten sich nur insignifikant in den Stichproben aus dem Normal- und Tumorgewebe unterscheidet. Möglicherweise sind die geringen Unterschiede zwischen den Geweben die Ursache dafür. Aus der Wordcloud in Abbildung 4.18b ist zu entnehmen, dass die meisten Themen denselben Genen eine hohe Wahrscheinlichkeit zuweisen, was die Gleichverteilung der Themen 2-4 über die Patienten erklärt. In Thema 1 sind jedoch 5 der 10 wahrscheinlichsten Gene einzigartig für das Thema. Die Gene *GSN*, *DYSF*, *CACNA1G*, *CACNA1C* und *ZMYND8* scheinen einen große Rolle in den Patienten 1, 24, 27, 43, 47, 62 und 69 einzunehmen. Eventuell ist durch die Interpretation dieser Gene eine Behandlung der genannten Patienten möglich. Die Vorhersage der Labels ist mit einer Genauigkeit von 50% sehr schlecht, da es insgesamt nur 2 unterschiedliche Labels gibt. Da bei beiden Datensätzen die Vorhersage nicht erfolgreich ist, ist anzunehmen dass es nicht an der Größe des Datensatzes, sondern an der geringen Differenz der Gewebe liegt.

Kapitel 6

Fazit

6.1 Zusammenfassung

In dieser Arbeit wurden genetische Daten mit den Textmodellen *Latent Dirichlet Allocation*, *Supervised Latent Dirichlet Allocation* und *Labeled Latent Dirichlet Allocation* untersucht. Die zu untersuchenden Datensätze wurden mit zwei Verfahren präpariert. Beim ersten Verfahren wurden alle Genmutationen und beim zweiten Verfahren nur die Differenz zweier Gewebearten in den Datensatz hinzugefügt. Es wurde die geeignete Themenanzahl für die Untersuchung der Datensätze mit den Modellen bestimmt und die Modelle auf die genetischen Daten angewandt. Außerdem wurde versucht, die Labels der Stichproben in beiden Datensätzen vorherzusagen.

Die Ergebnisse haben gezeigt, dass die Wahl von 5 Themen für die Untersuchung der genetischen Daten zu besseren Ergebnissen führt, als 10 oder 15 Themen. Die Anwendung der Modelle auf den Datensatz, der mit dem ersten Verfahren bestimmt wurde, zeigte, dass die Gewebe sich sehr ähnlich bezüglich der Genmutationen sind. Dies ist besonders gut bei den Themenverteilungen in den Ergebnissen von der Anwendung des sLDA und lLDA Modells aufgefallen. Bei der Untersuchung des Datensatzes, welcher mit dem zweiten Verfahren präpariert worden ist, sind jedoch plötzlich klare Strukturen zu erkennen. Die Themen beschreiben fast eindeutig Patienten, sodass diese, je nachdem welche Themen sie am wahrscheinlichsten behandeln, in Klassen aufgeteilt werden können. Vorhersagen über die Labels der Stichproben sind bei dem kleineren und größeren Datensatz erfolglos gewesen.

6.2 Fazit

Es hat sich ergeben, dass das Untersuchen vom Datensatz, welcher mit dem ersten Verfahren präpariert wurden, zu nur unbefriedigenden Ergebnissen führte. Da sich die Gewebearten eines Patienten nur wenig unterscheiden, gelingt es mit den Modellen nicht so gut,

diese voneinander zu unterscheiden. Hingegen sind die Ergebnisse bei der Untersuchung des Datensatzes, der mit dem zweiten Verfahren bestimmt worden ist, zufriedenstellend. Es sind klare Strukturen bei den Patienten zu erkennen, woraus gefolgert werden kann, dass die Modelle auf ein gut präparierten genetischen Datensatz zu guten Ergebnissen führen kann.

6.3 Ausblick

Um die Ergebnisse der Anwendung von sLDA auf den Datensatz zu verbessern, müssen die Unterschiede zwischen den Gewebearten größer werden. Dies kann durch die Vorfilterung der Genmutationen, von denen bekannt ist, dass diese bei gesunden Menschen ebenfalls auftreten, im Datensatz passieren. Der Exome Aggregation Consortuim (ExAC) [8] Datensatz enthält genau diese Informationen. Das gleiche kann gemacht werden, um die Ergebnisse der Analyse mit lLDA zu verbessern.

Die Datensätze können außerdem mit einer weiteren Methode, namens *Discriminative Latent Dirichlet Allocation* (discLDA) [7], untersucht werden. DiscLDA ist eine überwachtes Modell, welches sowie sLDA, voraussetzt, dass die Dokumente jeweils genau einem Label zugeordnet sind. Der Unterschied ist jedoch, dass bei DiscLDA versucht wird, die Themenverteilungen der Dokumente mit gleichem Label ungefähr gleich zu wählen, sodass bestimmte Themenverteilungen mit bestimmten Labels assoziiert werden können. Dadurch können Unterschiede in den Labels besser durch die Themenverteilungen erkannt werden als bei sLDA.

Anhang A

Weitere Informationen

Abbildungsverzeichnis

Literaturverzeichnis

[1] Blei, David M.: *Probabilistic Topic Models.* Communications of the ACM, 2012.

[2] Blei, David M. und Jon D. McAuliffe: *Supervised Topic Models.* 2010.

[3] Blei, David M., Andrew Y.Ng und Micheal I.Jordan: *Latent Dirichlet Allocation.* 2003.

[4] Danecek, Petr, Adam Auton, Goncalo Abecasis, Cornelis A. Albers1, Eric Banks, Mark A. DePristo, Robert E. Handsaker, Gerton Lunter, Gabor T. Marth, Stephen T. Sherry, Gilean McVean, Richard Durbin1, und 1000 Genomes Project Analysis Group: *The variant call format and VCFtools.* 2011.

[5] Evans, Audrey E., Giulio J.D'Angio und Judson Randolph: *A proposed staging for children with Neuroblastoma.* 1971.

[6] Hess, Sibylle und Katharina Morik: *C-SALT: Mining Class-Specic ALTerations in Boolean Matrix Factorization.* 2017.

[7] Lacoste-Julien, Simon, Fei Sha und Michael I. Jordan: *DiscLDA: Discriminative Learning for Dimensionality Reduction and Classification.* 2008.

[8] MacArthur, Daniel G., Monkol Lek, Konrad J. Karczewski1, Eric V. Minikel, Kaitlin E. Samocha, Eric Banks, Timothy Fennell, Anne H. O'Donnell-Luria, James S. Ware, Andrew J. Hill, Beryl B. Cummings, Taru Tukiainen1 und Daniel P. Birnbaum et al.: *Analysis of protein-coding genetic variation in 60,706 humans.* Nature, 2016.

[9] Ramage, Daniel, David Hall, Ramesh Nallapati und Christopher D. Manning: *Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora.* 2009.

[10] Schramm, Alexander, Johannes Köster, Yassen Assenov, Kristina Althoff1, Martin Peifer, Ellen Mahlow, Andrea Odersky, Daniela Beisser,

Corinna Ernst, Anton G Henssen1, Harald Stephan, Christopher Schröder, Lukas Heukamp, Anne Engesser, Yvonne Kahlert, Jessica Theissen, Barbara Hero, Frederik Roels, Janine Altmüller, Peter Nürnberg, Kathy Astrahantseff, Christian Gloeckner, Katleen De Preter, Christoph Plass, Sangkyun Lee, Holger N Lode, Kai-Oliver Henrich, Moritz Gartlgruber, Frank Speleman, Peter Schmezer, Frank Westermann, Sven Rahmann, Matthias Fischer, Angelika Eggert und Johannes H Schulte1: *Mutational dynamics between primary and relapse neuroblastomas.* 2015.

[11] Steyvers, Mark und Tom Griffiths: *Probabilistic Topic Models.*

[12] Westerveld, Thijs, Arjen de Vries, und Franciska de Jong: *Generative Probabilistic Models.* 2007.

[13] Zhao1, Weizhong, James J. Chen, Roger Perkins, Yuping Wang, Zhichao Liu1, Huixiao Hong, Weida Tong und Wen Zou1: *A novel procedure on next generation sequencing data analysis using text mining algorithm.* BMC Bioinformatics, 2016.