technische universität
dortmund

# Statistische Analyse von Modellen für die Krankheitsprogression

**Dissertation**

zur Erlangung des akademischen Grades
Doktor der Naturwissenschaften
an der Fakultät Statistik
der Technischen Universität Dortmund

vorgelegt von

**Katrin Hainke**

Dortmund, September 2017

Gutachter: Prof. Dr. Roland Fried
Prof. Dr. Jörg Rahnenführer

Tag der mündlichen Prüfung: 5. Dezember 2017

# Inhaltsverzeichnis

# Kapitel 1

# Einleitung

Die grundlegende Kenntnis über den Verlauf einer Krankheit und deren charakteristische Eigenschaften kann viel zu einer erfolgreichen Behandlung bzw. der medizinischen Versorgung eines Patienten beitragen. Ein großes Ziel ist daher, eine Krankheit so gut wie möglich zu verstehen. Dies ist sowohl für eine genauere und frühere Diagnose des Krankheitsstadiums als auch für geeignetere individuelle Therapieentscheidungen hilfreich.

Die Krankheit Krebs lässt sich z.B. durch die Anhäufung von Mutationen charakterisieren, die sich auf das Tumorwachstum, den klinischen Verlauf, die Immunevasion und das Entstehen von Medikamentenresistenzen auswirken (Beerenwinkel et al., 2015). Durch aufgetretene Mutationen kann das Zellwachstum außer Kontrolle geraten, so dass Wachstum, Teilung und Zerstörung von Zellen nicht mehr optimal aufeinander abgestimmt sind. Obwohl es sich bei Krebs um eine sehr heterogene Krankheit handelt, gibt es grundlegende funktionale Prinzipien, die oft als 'hallmarks of cancer' bezeichnet werden. Es gibt verschiedene Ansätze, die Entstehung von Krebs mit mathematischen und statistischen Methoden zu modellieren. Dazu gehören populationsdynamische Modelle, phylogenetische Methoden und wahrscheinlichkeitstheoretische grafische Modelle (Beerenwinkel et al., 2015).

Für einfache populationsdynamische Modelle wird oft angenommen, dass keine Interaktion zwischen Tumorzellen stattfindet und die Populationsstruktur unberücksichtigt bleibt. Unter diesen Annahmen können durch mathematische Modelle interessierende Merkmale wie die genetische Diversität innerhalb des Tumors, die Wahrscheinlichkeit, mit der sich eine Mutation durchsetzt, und das Alter des Tumors berechnet werden. Phylogenetische Methoden berechnen Bäume, die die evolutionäre Historie des Tumors darstellen und anhand derer verschiedene Hypothesen zur Tumorentstehung getestet werden können. Den grafischen Progressionsmodellen liegt die Annahme zugrunde, dass die genetischen Ereignisse einer Krebserkrankung nicht unabhängig voneinander auftreten. Ihr Ziel ist es, diese Abhängigkeiten zu schätzen.

Für alle drei genannten Ansätze werden verschiedene Methoden in Beerenwinkel et al. (2015) benannt. In dieser Arbeit wird der Fokus jedoch auf die grafischen Krankheitsprogressionsmodelle gelegt. Es ist von großer Bedeutung, Informationen über die einzelnen Schritte einer Krankheit zu gewinnen, um sie so besser zu verstehen und gezielter behandeln zu können.

Allen hier betrachteten Modellen zur Beschreibung der Krankheitsprogression liegen genetische Ereignisse zugrunde. Über diese sollen Aussagen bezüglich Abhängigkeit, Unabhängigkeit und einer möglichen Reihenfolge getroffen werden. Ein genetisches Ereignis lässt sich dabei durch eine binäre Zufallsvariable darstellen, die angibt, ob das Ereignis eingetreten ist oder nicht. Konkret bedeutet das Eintreten eines genetischen Ereignisses, dass eine Mutation stattgefunden hat. Dabei ist es je nach zugrunde liegendem Datensatz ganz unterschiedlich, ob sich diese Mutation auf ein ganzes Chromosom, einen Chromosomenarm oder sogar konkret auf einen bestimmten DNA-Abschnitt und damit auf ein bestimmtes Gen bezieht. Es kann ebenfalls unterschieden werden, ob durch eine aufgetretenen Mutation Informationen im Erbmaterial verloren gegangen oder z.B. durch Verdopplung hinzugekommen sind.

Auch für Krankheiten wie HIV lassen sich Progressionsmodelle berechnen. Dabei wird nicht wie beim Krebs die DNA der Tumorzellen betrachtet, sondern die DNA des HI-Virus. Dieses Virus besitzt eine schnelle Evolutionsrate und verändert sich somit häufig. Die bei der Vermehrung aufgetretenen zufälligen Mutationen können dabei zu einem Selektionsvorteil führen, wenn eine Variante des Virus' dadurch resistent gegenüber einer bestimmten Medikamentenbehandlung ist. Im Verlauf einer Behandlung können somit auch die Veränderungen der Virus-DNA beobachtet werden, die letztendlich zu einer Medikamentenresistenz führen können. Auch hier ist es wieder von großer Bedeutung, mehr über die Abhängigkeiten und Reihenfolge dieser Veränderungen zu erfahren.

Generell ist es möglich, Krankheitsprogressionsmodelle an jede Art binären Datensatz anzupassen um die Abhängigkeitsstruktur der einzelnen Ereignisse zu schätzen. Häufig werden für eine Krankheit jedoch Genexpressionswerte ermittelt, die als stetige Variablen vorliegen. Hier besteht die Möglichkeit der Dichotomisierung, wenn die Genexpressionswerte z.B. einen bestimmten, vorher zu bestimmenden Grenzwert überschritten haben. Natürlich entsteht durch diese Dichotomisierung ein Informationsverlust. Trotzdem kann durch das Berechnen einzelner konkreter Krankheitsschritte viel Information über eine Krankheit gewonnen werden.

Meistens liegen den Krankheitsprogressionsmodellen Querschnittsdaten zugrunde. Das bedeutet, dass nicht bekannt ist, zu welchem Zeitpunkt im Krankheitsverlauf die Beobachtung gemacht wurde. Der Zeitpunkt der Beobachtung (bei Krebs z.B. die operative Entfernung) kann ganz zu Beginn der Krankheit oder auch schon in einem weit fortgeschrittenen Stadium liegen. Der Datensatz gibt darüber keine Auskunft.

Es gibt jedoch auch Ansätze, denen longitudinale Daten zugrunde liegen. Hier ist z.B. das 'mutagenetic tree hidden Markov model' von Beerenwinkel und Drton (2007) zu nennen. Weiterhin gibt es auch noch das Konzept der gegenseitigen Exklusivität (engl. mutually exclusive). Hier werden genetische Ereignisse zu Gruppen zusammengefasst, wo jeweils nur ein einziges Ereignis der Gruppe eintreten kann. Wie sich eine Reihenfolge dieser Gruppen schätzen lässt, ist in Cristea et al. (2017) nachzulesen.

Insgesamt gibt es für die Modellierung der Krankheitsprogression auf Querschnittsdaten viele verschiedene Ansätze, die in Kapitel 2 dieser Arbeit beschrieben werden. Eine umfassende Bewertung bzw. ein Vergleich dieser Modelle fehlte bislang in der Literatur. In Li (2009) werden

zwar einige Vor- und Nachteile einzelner Modelle benannt, ohne diese jedoch konkret zu vergleichen und in Relation zu setzen. In Kapitel 3 wird daher auf die Frage eingegangen, ob es ein Modell gibt, das die Krankheitsprogression immer am besten beschreiben kann bzw. für welche Datensituationen welches Modell am besten geeignet ist. Außerdem ist unklar, wie sich Modellklassen verhalten und wie die berechneten Ergebnisse zu bewerten sind, wenn einige Modellannahmen verletzt sind. Obige Fragen lassen sich leicht beantworten, wenn das wahre Modell bekannt ist. Ist dies jedoch nicht der Fall, werden geeignete Modellwahlstrategien benötigt, die aus einer Menge von Modellklassen ein geeignetes Modell auswählen. Diese Problemstellung wird in Kapitel 4 behandelt. Ein weiterer wichtiger Punkt vor dem Anpassen eines Progressionsmodells ist die Auswahl der Ereignisse. Nur die Ereignisse, die für den Krankheitsverlauf eine entscheidende Rolle spielen, sollten in das Modell aufgenommen werden. In Kapitel 5 werden verschiedene Variablenselektionsmethoden vorgestellt, bewertet und auf echte Datensätze angewendet. Eine abschließende Zusammenfassung der Ergebnisse sowie ein Ausblick erfolgt in Kapitel 6.

Die Ergebnisse aus den Kapiteln 3 und 5 sind bereits in Hainke et al. (2012) bzw. Hainke et al. (2017) veröffentlicht worden.

# Kapitel 2

# Modelle zur Beschreibung von Krankheitsprogression

Für die Modellierung von Krankheitsverläufen basierend auf Querschnittsdaten gibt es bislang verschiedene Ansätze, die im Folgenden erläutert werden sollen. Ein einfaches Pfadmodell wurde bereits 1988 von Vogelstein et al. vorgestellt. Ausgehend von diesem einfachsten Baummodell entwickelten Desper et al. (1999) die onkogenetischen Bäume. Desper et al. (2000) beschäftigten sich ebenfalls mit Distanzbäumen. Auch gerichtete azyklische Graphen (Simon et al., 2000) und Kontingenzbäume (Radmacher et al., 2001) kommen in Frage, das Fortschreiten einer Krankheit zu modellieren. Ausgehend von der Arbeit von Desper et al. (1999) verwenden Beerenwinkel et al. (2005) eine Mischung onkogenetischer Bäume, die bessere Ergebnisse erzielen soll. Darüber hinaus werden Verallgemeinerungen solcher Bäume zu Netzwerk Modellen (Hjelm et al., 2006) und verbindenden Bayes Netzen (Beerenwinkel et al., 2007) erläutert.

Bevor die einzelnen Modelle jedoch beschrieben werden können, werden in Abschnitt 2.1 zunächst einige Grundlagen der Graphentheorie behandelt. In den Abschnitten 2.2 bis 2.9 werden dann die verschiedenen Modelle vorgestellt, die zur Modellierung von Krankheitsprogression verwendet werden können. Abschließend wird in Abschnitt 2.10 noch kurz auf weitere Modellklassen eingegangen, die parallel zur Entstehung dieser Dissertation entwickelt wurden.

## 2.1 Grundlagen der Graphentheorie

In der Graphentheorie ist der Baum eine spezielle Form eines Graphen. Ein *Graph G* besteht dabei aus einer endlichen Menge $V$ von *Knoten* (engl.: vertex) und einer Menge $E \subseteq V \times V$ von *Kanten* (engl.: edge), $G = (V, E)$. Eine Kante ist die Verbindung von zwei Knoten und wird daher als Tupel zweier Knoten dargestellt. Graphen werden normalerweise in grafischer Form dargestellt. Knoten werden dabei durch Punkte repräsentiert und Kanten durch Linien, die zwei Punkte verbinden. Es gibt *gerichtete* und *ungerichtete* Kanten. Gerichtete Kanten $(a, b)$ verbinden die Knoten $a$ und $b$ nur in einer gegebenen Reihenfolge, wobei der erste Eintrag im Tupel den Start- und der zweite den Endpunkt angibt. Der Startpunkt wird dabei als *Elter* und der Endpunkt als *Kind* bezeichnet. Besitzt ein Knoten $b$ mehrere eingehende

Kanten, d.h. mehrere Eltern, werden diese oft als Menge $pa(b) = parents(b)$ angegeben. Analog werden alle Kinder des Knotens $a$ in der Menge $ch(a) = children(a)$ zusammengefasst. Ungerichtete Kanten $\{a, b\}$ können in beiden Richtungen beschritten werden. Jede Kante kann ein ihr zugewiesenes Gewicht besitzen, das nie kleiner als Null ist. Unter dem *Gewicht eines Graphen* versteht man die Summe der Gewichte von allen im Graph enthaltenen Kanten. Ein ungerichteter Graph heißt *vollständig*, wenn zu jedem Knotenpaar eine Kante existiert. Ein *Teilgraph* $G' = (V', E')$ mit $V' \subseteq V$ und $E' \subseteq E$ ist ein Graph, in dem nur einige Knoten und Kanten von $G$ ausgewählt werden. Ein Graph heißt *unzusammenhängend*, wenn sich die Menge aller Knoten in zwei disjunkte Teilmengen $V_1$ und $V_2$ so unterteilen lässt, dass keine Kante existiert, die in $V_1$ beginnt und in $V_2$ endet oder umgekehrt. Ist dies nicht möglich, ist der Graph *zusammenhängend* (Wallis, 2000).

Ein *Weg* ist eine endliche Folge von Knoten $v_0, v_1, ..., v_n$ und Kanten $e_1, ..., e_n$ aus $G$: $v_0, e_1, v_1, e_2, ..., e_n, v_n$. Dabei ist $e_i$ bei ungerichteten Graphen eine Kante zwischen $v_{i-1}$ und $v_i$ und bei gerichteten Graphen eine Kante von $v_{i-1}$ nach $v_i$. Üblicherweise wird ein Weg aber nur durch seine Kantenfolge $e_1, ..., e_n$ angegeben, die ausreicht, den Weg eindeutig zu definieren. Ein *einfacher Weg* ist ein Weg, auf dem keine Kante zweimal beschritten wird. Ein *Pfad* ist ein Weg, auf dem kein Knoten zweimal besucht wird. Um zu kennzeichnen, dass gerichtete Kanten vorliegen, verwendet man auch die Bezeichnungen *gerichteter Weg* und *gerichteter Pfad*. Ein Knoten $a$ ist *Vorfahre* (auch Vorgänger) eines Knotens $b$, falls es einen gerichteten Pfad von $a$ nach $b$ gibt. Entsprechend ist $b$ ein *Nachfahre* (auch Nachfolger) von $a$. Wege, deren Start- und Zielknoten gleich sind, also $v_0 = v_n$, werden als *Kreise* oder *Zyklen* bezeichnet. In gerichteten Graphen kann ein Zyklus aus nur 2 Knoten gebildet werden, bei ungerichteten Graphen benötigt man dazu mindestens 3 Knoten.

Ein ungerichteter Graph ist, wie oben schon beschrieben, zusammenhängend, wenn es zwischen zwei beliebigen Knoten einen Weg gibt. Ein maximal zusammenhängender Teilgraph heißt *Zusammenhangskomponente*. Maximal zusammenhängend bedeutet dabei, dass es keine Kante in $G$ gibt, die einen Knoten, der nicht im Teilgraph enthalten ist, mit diesem verbindet. Andere, nicht im Teilgraph enthaltene Knoten, sind also von Knoten dieses Teilgraphen nicht zu erreichen.

Ein *Baum* ist nun ein (un)gerichteter zusammenhängender Graph, der keinen Kreis enthält. Je nachdem, ob gerichtete oder ungerichtete Kanten vorliegen, wird unterschieden zwischen ungerichteten und gerichteten bzw. gewurzelten Bäumen. Die Charakterisierung eines *ungerichteten Baums* kann zum einen über Brücken und zum anderen über die Anzahl der Kanten erfolgen. Eine *Brücke* ist dabei eine Kante, dessen Entfernung die Zahl der Zusammenhangskomponenten erhöht. Ein zusammenhängender Graph ist genau dann ein Baum, wenn alle Kanten Brücken sind. Alternativ ist ein endlicher, zusammenhängender Graph mit $v$ Knoten genau dann ein Baum, wenn er exakt $v - 1$ Kanten enthält (Wallis, 2000). Diese Definition gilt auch für gerichtete Graphen bzw. Bäume.

Werden nur gerichtete Kanten betrachtet, kann ein *gewurzelter Baum* $T = (V, E, r)$ wie folgt definiert werden. Wenn $r \in V$ der Wurzelknoten des Baums ist, können alle anderen Knoten durch genau einen gerichteten Pfad ausgehend von $r$ erreicht werden. Insgesamt muss also gelten:

(1) für jeden Knoten $v \in V$ gibt es genau eine Kante $(u, v) \in E$ mit $v$ als zweiter Komponente;
(2) es gibt keine eingehende Kante für den Wurzelknoten, d.h. für alle Knoten $u \in V$ gibt es keine Kante $(u, r) \in E$;
(3) der Graph enthält keinen Kreis.

Spezielle Formen von gewurzelten Bäumen sind Pfade und Sterne. Ein *Pfad* (bzw. genauer ein gerichteter Pfad) ist, wie oben schon definiert, ein Baum, bei dem ein Knoten höchstens eine ausgehende Kante besitzt. Ein *Stern* ist ein Baum, bei dem alle Kanten von der Wurzel ausgehen. Knoten in einem Baum, die keine ausgehende Kante besitzen, werden als *Blätter* bezeichnet.

## 2.2 Einfaches Pfadmodell

Vogelstein et al. (1988) haben bei ihrer Untersuchung von Darmkrebs-Proben kein konkretes Modell im Hinterkopf, das die Tumorprogression beschreiben soll. Ihr Ziel ist nur, etwas mehr über den Prozess der genetischen Veränderungen beim Fortschreiten der Krankheit zu erfahren. Sie verwenden dazu Darmkrebs-Proben, da diese sich gut in verschiedene Stadien einteilen lassen. Dabei wurden vier genetische Veränderungen (Mutationen und Allelverluste) genauer betrachtet.

Das Auftreten dieser genetischen Ereignisse wurde zunächst unabhängig von anderen Ereignissen untersucht. Vogelstein et al. (1988) berechnen für jede der vier Veränderungen, wie oft sie in jedem Tumorstadium auftritt. Hierbei fällt auf, dass manche genetische Veränderungen schon in frühen Krebs-Stadien auftreten und manche erst in späten. Ein weiterer Aspekt, der betrachtet wurde, ist die Anzahl der Veränderungen in den einzelnen Krebs-Stadien. Zu Beginn der Erkrankung sind nur wenige genetische Ereignisse eingetreten, während sie im Verlauf der Krankheit immer mehr zunehmen. Anhand dieser Ergebnisse zeigt sich eindeutig der fortschreitende Charakter von genetischen Veränderungen. Um diese Ereignisse in eine eventuelle Reihenfolge zu bringen, wird für jede der vier betrachteten Veränderungen berechnet, wie hoch der Prozentsatz der anderen Veränderungen ist, wenn das jeweilige Ereignis eingetreten ist oder nicht. Aus diesen Berechnungen kann jedoch keine eindeutige Reihenfolge abgeleitet werden, in der die Veränderungen auftreten könnten, da es für jede der zwei möglichen Richtungen immer noch viele Beobachtungen gibt, die dagegen sprechen.

Das Finden einer Reihenfolge von genetischen Ereignissen kann mit dem Anpassen eines Pfadmodells bzw. Pfades gleichgesetzt werden. Ein Knoten kann so zu höchstens einem anderen Knoten führen, d.h. die Ereignisse finden alle nacheinander statt. Vogelstein et al. (1988) geben jedoch kein konkretes Pfadmodell an, sondern zeigen nur mögliche Berechnungswege, die Hinweise auf eine Reihenfolge geben könnten. In Fearon und Vogelstein (1990) wird nochmals verdeutlicht, dass weniger die Reihenfolge der Ereignisse als vielmehr ihre Anzahl für die Tumorprogression verantwortlich ist.

## 2.3 Onkogenetische Bäume

Mit Hilfe der vergleichenden genomischen Hybridisierung (engl.: comparative genome hybridization, CGH) kann für verschiedene Tumore in unterschiedlichen, nicht bekannten Stadien die Informationen gewonnen werden, an welchen Chromosomen Zugewinne oder Verluste aufgetreten sind. Oft wird dabei auch noch nach p-Arm und q-Arm unterschieden, d.h. dem kurzen oder langen Arm der Chromosomen. Für jeden Tumor und damit für jeden Patienten können daher Mengen von aufgetretenen genetischen Ereignissen angegeben werden. Ein Modell für die Onkogenese[1] soll nun ein Prozess sein, der solche Mengen genetischer Ereignisse erzeugt und somit eine Verteilung über allen Mengen genetischer Ereignisse definiert (Desper et al., 1999). Sei $V$ die endliche Menge genetischer Ereignisse, die als Knotenmenge eines onkogenetischen Baums (engl.: branching tree) dienen soll. Als Wurzelknoten dient dabei eine Art 'Nullereignis', das allen Tumoren zugrunde liegt und als Krankheitsbeginn verstanden werden kann. Eine Verteilung auf $2^V$ ist eine Funktion $p$, die jeder Teilmenge $S \subseteq V$, und damit jeder möglichen Kombination von eingetretenen genetischen Ereignissen, eine nichtnegative Zahl $p(S)$ zuordnet und für die außerdem gilt $\sum_{S \subseteq V} p(S) = 1$.

Ein *markierter Baum* (engl.: labeled tree) $T = (V, E, r, \alpha)$ ist ein gewurzelter Baum mit einer Funktion $\alpha : E \to \mathbb{R}$, so dass $\alpha(e) > 0$ für alle Kanten $e \in E$. Gilt zusätzlich $0 < \alpha(e) \leq 1$, so wird der Baum $T$ als onkogenetischer Baum bezeichnet (Desper et al., 1999). Der Wert $\alpha(e)$ kann dabei als Wahrscheinlichkeit interpretiert werden, dass die Kante $e$ im Baum enthalten ist. Die Ereignisse 'Kante $e$ ist enthalten' sind dabei unabhängig voneinander.

Dieses Modell lässt sich vom Kontext der Tumorprogression insofern abstrahieren, als dass man die genetischen Ereignisse $r, 1, ..., n$ als binäre Zufallsvariablen $X_r, X_1, ..., X_n$ interpretieren kann ($n \in \mathbb{N}$), wobei diese entweder den Wert Null (Ereignis ist nicht aufgetreten) oder den Wert Eins (Ereignis ist aufgetreten) annehmen. Eine Ausnahme bildet hier der Wurzelknoten $r$, für den $P(X_r = 1) = 1$ gilt. Ein Datensatz, an den ein Progressionsmodell angepasst werden soll, wird typischerweise als eine $(N \times n)$-Matrix dargestellt, wobei $N \in \mathbb{N}$ die Anzahl der Beobachtungen, also der untersuchten Patienten bzw. Tumorproben ist. Für jede Beobachtung wird das Muster genetischer Ereignisse als Zeilenvektor $x_i = (x_{i1}, ..., x_{in})$ beschrieben mit $x_{ij} = 1$, falls das Ereignis $j$ in der $i$-ten Beobachtung eingetreten ist und $x_{ij} = 0$ sonst.

Mit Hilfe eines markierten Baums $T = (V, E, r, \alpha)$ kann eine Wahrscheinlichkeitsverteilung auf $2^V$ für $S \subseteq V$ wie folgt beschrieben werden. Falls $r \in S$ und $E' \subseteq E$ so gewählt werden kann, dass $S$ gerade die Knoten enthält, die im Baum $(V, E', r)$ von $r$ aus erreicht werden können, dann gilt

$$p(S) = \prod_{e \in E'} \alpha(e) \cdot \prod_{\substack{e=(u,v) \in E \\ u \in S, v \notin S}} (1 - \alpha(e)) \tag{2.1}$$

Falls keine Menge $E'$ gefunden werden kann, die die oben genannten Bedingungen erfüllt, gilt $p(S) = 0$. Das bedeutet, dass bestimmte Muster genetischer Ereignisse durch ein gegebenes

[1] der stochastische Prozess im Genom, der letztendlich zu Krebs führt

Baummodell nicht beschrieben werden können. Ein Lösungsansatz hierzu wird jedoch in Abschnitt 2.7 vorgestellt.

Die Wahrscheinlichkeit für das Auftreten eines genetischen Musters $x$ kann mit Hilfe von (2.1) wie folgt berechnet werden. Sei $S$ die zu $x$ gehörige Menge der aufgetretenen Ereignisse, also $S = \{x_i | x_i = 1\}$. Die Wahrscheinlichkeit dieses Musters $x$ ist dann die Wahrscheinlichkeit, dass der zugrunde liegende onkogenetische Baum $T$ das Muster $x$ erzeugt, also $P(x \mid T) = p(S)$.

Ein Beispiel für einen onkogenetischen Baum mit $n = 8$ Ereignissen ist in Abbildung 2.1 gegeben. Die Wurzel $r$ ist hier mit $E0$ bezeichnet und die einzelnen Ereignisse mit $E1, ..., E8$. Sei nun z.B. das Ereignismuster $x = (1, 0, 1, 0, 0, 0, 1, 0)$ gegeben. Die zugehörige Ereignismenge lautet $S = \{E0, E1, E3, E7\}$ und die für die Formel (2.1) benötigte Kantenmenge $E' = \{E0 \to E1, E0 \to E7, E7 \to E3\}$. Die Wahrscheinlichkeit, dass nun genau die Ereignisse aus $S$ eintreten, ein Patient also das genetische Muster $x$ besitzt, lässt sich damit berechnen zu:

$$p(\{E0, E1, E3, E7\}) = 0.6 \cdot 0.3 \cdot 0.5 \cdot (1 - 0.7) \cdot (1 - 0.3) \cdot (1 - 0.2) = 0.015 \qquad (2.2)$$

Das Auftreten einer Ereignismenge $S = \{E0, E5, E6, E8\}$ kann mit diesem Baum jedoch nicht modelliert werden und erhält die Wahrscheinlichkeit Null, da das Ereignis $E6$ erst eintreten kann, wenn $E7$ eingetreten ist.

Das bisher beschriebene Modell ist ein einfacher zeitloser onkogenetischer Baum. Desper et al. (1999) beschreiben jedoch auch ein Modell, das die Zeit miteinbezieht. Ein *zeitlicher onkogenetischer Baum* ist ein markierter Baum $T = (V, E, r, \lambda)$ mit einer zusätzlichen Verteilung $\phi$ auf den positiven reellen Zahlen. Mit Hilfe dieses Modells können auf folgende Weise Mengen genetischer Ereignisse erzeugt werden. Zunächst wird für jede Kante $e \in E$ eine exponentialverteilte Zufallsvariable $t(e)$ mit Parameter $\lambda(e)$ gezogen. Anschließend wird $t_{total}$ als eine Realisierung der Verteilung $\phi$ bestimmt. Ein Knoten $v \in V$ wird genau dann in die Menge $S$ stattgefundener genetischer Ereignisse aufgenommen, falls es einen Weg in $T$ von $r$ zu $v$ gibt, so dass die Summe aller $t(e)$'s für Kanten $e$ auf diesem Weg höchstens $t_{total}$ ist.

In einem zeitlichen onkogenetischen Baum wird angenommen, dass das Ereignis $r$ zum Zeitpunkt 0 eintritt. Ist bekannt, dass ein Ereignis $u$ eingetreten ist, so ist für alle Kanten $(u, v) \in E$ das Ereignis $v$ ein Poisson Ereignis mit Rate $\lambda$. Es werden gerade die Ereignisse ausgewählt, die bis zum Zeitpunkt $t_{total}$ aufgetreten sind, wobei $t_{total}$ als die Zeit interpretiert werden kann, zu der der Tumor bei einem bestimmten Patienten untersucht wurde.

Da als zugrunde liegende Struktur dieses Progressionsmodells ein Baum gewählt wurde, besitzt jeder Knoten nur genau einen Elternknoten. Die Wahrscheinlichkeit für das Auftreten eines genetischen Ereignisses ist somit ausschließlich von seinem Vorgänger im Baum abhängig. Obwohl diese Annahmen in Frage gestellt werden können, werden sie der Einfachheit halber getroffen und weil die Hoffnung besteht, dass trotzdem die dominanten Faktoren der Onkogenese aufgedeckt werden können. Außerdem könnte statistische Abhängigkeit genetischer Ereignisse adäquat durch unabhängige Kanten approximiert werden. Ein weiterer Einwand,

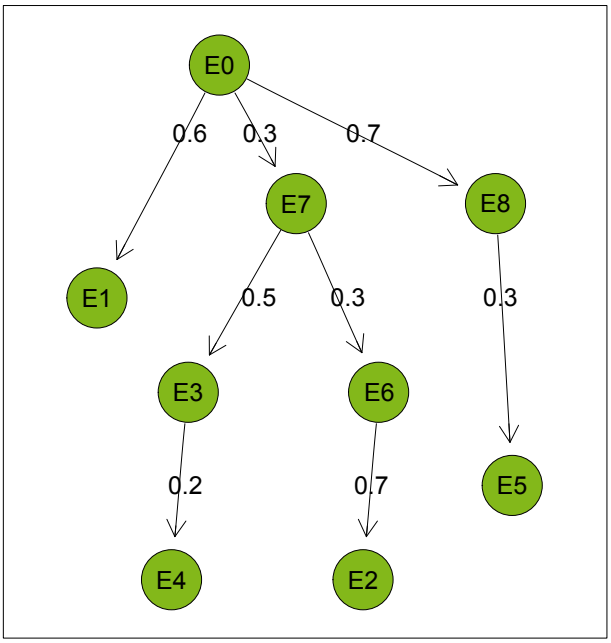



Abbildung 2.1: Beispiel für einen onkogenetischen Baum mit $n = 8$ Ereignissen

der beachtet werden muss, ist, dass die zugrunde liegenden Daten auch falsch positive oder falsch negative Ereignisse enthalten können. Falsch negative Ereignisse können durch eine Hilfsstruktur im Modell mit bedacht werden. Es besteht jedoch keine einfache Möglichkeit, falsch positive mit zu berücksichtigen (Desper et al., 1999).

### Schätzen des Modells

Liegen nun Mengen von genetischen Ereignissen als Beobachtungen vor, stellt sich die Frage, wie man aus diesen Daten einen geeigneten onkogenetischen Baum bestimmen kann. Diese Frage kann durch das Rekonstruktionsproblem mit maximaler Verzweigung (engl.: maximum branching) beantwortet werden. Ausgehend von den zugrunde liegenden Daten kann für jedes Paar $(u, v)$ genetischer Ereignisse ein Gewicht $w_{uv}$ bestimmt werden, dessen Berechnung nachfolgend beschrieben wird. Das Gewicht soll dabei ausdrücken, wie erwünscht es ist, dass im Baum das Ereignis $v$ direkt nach dem Ereignis $u$ auftritt. Mit Hilfe der logarithmierten Gewichte wird dann der gewurzelte Baum gefunden, dessen Gewicht maximal ist. Hierzu kann der Branching Algorithmus von Edmonds (Edmonds, 1967) verwendet werden.

Das Gewicht einer Kante wird ausgehend von der Wahrscheinlichkeitsfunktion $p$ wie folgt bestimmt:

$$w_{uv} = \frac{p_u}{p_u + p_v} \cdot \frac{p_{uv}}{p_u p_v}, \tag{2.3}$$

wobei $p_u = P(X_u = 1)$ und $p_{uv} = P(X_u = 1, X_v = 1)$, $u, v \in V = \{r, 1, ..., n\}$. Der erste Term des Gewichts stellt eine Asymmetrie her, so dass die Kante in Richtung des häufigeren Ereignisses ein kleineres Gewicht bekommt. Häufige Ereignisse treten damit früh im Baum auf, da Kanten, die in Richtung dieser häufigen Ereignisse zeigen, nur ein kleines Gewicht

besitzen und somit eher nicht vom Branching Algorithmus ausgewählt werden. Der zweite Ausdruck beschreibt die Korrelation zweier Ereignisse, d.h. Kanten, deren zwei Ereignisse häufig gemeinsam vorkommen, werden größere Gewichte zugewiesen, je kleiner ihre jeweiligen relativen Häufigkeiten sind. Für den Fall der Unkorreliertheit beträgt der zweite Ausdruck 1.

Im Allgemeinen müssen $p_u$ und $p_{uv}$ über die jeweiligen relativen Häufigkeiten aus den Daten geschätzt werden. Desper et al. (1999) zeigen jedoch, wenn $p$ die tatsächlich von $T$ erzeugte Wahrscheinlichkeitsverteilung ist, dann ist $\hat{T} = T$ . Es wird ebenfalls bewiesen, wenn $\hat{p}$ der Schätzer der wahren Verteilung $p$ aus $N$ Beobachtungen ist, so liefert der Algorithmus den wahren Baum $T$ mit hoher Wahrscheinlichkeit, wenn N genügend groß ist. Edmonds' Branching Algorithmus ist dabei jedoch kein echtes Maximum-Likelihood-Verfahren.

## 2.4 Distanzbäume

Desper et al. (2000) entwickelten neben onkogenetischen Bäumen auch Bäume, die auf einer gemessenen Distanz zwischen je zwei Ereignissen basieren. Die zugrunde liegende Struktur stellt zwar ebenfalls einen Baum dar, jedoch sind hier die genetischen Ereignisse nur in den Blättern des Baums zu finden. Den Vorteil dieser Betrachtungsweise sehen Desper et al. (2000) darin, dass ein Distanzbaum Auskunft über den Zusammenhang zwischen zwei beliebigen Ereignispaaren geben kann und nicht nur über einige ausgewählte Kombinationen, wie sie in onkogenetischen Bäumen vorkommen. Kein Ereignis ist somit notwendigerweise Vorgänger eines anderen Ereignisses und ein Ereignis kann nicht erst dann eintreten, wenn bestimmte Voraussetzungen erfüllt sind. Des Weiteren können aufgrund dieser etwas anderen Baumstruktur die umfangreichen Methoden genutzt werden, die im Kontext von phylogenetischen Bäumen[2] entwickelt wurden.

Ein Distanzbaum enthält also die genetischen Ereignisse in den Blättern, während die inneren Knoten beliebige unbekannte Ereignisse sind, die nicht beobachtet werden können. Es wird hier also keine Reihenfolge-Beziehung für die genetischen Ereignisse unterstellt und somit ist kein beobachtetes Ereignis Vorgänger eines anderen beobachteten Ereignisses. Die Distanz zur Wurzel gibt jedoch an, ob ein Ereignis früh oder spät eintritt. Formal lässt sich solch ein Baum als 5-Tupel $T = (V, E, r, p, L)$ darstellen. $V, E, r$ und $p$ sind dabei wie bei onkogenetischen Bäumen die Menge der Knoten bzw. Kanten, der Wurzelknoten und die Wahrscheinlichkeit $p(e)$ für jede Kante $e \in E$. Die Menge $L \subseteq V$ ist eine nichtleere Menge von Blättern, d.h. $L$ ist die Menge der beobachteten genetischen Ereignisse. Es gilt damit nicht mehr $V = L \cup \{r\}$ wie bei onkogenetischen Bäumen. Zusätzliche Ereignisse (innere Knoten) sind erlaubt.

[2] Bäume, die die evolutionären Beziehungen zwischen verschiedenen Arten darstellen, von denen vermutet wird, dass sie einen gemeinsamen Vorfahren haben; Phylogenese = stammesgeschichtliche Entwicklung aller Lebewesen

### Schätzen des Modells

Gegeben ist die Menge $L$ genetischer Ereignisse und $N$ Beobachtungen, die jeweils angeben, welche Ereignisse eingetreten sind und welche nicht. Diese Beobachtungen können als Stichprobe einer Verteilung $p$ über $2^L$ angesehen werden. Aus diesen Informationen soll nun ein Distanzbaum $T = (V, E, r, p_T, L)$ erzeugt werden, so dass $p_T$ eine gute Approximation von $p$ ist. Zunächst wird dabei aus der vorliegenden Stichprobe aus der Verteilung $p$ eine Distanzmetrik $T_L$ zwischen den Ereignissen in $L$ abgeleitet. Die Distanz zwischen zwei Ereignissen soll dabei die Stärke des Zusammenhangs ausdrücken. Anschließend soll ein Baum mit einer entsprechenden Pfadmetrik $d_T$ gefunden werden, die sehr ähnlich zu $T_L$ ist. Eine Pfadmetrik gibt dabei den Abstand zwischen je zwei Knoten im Baum an und berechnet sich wie folgt:

$$d_T(x, y) = \sum_{e \in \mathcal{P}_{xy}} d(e) \tag{2.4}$$

Dabei sind $x, y$ Knoten aus $V$, $\mathcal{P}_{xy}$ beschreibt den eindeutigen Weg in $T$ von $x$ nach $y$ und $d(e)$ gibt das Gewicht für jede Kante $e \in E$ an, wobei die Kanten hier zunächst als ungerichtet angesehen werden. Abgeleitet ist dieser Ansatz von Cavender-Farris Bäumen (Farach und Kannan, 1996), die hier jedoch nicht weiter erläutert werden sollen.

Sei also $T$ der unbekannte Distanzbaum mit $L$ als Menge aller Blätter. Gegeben seien $N$ Teilmengen von $L$, die von $T$ erzeugt worden sind. Das Rekonstruktionsproblem fordert nun, dass ein Baum $T^*$ geschätzt wird, der die gleiche Menge $L$ von Blättern enthält und dessen Verteilung $p_{T^*}$ die Verteilung $p_T$ des wahren Baums gut approximiert.

Um einen Distanzbaum erzeugen zu können, muss aus der Verteilung $p_T$ eine Pfadmetrik abgeleitet werden. Da die Kantenwahrscheinlichkeiten multiplikativ sind, bietet es sich an, als Kantengewicht den negativen Logarithmus dieser zu verwenden, $d(e) = -\log p(e)$. Aus diesen Gewichten ergibt sich dann die Pfadmetrik $d_T$. Für zwei Blätter $x$ und $y$ lässt sich die Distanz wie folgt berechnen:

$$d_T(x, y) = -2 \log p_{xy} + \log p_x + \log p_y \tag{2.5}$$

Dabei sind $p_x$ bzw. $p_{xy}$ die Wahrscheinlichkeiten, dass Ereignis $x$ bzw. die beiden Ereignisse $x$ und $y$ eintreten, die sich bei bekannter Verteilung $p$ auf $2^L$ berechnen lassen als

$$p_x = \sum_{S \subseteq L, x \in S} p(S) \qquad \text{und} \qquad p_{xy} = \sum_{S \subseteq L, \{x,y\} \subseteq S} p(S) \tag{2.6}$$

Aus den $N$ zugrunde liegenden Daten wird nun für alle $x \in L$ der Schätzer $\hat{p}_x$ berechnet als Anzahl der Teilmengen, die $x$ enthalten, dividiert durch $N$. Analog ist $\hat{p}_{xy}$ der Anteil der Teilmengen, die sowohl $x$ als auch $y$ enthalten. Für jedes $x, y$ ist dann der zugehörige Eintrag der Pfadmetrik $\hat{d}_T(x, y) = -2 \log \hat{p}_{xy} + \log \hat{p}_x + \log \hat{p}_y$. Anschließend kann über einen ausgewählten Algorithmus der Baum $T^*$ bestimmt werden, dessen Pfadmetrik $d_{T^*}$ sehr ähnlich zu $\hat{d}_T$ ist. Wenn die Beobachtungen wirklich von einem Distanzbaum stammen, konvergiert $\hat{d}_T$ mit wachsendem Stichprobenumfang gegen $d_T$. Ein möglicher Algorithmus zum Anpassen eines Distanzbaumes ist die Pivotmethode von Agarwala et al. (1999), wobei oft auch mehrere Algorithmen verwendet werden und dann der beste Baum $T^*$ gewählt wird. Alternativ zu

diesem Vorgehen stellen von Heydebreck et al. (2004) einen Maximum Likelihood Ansatz für Distanzbäume vor.

Ein Beispiel für einen Distanzbaum mit $n = 3$ Ereignissen ist in Abbildung 2.2 gegeben. Der Baum auf der linken Seite kann wie ein onkogenetischer Baum interpretiert werden. Der Baum rechts ist der zugehörige Baum mit den Distanzen, die jeweils der negative Logarithmus der entsprechenden Wahrscheinlichkeit sind. Der Abstand zwischen zwei Ereignissen berechnet sich als Distanz von Blatt zu Blatt. Für die Ereignisse $E1$ und $E2$ ergibt sich z.B. ein Abstand von $2.34 = 0.92 + 0.22 + 1.20$. Vertikale Abstände sind nur der Übersichtlichkeit halber eingefügt. Je weiter ein Ereignis von der Wurzel entfernt ist, desto später tritt es ein. In diesem Beispiel würden also die Ereignisse $E1$ und $E3$ eher früher und das Ereignis $E2$ eher später eintreten. Für größere Ereignismengen kann ein Distanzbaum auch noch Auskunft darüber geben, welche Ereignisse häufig zusammen auftreten. Diese sind als Untergruppen bzw. Cluster im Baum erkennbar.

## 2.5 Gerichtete azyklische Graphen

Eine Verallgemeinerung von Baummodellen sind gerichtete azyklische Graphen (engl.: directed acyclic graphs, DAGs). In DAG Modellen darf ein Knoten mehrere Elternknoten besitzen und nicht nur einen wie bei den bisher vorgestellten Baummodellen. Für die Modellierung der Tumorprogression bedeutet das, dass ein genetisches Ereignis von mehreren anderen Ereignissen abhängen darf. Ein Ereignis kann somit über verschiedene Pfade im Graphen erreicht werden.

Die DAG Modelle, die Simon et al. (2000) verwenden, sind wie folgt definiert. Jeder Knoten des gerichteten azyklischen Graphen ist eine Teilmenge der möglichen Ereignisse $\{1, ..., n\}$. Zusätzlich wird jedem Knoten eine Ebene zugewiesen. Auf Ebene 0 befindet sich nur ein einziger Knoten, nämlich die Wurzel $r$, also das Nullereignis, das immer stattfindet. Ausgehend von der Wurzel verlaufen $n$ gerichtete Kanten zu den verschiedenen genetischen Ereignissen. Diese erste Ebene steht für die Ereignisse, die als erstes eintreten können. Jede Kante $r \to i$ für $i \in \{1, ..., n\}$ ist dabei mit einem bestimmten Gewicht versehen, das die Wahrscheinlichkeit $p_i(1)$ angibt, dass das entsprechende Ereignis als erstes eintritt. Die Knoten auf der zweiten Ebene des DAG stellen Paare von genetischen Ereignissen dar. Für jeden Knoten bzw. jedes Ereignis $i$ auf der ersten Ebene gibt es eine Kante zu jedem Knoten $\{i, j\}$ auf der zweiten Ebene ($i \in \{1, ..., n\}, j \in \{1, ..., n\} \setminus \{i\}$). Diese Kante beschreibt die Möglichkeit, dass das Ereignis $i$ als erstes auftritt und $j$ als zweites und ist mit der bedingten Wahrscheinlichkeit $p_{i,j}(2)$ versehen, dass $j$ als zweites Ereignis eintritt gegeben, dass $i$ als erstes eingetreten ist. Von jeder 2er-Ereignismenge der zweiten Ebene gibt es gerichtete Kanten zur den 3er-Ereignismengen auf der dritten Ebene. Das Gewicht einer Kante $\{i, j\} \to \{i, j, k\}$ ($i \neq j \neq k$) ist dabei die Wahrscheinlichkeit $p_{i,j,k}(3)$, dass $k$ als drittes Ereignis eintritt, wenn $i$ und $j$ zuvor aufgetreten sind. Analog werden auch die weiteren Ebenen gebildet. Insgesamt besteht der Graph aus $n + 1$ Ebenen, wobei auf der letzten Ebene wieder nur ein Knoten existiert, der alle Ereignisse enthält. Die Ereignisse in den Knoten sind dabei ungeordnet, die Indizes an den

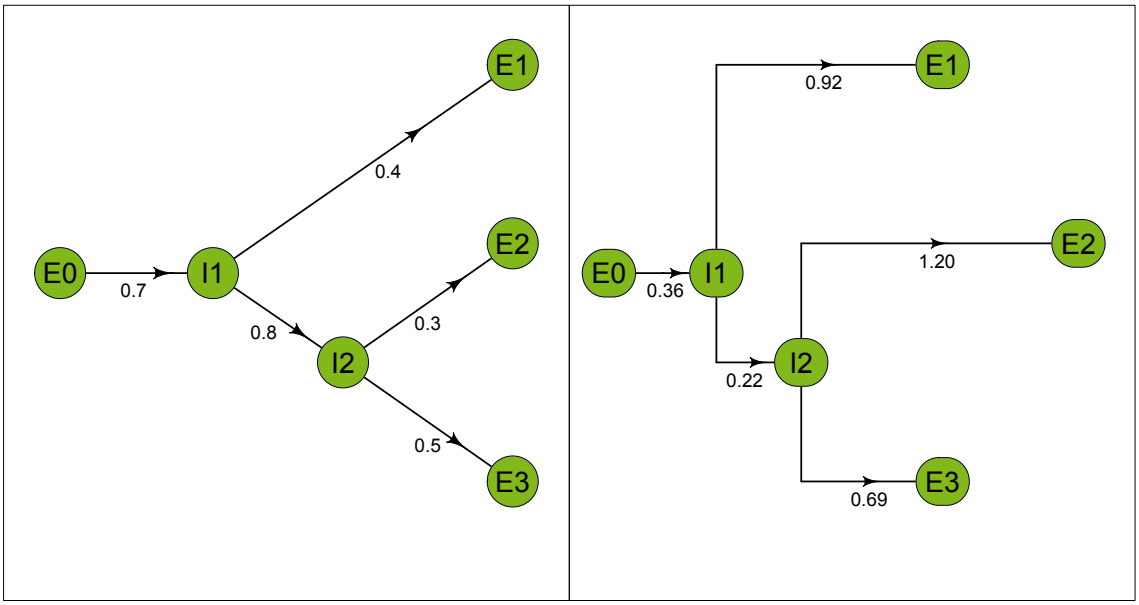



Abbildung 2.2: Beispiel für einen Distanzbaum mit $n = 3$ Ereignissen (links: mit Wahrscheinlichkeiten, rechts: mit Distanzen)

Wahrscheinlichkeiten jedoch geordnet. Zusätzlich zu den Wahrscheinlichkeiten für jede Kante müssen auch noch die Wahrscheinlichkeiten $p_0(m), m = 1, ..., n$ berücksichtigt werden, die angeben, dass höchstens $m - 1$ genetische Ereignisse auftreten.

Für $n = 3$ Ereignisse ist in Abbildung 2.3 ein Beispiel für ein DAG Modell gegeben. Wie leicht zu erkennen ist, können im Gegensatz zu onkogenetischen Bäumen alle Kombinationen und eine beliebige Reihenfolge von Ereignissen realisiert werden.

### Schätzen des Modells

Die oben genannten Wahrscheinlichkeiten können jeweils aus den Daten geschätzt werden. Ein vollständiger DAG besitzt jedoch deutlich mehr Kanten als ein Baum und dementsprechend sind auch viel mehr Parameter zu schätzen. Daher kann zum einen das Problem bestehen, dass der zugrunde liegende Datensatz zu klein ist, um ein Modell anzupassen und zum anderen, dass das Modell überangepasst wird. Simon et al. (2000) verwenden das DAG Modell aber nur, um die mit Baummodellen erzielten Ergebnisse zu überprüfen, und können daher die genannten Probleme wie folgt umgehen. Zunächst wird ein Modell angepasst, das die Unabhängigkeit der genetischen Ereignisse zugrunde legt. Falls das so berechnete Modell ungeeignet ist (z.B. in Bezug auf die Fehlerquadratsumme der im Unabhängigkeitsmodell erwarteten Anzahl genetischer Ereignisse eines bestimmten Pfades und der tatsächlich beobachteten Anzahl), werden ausgewählte Abhängigkeiten in das Modell mit aufgenommen, die die Anpassung bestmöglich optimieren. Dieses Verfahren ist analog zu dem Vorgehen der schrittweisen Regression.

Für das Unabhängigkeitsmodell müssen nur $2n - 2$ Parameter geschätzt werden. Zum einen sind das die Wahrscheinlichkeiten $p_i(1), i = 1, ..., n$, dass das Ereignis $i$ als erstes eintritt,

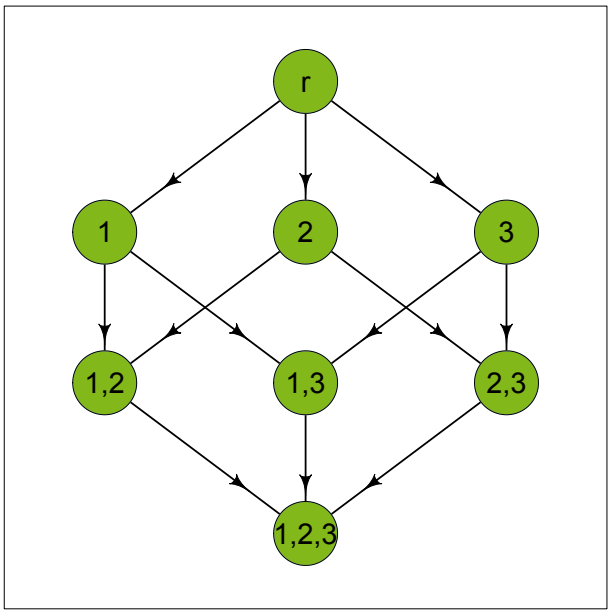



Abbildung 2.3: Beispiel für einen gerichteten azyklischen Graphen mit $n = 3$ Ereignissen

wobei $\sum_{i=1}^{n} p_i(1) = 1$, und zum anderen die Wahrscheinlichkeiten $p_0(m), m = 2, ..., n$, dass kein weiteres $m$-tes Ereignis eintritt. Beobachtungen, die überhaupt kein genetisches Ereignis aufweisen, werden ausgeschlossen. Alle weiteren Wahrscheinlichkeiten lassen sich aus diesen Parametern schätzen. Für die zweite Ebene geschieht das wie folgt und für alle anderen Ebenen analog:

$$p_{i,j}(2) = \frac{(1 - p_0(2))p_j(1)}{1 - p_i(1)} \tag{2.7}$$

Die Wahrscheinlichkeit, dass das zweite Ereignis $j$ ist unter der Bedingung, dass $i$ das erste Ereignis war, berechnet sich also als Produkt von zwei Termen. Der erste Term ist die Wahrscheinlichkeit, dass überhaupt ein zweites Ereignis auftritt, also $(1 - p_0(2))$. Der zweite Term ist die bedingte Wahrscheinlichkeit, dass $j$ das zweite Ereignis ist unter der Bedingung, dass $i$ als erstes eingetreten ist und überhaupt zwei Ereignisse auftreten. Für diesen zweiten Term werden die Ereigniswahrscheinlichkeiten für erste Ereignisse verwendet, die nur dadurch angepasst werden, dass schon das Ereignis $i$ eingetreten ist. Die Wahrscheinlichkeiten für die Ereignisse, die möglicherweise eintreten, muss sich immer zu 1 aufsummieren. Da das Ereignis $i$ schon eingetreten ist, kann es sozusagen nicht noch ein zweites Mal eintreten. Die Wahrscheinlichkeiten der anderen Ereignisse müssen also angepasst werden. Dies erfolgt so, dass die Wahrscheinlichkeit für alle Ereignisse, die noch nicht eingetreten sind, dividiert wird durch die Summe der Wahrscheinlichkeiten für die Ereignisse, die bereits eingetreten sind. Die Ereignisse der ersten Ebene haben damit keinen Einfluss darauf, welches das zweite Ereignis sein wird. Es werden schließlich die Parameter ausgewählt, die die Likelihood-Funktion

$$\prod_{v} \Big( \sum_{\mu \in P(v)} P(\mu) \Big)^{t(v)} \tag{2.8}$$

maximieren. Dabei läuft das Produkt über alle Knoten $v$ des DAG und damit über alle möglichen Ereignismengen und die Summe über alle Wege $\mu$, die von der Wurzel $r$ zum Knoten $v$ führen. Die Summanden sind dabei die Produkte der Kantenwahrscheinlichkeiten entlang der Pfade

$\mu$. Der Exponent $t(v)$ gibt an, wie oft die Ereignisse des Knotens $v$ gemeinsam in den Daten vorkommen.

Die Likelihood berechnet sich generell als Produkt der Wahrscheinlichkeiten für alle beobachteten Datenpunkte. Die Wahrscheinlichkeit für das Auftreten einer bestimmten Kombination von Ereignissen berechnet sich in diesem Fall als Summe der Wahrscheinlichkeiten über alle Wege, die diese Ereigniskombination darstellen können. Manche Kombinationen werden häufiger beobachtet, so dass das Produkt über alle Daten gleiche Faktoren enthält. Diese Tatsache führt zur genannten Formel der Likelihood-Funktion. Es wird nicht mehr das Produkt über alle Datenpunkte, sondern das Produkt über alle Ereignismengen betrachtet. Für jede Ereignismenge wird berechnet, mit welcher Wahrscheinlichkeit sie auftritt. Und zuletzt wird nur noch gezählt, wie oft diese Ereigniskombination in den beobachteten Daten auftritt.

Gerichtete azyklische Graphen werden auch von Radmacher et al. (2001) verwendet. Sie schätzen ebenfalls zum einen ein Modell unter Unabhängigkeitsannahme und zum anderen ein Modell, in das bestimmte Abhängigkeiten aufgenommen werden. Die Berechnung dieser Abhängigkeiten erfolgt jedoch auf eine andere Weise. Simon et al. (2000) erlauben in ihrer Abhängigkeitsstruktur nur, dass ein Ereignis die Wahrscheinlichkeit für das Auftreten eines direkten Nachfolgers erhöhen kann. Radmacher et al. (2001) ermöglichen jedoch auch, dass ein Ereignis die Wahrscheinlichkeit eines beliebigen anderen Ereignisses erhöhen kann, das bisher noch nicht eingetreten ist, egal wie viele andere Ereignisse noch dazwischenliegen. Die Berechnung der bedingten Wahrscheinlichkeit, dass ein Ereignis $i_2$ eintritt nachdem als erstes Ereignis $i_1$ eingetreten ist, lässt sich nicht mehr anhand von (2.7) bestimmen, sondern muss wie folgt berechnet werden:

$$p_{i_1 i_2}(2) = \frac{(1 - p_0(2)) r_{i_1 i_2} p_{i_2}(1)}{\sum_{j \neq i_1} r_{i_1 j} p_j(1)} \tag{2.9}$$

Das Skalar $r_{ij}$ gibt dabei den Grad des Zusammenhangs zwischen den Ereignissen $i$ und $j$ an. Ein Wert von 1 bedeutet, dass das Eintreten von $j$ nicht vom vorangehenden Eintreten von $i$ beeinflusst ist, während ein Wert größer als 1 einen positiven Zusammenhang kennzeichnet. Für den allgemeinen Fall, dass $i_m$ als $m$-tes Ereignis eintritt, müssen alle Einflüsse von Ereignissen berücksichtigt werden, die zuvor eingetreten sind:

$$p_{i_1 \ldots i_m}(m) = \frac{(1 - p_0(m))(\prod_{k=1}^{m-1} r_{i_k i_m}) p_{i_m}}{\sum_{j: j \notin \{i_1, \ldots, i_m\}} [(\prod_{k=1}^{m-1} r_{i_k j}) p_j]} \tag{2.10}$$

Um auch hier nicht zu viele Parameter schätzen zu müssen, gilt $r_{ij} = 1$, falls der exakte Test nach Fisher auf Unabhängigkeit einen $p$-Wert größer oder gleich 0.1 liefert. Nur für Ereignispaare mit einem kleineren $p$-Wert werden die $r_{ij}$ und, da diese nicht symmetrisch sind, auch die $r_{ji}$ berechnet.

## 2.6 Kontingenzbäume

Die Struktur der Kontingenzbäume gleicht dem Aufbau von onkogenetischen Bäumen. Es gibt eine Wurzel, die als Nullereignis interpretiert werden kann und ausgehend davon ergeben sich je nach Abhängigkeit der einzelnen Ereignisse bestimmte Pfade im Baum. Die Berechnung dieses Baummodells erfolgt allerdings auf eine andere Weise.

#### Schätzen des Modells

Zum Erzeugen eines Kontingenzbaums aus einer vorliegenden Datenmenge werden nur paarweise Abhängigkeiten betrachtet. Die Idee ist, über Fishers exakten Test für Kontingenztafeln alle paarweisen Abhängigkeiten aufzudecken. Liefert der Test Hinweise auf einen signifikant positiven Zusammenhang, so wird eine Kante zwischen den entsprechenden Ereignissen im Baum eingefügt.

Das Einfügen der Kanten in den Baum nach diesem Prinzip kann aber dazu führen, dass ein Knoten über mehrere Wege erreicht werden kann. Die Bedingungen, die ein Baum erfüllen muss, werden dadurch verletzt. Um die geforderte Baumstruktur wieder herzustellen, schlagen Radmacher et al. (2001) vor, so viele Kanten aus dem Graphen zu entfernen, so dass die Voraussetzungen an einen Baum erfüllt sind und die Likelihood Funktion des Kontingenzbaums maximiert wird. Des Weiteren liefert der exakte Test nach Fisher zwar die Information, zwischen welchen Knoten ein Zusammenhang besteht, trifft aber keine Aussage über die Richtung dieser Abhängigkeitsstruktur. Das heißt, die im Baum eingefügten Kanten besitzen noch keine Richtung. Es wird daher das Ereignis eines zusammenhängenden Ereignispaares nahe der Wurzel platziert, welches häufiger in den Daten auftritt. Diese Methode zur Bestimmung der Reihenfolge von genetischen Ereignissen entspricht dabei der Maximierung der Kantengewichte bei onkogenetischen Bäumen.

Radmacher et al. (2001) berechnen Kontingenzbäume nicht von Grund auf, sondern verwenden die Information, die ein zuvor geschätzter onkogenetischer Baum geliefert hat. Nur für die Ereignispaare, die im onkogenetischen Baum durch eine Kante verbunden sind, wird eine Kontingenztafel erstellt und Fishers exakter Test durchgeführt. In den Kontingenzbaum werden nur die Kanten übernommen, für die der Test einen $p$-Wert kleiner als 0.10 liefert. Unzuverlässige Abhängigkeiten von zwei Ereignissen werden so aus dem Baum entfernt.

## 2.7 Onkogenetische Baum-Mischungs-Modelle

Ein Problem onkogenetischer Bäume, wie sie in Abschnitt 2.3 beschrieben wurden, ist, dass bestimmte Muster genetischer Ereignisse durch ein gegebenes Baummodell nicht beschrieben werden können. Eine spezielle Baumstruktur passt häufig nur zu einem Teil des Datensatzes. Beerenwinkel et al. (2005) erklären dieses Defizit damit, dass die Daten aus mehr als

einem genetischen Prozess stammen. Sie entwickeln daher das Konzept der Baum-Mischungs-Modelle (engl.: tree mixture model). Die unterschiedlichen evolutionären Prozesse, die in einem Gen oder Genom stattfinden, sollen jeweils durch ein onkogenetisches Baummodell beschrieben werden. Um wirklich alle Kombinationen genetischer Ereignisse beschreiben zu können, wird zusätzlich eine Rauschkomponente eingeführt. In diesem Fall ist das zugrunde liegende Baummodell ein Stern, d.h. alle genetischen Ereignisse sind voneinander unabhängig.

Eine formale Definition des onkogenetischen Baum-Mischungs-Modells $M$ lautet wie folgt:

$$M = \sum_{k=1}^{K} \delta_k T_k \qquad \text{mit } \delta_k \in [0,1] \text{ und } \sum_{k=1}^{K} \delta_k = 1, \tag{2.11}$$

wobei $T_k = (V, E_k, r, \alpha_k)$, $k = 1, ..., K$, onkogenetische Bäume wie in Abschnitt 2.3 sind. Der erste Baum $T_1$ besitzt dabei allerdings die spezielle Struktur eines Sterns. Dabei wird angenommen, dass die Wahrscheinlichkeit des Eintretens für alle Ereignisse gleich ist, nämlich $\beta$, und dass alle Ereignisse unabhängig voneinander eintreten können, d.h. $\alpha_1(e) = \beta$ für alle $e \in E_1$. Durch die spezielle Wahl von $T_1$ wird sichergestellt, dass jedes Muster $x$ genetischer Ereignisse eine positive Wahrscheinlichkeit besitzt, die sich wie folgt berechnen lässt:

$$P(x \mid M) = \sum_{k=1}^{K} \delta_k P(x \mid T_k) \tag{2.12}$$

Ein Beispiel für einen onkogenetischen Misch-Baum mit $n = 5$ Ereignissen und $K = 2$ Baumkomponenten ist in Abbildung 2.4 aufgeführt. Ein Anteil von 71% der Daten kann mit Hilfe des rechten Baums abgedeckt werden. Da jedoch nicht alle Ereigniskombinationen durch diesen Baum dargestellt werden können und manche Muster somit die Wahrscheinlichkeit Null besitzen, gibt es zusätzlich die erste Baumkomponente. Diese garantiert durch ihre Sternform, dass alle Ereignisse unabhängig voneinander eintreten können und somit jede Kombination eine positive Wahrscheinlichkeit besitzt.

### Schätzen des Modells

Anhand von beobachteten Mustern genetischer Ereignisse $X = (x_{ij})_{1 \le i \le N,\ 1 \le j \le n}$ und der Anzahl $K \ge 2$ onkogenetischer Bäume soll nun ein $K$-onkogenetisches Baum-Mischungs-Modell angepasst werden. Ist bekannt, welches Muster aus welcher Baumkomponente des Mischungs-Modells stammt, kann das Verfahren aus Abschnitt 2.3 einfach $K$-mal angewendet werden, um $K$ einfache onkogenetische Bäume zu erhalten. Da diese Information aber fehlt, muss auch die Zugehörigkeit zu den Baumkomponenten aus den Daten geschätzt werden. Beerenwinkel et al. (2005) verwenden dazu einen Algorithmus, der ähnlich zu einem EM-Algorithmus (engl.: expectation maximization) ist. Ziel ist dabei die Baumkomponenten $T_1, ..., T_K$ mit zugehörigen Parametern $\delta_1, ..., \delta_K$ so zu bestimmten, dass die logarithmierte Likelihood-Funktion

$$\sum_{i=1}^{N} \log \sum_{k=1}^{K} \delta_k L(x_i \mid T_k) \tag{2.13}$$

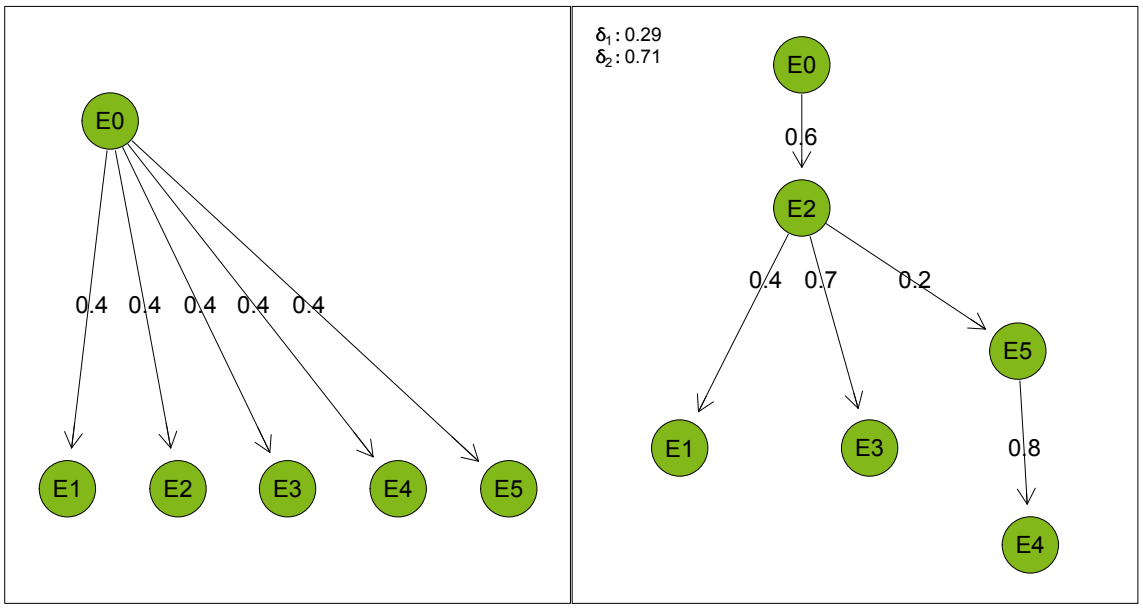



Abbildung 2.4: Beispiel für ein onkogenetisches Baum-Mischungs-Modell mit $n = 5$ Ereignissen und $K = 2$ Komponenten

maximiert wird. Dabei wird davon ausgegangen, dass die $x_i$ unabhängig voneinander sind.

Seien $\Delta_1, ..., \Delta_K$ binäre Zufallsvariablen mit $P(\Delta_k = 1) = \delta_k$, $k = 1, ..., K$. Die *Responsibility* $\gamma_{ik}$ der $k$-ten Baumkomponente $T_k$ für die $i$-te Beobachtung $x_i$ ist dann die Wahrscheinlichkeit, dass $T_k$ das Muster $x_i$ generiert, wenn das Modell $M$ zugrunde liegt, also $\gamma_{ik} = P(\Delta_k = 1 \mid M, x_i)$. Sind die Responsibilities bekannt, so kann das zugehörige Baummodell geschätzt werden und ist umgekehrt das Baummodell bekannt, so können die Responsibilities geschätzt werden. Dieser iterative Vorgang wird durch den EM-ähnlichen Algorithmus realisiert. Zunächst muss jedoch eine Initialisierung bzw. Startschätzung der Responsibilities erfolgen. Da eine zufällige Wahl der $\gamma_{ik}$ zu schlechten Ergebnissen führt, wird ein $k$-means Clusterverfahren mit $k = K - 1$ auf der Menge der genetischen Ereignisse durchgeführt. Der Wert der Responsibilities wird dann wie folgt initialisiert:

$$\gamma_{ik} = \begin{cases} \frac{1}{2}, & \text{falls } x_i \text{ zum Cluster } k-1 \text{ gehört} \\ \frac{1}{2(K-1)}, & \text{sonst} \end{cases} \tag{2.14}$$

Im Maximierungsschritt (M-Schritt) des Algorithmus werden die Modellparameter $\delta_k$ und $T_k$ ($k = 1, ..., K$) aus den Responsibilities bestimmt. Sei dazu $N_k = \sum_{i=1}^{N} \gamma_{ik}$. Da $T_1$ die spezielle Struktur eines Sterns aufweisen soll, werden hier die Gewichte der Kanten über

$$\beta = \frac{1}{nN_1} \sum_{j=1}^{n} \sum_{i=1}^{N} \gamma_{i1} x_{ij} \tag{2.15}$$

berechnet. Der Mischungs-Parameter für diese Baumkomponente beträgt $\delta_1 = \frac{N_1}{N}$. Für alle weiteren Baumkomponenten $T_k$, $k = 2, ..., K$, werden folgende Berechnungsschritte durch-

geführt. Für alle Ereignispaare $(u, v)$, $1 \leq u, v \leq n$, wird die gemeinsame Wahrscheinlichkeit

$$p_{k_{uv}} = \frac{1}{N_k} \sum_{i=1}^{N} \gamma_{ik} x_{iu} x_{iv} \tag{2.16}$$

berechnet. Aus diesen $p_{k_{uv}}$ können die $p_{k_u}$ bzw. $p_{k_v}$ und somit auch die Gewichte $w_{k_{uv}}$ für die $k$-te Baumkomponente bestimmt werden. Mit Hilfe von Edmonds' Branching Algorithmus kann dann $T_k$ genauso wie in Abschnitt 2.3 als 'maximum weight branching' ermittelt werden. Die Mischungs-Parameter berechnen sich zu $\delta_k = \frac{N_k}{N}$.

Ausgehend von einem gegebenen Baum-Mischungs-Modell können im Expectation-Schritt (E-Schritt) des Algorithmus' die Responsibilities über

$$\gamma_{ik} = \frac{\delta_k P(x_i \mid T_k)}{\sum_{m=1}^{K} \delta_m P(x_i \mid T_m)} \tag{2.17}$$

berechnet werden. Der E- und M-Schritt werden solange wiederholt, bis die log-Likelihood-Funktion nicht mehr größer wird.

## 2.8 Netzwerk Modelle

Hjelm et al. (2006) verwenden wahrscheinlichkeitstheoretische Netzwerk Modelle zur Modellierung der Krankheitsprogression. Als Grundlage dazu dienen Markov Modelle. Diese wurden auch schon von Simon et al. (2000) verwendet, die ihr Modell jedoch aufgrund der grafischen Darstellung DAG-Modell genannt haben (siehe Abschnitt 2.5).

Ein Markov Modell für die Anhäufung von genetischen Veränderungen lautet wie folgt. Sei $\{X(t) : t \geq 0\}$ ein Zufallsprozess, wobei $X(t)$ die Menge von Ereignissen beschreibt, die zum Zeitpunkt $t$ bereits aufgetreten sind. Zu Beginn ist noch kein Ereignis eingetreten, so dass $X(0) = \emptyset$. Da genetische Ereignisse nach und nach eintreten und irreversibel sind, gilt außerdem $X(t_1) \subseteq X(t_2)$, falls $t_1 < t_2$. Zusätzlich gibt es ein Stoppereignis $\mathcal{S}$, das bewirkt, dass nach dem Eintreten von $\mathcal{S}$ zum Zeitpunkt $t_i$ keine weiteren genetischen Ereignisse mehr stattfinden können, so dass $X(t_j) = X(t_i)$ für alle $t_j > t_i$.

Sei nun bekannt, dass eine Menge $Q$ von Ereignissen mit $\mathcal{S} \notin Q$ eingetreten ist. Alle Ereignisse, die bislang noch nicht aufgetreten sind, d.h. alle Ereignisse in $Q^C = (\{1, ..., n\} \setminus Q) \cup \{\mathcal{S}\}$, sind mögliche Kandidaten für das nächste Ereignis, das eintritt. Es werden folgende Annahmen gemacht:

(1) Die Zeit, bis ein Ereignis $x \in Q^C$ eintritt, wird mit $T_x^Q$ bezeichnet und ist exponentialverteilt mit Parameter $\Lambda_x(Q)$.
(2) Die $T_x^Q$ sind unabhängig für alle $x \in Q^C$.

Die erste Annahme bedeutet, dass der Prozess $X(t)$ gedächtnislos ist, d.h. $P(T_x^Q > s+t \mid T_x^Q > t) = P(T_x^Q > s)$. Obwohl $T_x$ und $T_y$ als Zeitpunkte des Eintretens von $x$ bzw. $y$ nicht unabhängig voneinander sein müssen, da $x$ das Eintreten von $y$ beeinflussen könnte, und umgekehrt, ist die zweite Annahme gerechtfertigt, da in einem gegebenen Zustand $Q$ die Anzahl der eingetretenen

Ereignisse konstant bleibt. Insgesamt wird durch diese beiden Annahmen festgelegt, dass $\{X(t) : t \geq 0\}$ ein Markov Prozess ist mit folgender Übergangswahrscheinlichkeit vom Zustand $Q$ zu $Q \cup \{x\}$:

$$p_{Q,Q\cup\{x\}} = \frac{\Lambda_x(Q)}{\sum_{y\in Q^C} \Lambda_y(Q)} \tag{2.18}$$

Sei $D = \{d_1, ..., d_k\}$ eine Teilmenge der Ereignismenge $\{1, ..., n\}$. Eine mögliche Reihenfolge der Ereignisse kann durch $d_{\sigma_1}, ..., d_{\sigma_k}$ angegeben werden, wobei sich die Wahrscheinlichkeit für das Auftreten genau dieser Reihenfolge wie folgt berechnen lässt:

$$P(d_{\sigma_1}, ..., d_{\sigma_k}) = p_{\emptyset,\{d_{\sigma_1}\}} \cdot ... \cdot p_{\{d_{\sigma_1},...,d_{\sigma_{i-1}}\},\{d_{\sigma_1},...,d_{\sigma_i}\}} \cdot ... \cdot p_{\{d_{\sigma_1},...,d_{\sigma_{k-1}}\},D} \cdot p_{D,D\cup\mathcal{S}} \tag{2.19}$$

Ein Beispiel für eine Markov Kette mit zwei Ereignissen ist in Abbildung 2.5 gegeben.

Das oben beschriebene Markov Modell benötigt einen Parameter pro Zustandspaar, was zu einer exponentiellen Anzahl von Parametern in $n$ führt. Um dies zu verhindern berücksichtigen Hjelm et al. (2006) in ihrem Netzwerk Modell nur paarweise Abhängigkeiten zwischen zwei Ereignissen, was nur zu einer quadratischen Anzahl von zu schätzenden Parametern führt.

Ein Netzwerk Modell NAM (engl.: network aberration model) ist definiert als ein Tripel $M = (\lambda, \delta, \psi)$. Dabei gibt $\lambda = \{\lambda_1, ..., \lambda_n\}$ die Veränderungsrate für jedes Ereignis zum Startzeitpunkt an. Mit $\delta = \{\delta_{ij} : i, j = 1, ..., n; i \neq j\}$ wird die Menge aller paarweisen Abhängigkeiten bezeichnet, die angibt, wie sich die Veränderungsrate für ein Ereignis $j$ ändert, wenn Ereignis $i$ eingetreten ist. Es wird angenommen, dass ein Ereignis $i$ die Veränderungsrate für ein Ereignis $j$ immer um den gleichen Faktor ändert, egal, wann das Ereignis $i$ eintritt. Für $j \in Q^C \setminus \{S\}$ gilt dann allgemein:

$$\Lambda_j(Q) = \lambda_j \cdot \prod_{i\in Q} \delta_{ij} \tag{2.20}$$

Des Weiteren wird angenommen, dass ein Ereignis die Wahrscheinlichkeit für das Auftreten eines anderen Ereignisses nur erhöhen oder unverändert lassen kann. Dass die Wahrscheinlichkeit des Auftretens verringert werden kann, wird ausgeschlossen. Es gilt daher $\delta_{ij} \geq 1$ für alle $i, j = 1, ..., n; i \neq j$. Falls $\delta_{ij} = \delta_{ji} = 1$, werden die Ereignisse $i$ und $j$ als unabhängig voneinander angesehen. Die Menge $\psi = \{\psi_1, ..., \psi_n\}$ gibt die Stoppraten an, die von der Anzahl der bisher eingetretenen Ereignisse abhängen:

$$\Lambda_{\mathcal{S}}(Q) = \psi_{|Q|} \tag{2.21}$$

Die Likelihood eines NAM $M$ für eine beobachtete Ereignismenge $D = \{d_1, ..., d_k\}$ wird über das Aufsummieren aller denkbaren Szenarien berechnet:

$$L_D(M) = P(D \mid M) = \sum_{\sigma\in S_k} p_{\emptyset,\{d_{\sigma_1}\}} \cdot ... \cdot p_{\{d_{\sigma_1},...,d_{\sigma_{k-1}}\},D} \cdot p_{D,D\cup\mathcal{S}} \tag{2.22}$$

wobei $S_k$ die Menge aller möglichen Permutationen von $\{1, ..., k\}$ ist.

Die grafische Repräsentation dieses Modells erfolgt über einen Abhängigkeitsgraphen. Jeder Knoten stellt dabei ein genetisches Ereignis dar und es gibt gerichtete Kanten von Knoten $i$ zu $j$ mit Gewicht $\delta_{ij}$ genau dann, wenn $\delta_{ij} > 1$. Zyklen sind dabei erlaubt.

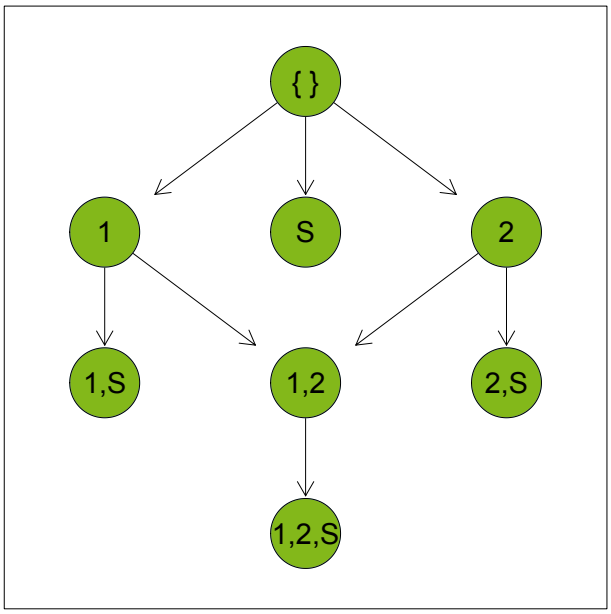


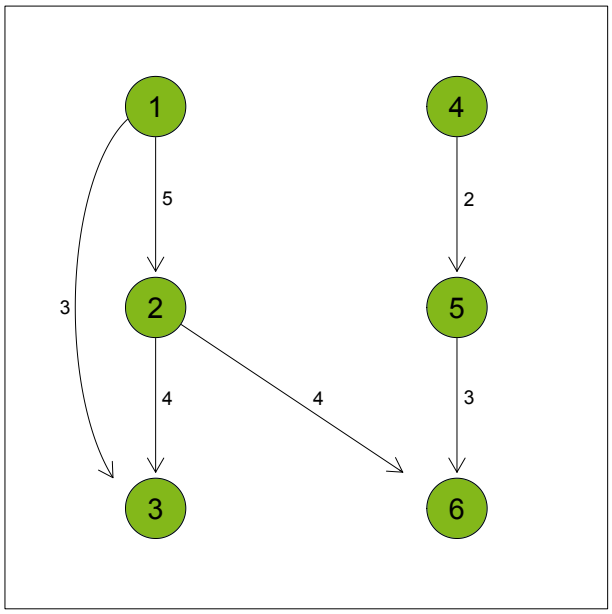

Abbildung 2.5: Beispiel für eine Markov Kette mit $n = 2$ Ereignissen

Ein Beispiel für einen einfachen Abhängigkeitsgraphen ist in Abbildung 2.6 gegeben. Man kann z.B. erkennen, dass Ereignis 1 die Wahrscheinlichkeit für das Auftreten von Ereignis 2 erhöht und Ereignis 2 die Wahrscheinlichkeit für die Ereignisse 3 und 6 erhöht, und so weiter. Ist z.B. die Ereignismenge $\{1, 2, 3\}$ beobachtet worden, kann man mit Hilfe der angegeben Werte für $\lambda, \psi$ und $\delta$ berechnen, dass die Wahrscheinlichkeit für die Reihenfolge $1 \rightarrow 2 \rightarrow 3$ ungefähr 25 mal so hoch ist wie die Wahrscheinlichkeit für die Reihenfolge $3 \rightarrow 2 \rightarrow 1$.

Im Gegensatz zu Baummodellen sind also die Annahmen, die an Netzwerk Modelle gestellt werden, nicht so stark. Der zugrunde liegende Abhängigkeitsgraph ist keiner Baumstruktur unterworfen, ein Knoten darf mehrere Vorgänger besitzen und der Graph auch Kreise enthalten. Weiterhin liegt bei Netzwerk Modellen auch keine deterministische, sondern eine probabilistische Modellierung der genetischen Ereignisse zugrunde. Ein Ereignis kann nicht erst dann eintreten, wenn bestimmte Vorgänger-Ereignisse eingetreten sind, sondern zu jedem beliebigen Zeitpunkt. Andere Ereignisse können jedoch das Eintreten positiv beeinflussen, also die Wahrscheinlichkeit für ein Ereignis erhöhen. Berücksichtigt werden hierbei aber nur paarweise Abhängigkeiten.

### Schätzen des Modells

Sei nun eine Menge $D = \{D_1, ..., D_m\}$ von Ereignismengen gegeben. Im Folgenden wird beschrieben, wie aus diesen Beobachtungen ein Netzwerk Modell geschätzt werden kann. Dazu wird der Maximum Likelihood Ansatz verwendet, d.h. es wird ein Netzwerk Modell $M$ gesucht, das die Likelihood $L_D(M) = \prod_{i=1}^{m} L_{D_i}(M)$ maximiert. Da die Likelihood Funktion sehr komplex ist, werden zur Berechnung von $M$ heuristische Methoden herangezogen. Das Vorgehen beschreiben Hjelm et al. (2006) wie folgt:




| i | $\lambda_i$ | $\psi_i$ |
|---|---|---|
| 0 | - | 1 |
| 1 | 8 | 10 |
| 2 | 5 | 15 |
| 3 | 3 | 24 |
| 4 | 8 | 27 |
| 5 | 5 | 35 |
| 6 | 3 | 42 |

Abbildung 2.6: Beispiel für einen einfachen Abhängigkeitsgraphen mit $n = 6$ Ereignissen

1. Initialisiere die Parameter $\lambda, \delta$ und $\psi$ (Initialisierungsschritt).
2. $(\lambda^{opt}, \delta^{opt}, \psi^{opt}) \leftarrow (\lambda, \delta, \psi)$
3. $\delta \leftarrow \delta^{opt}$
4. Verändere einen Parameter in $\delta$ (Modifikationsschritt).
5. Finde mit einer heuristischen Methode Werte für $\lambda$ und $\psi$, so dass $L_D(\lambda, \delta, \psi)$ nahe an $\max_{\lambda,\psi} L_D(\lambda, \delta, \psi)$ liegt (Kalibrierungsschritt).
6. Falls $L_D(\lambda, \delta, \psi) > L_D(\lambda^{opt}, \delta^{opt}, \psi^{opt})$, $(\lambda^{opt}, \delta^{opt}, \psi^{opt}) \leftarrow (\lambda, \delta, \psi)$.
7. Wiederhole die Schritte 3 - 6, bis ein lokales Maximum erreicht ist.

Initialisierung, Modifikation und Kalibrierung werden im Folgenden genauer erläutert. Im Initialisierungsschritt werden alle $\delta_{ij}$ auf 1 gesetzt, so dass zu Beginn ein Unabhängigkeitsmodell vorliegt. Die Veränderungs- und Stoppraten werden zunächst alle auf den gleichen Wert $I$ gesetzt, d.h.

$$\lambda_i \leftarrow I \quad \forall i = 1, ..., n \qquad \psi_i \leftarrow I \quad \forall i = n+1, ..., 2n+1 \tag{2.23}$$

Warum für die $\psi_i$ nicht ebenfalls die Indizes $1, ..., n$ gewählt werden, wird im weiteren Verlauf deutlich. Der Initialisierungsschritt endet mit dem Aufruf des Kalibrierungsschrittes.

Im Modifikationsschritt wird zufällig gleichverteilt ein Parameter $\delta_{ij}$ ausgewählt, der verändert werden soll. Eine Veränderung sieht dabei so aus, dass mit gleicher Wahrscheinlichkeit eins zu dem Wert $\delta_{ij}$ addiert, oder eins subtrahiert wird. Eine Ausnahme gibt es dabei für $\delta_{ij} = 1$. Hier wird mit Wahrscheinlichkeit 1 der Wert um eins erhöht, da die Abhängigkeitsparameter positive ganze Zahlen sein sollen.

Im Kalibrierungsschritt wird es aus Effizienzgründen vermieden, die Likelihood für verschiedene Parameterwerte konkret zu berechnen. Stattdessen wird mit Hilfe von künstlich erzeugten Mengen von Ereignismengen eine Über- bzw. Unterrepräsentation berechnet, die angibt, wie gut die Wahl von $\lambda$ und $\psi$ zu den Daten $D$ passt, wenn $\delta$ als bekannt vorausgesetzt

wird. Es werden $r$ künstliche Mengen $S^1, ..., S^r$ aus $(\lambda, \delta, \psi)$ erzeugt, die wie $D$ jeweils $m$ Ereignismengen enthalten. Sei $X = (X_1, ..., X_k)$ eine solche Menge von Ereignismengen. Definiere

$$X_{[j]} = \begin{cases} \{X_i \in X : j \in X_i\}, & \text{falls} \quad 1 \leq j \leq n \\ \{X_i \in X : |X_i| = j - (n+1)\}, & \text{falls} \quad (n+1) \leq j \leq 2n+1 \end{cases} \tag{2.24}$$

Für $j = 1, ..., n$, und damit für die $\lambda_i$, enthält $X_{[j]}$ gerade die Ereignismengen, die das Ereignis $j$ beinhalten. Und für $j = n+1, ..., 2n+1$, und damit für die $\psi_i$, enthält $X_{[j]}$ gerade die Ereignismengen mit genau $j-(n+1)$ Ereignissen ($j-(n+1) \in \{1, ..., n\}$). Bezeichne $\lambda_j$ als unterrepräsentiert in $S^i$, falls $|S^i_{[j]}| < |D_{[j]}|$, und als überrepräsentiert, falls $|S^i_{[j]}| > |D_{[j]}|$. Analog gilt $\psi_j$ als unterrepräsentiert, falls $|S^i_{[j+(n+1)]}| < |D_{[j+(n+1)]}|$ und als überrepräsentiert, falls $|S^i_{[j+(n+1)]}| > |D_{[j+(n+1)]}|$. Definiere

$$I(S^i, j) = \begin{cases} 1, & \text{falls} \quad |S^i_{[j]}| > |D_{[j]}| \\ -1, & \text{falls} \quad |S^i_{[j]}| < |D_{[j]}| \\ 0, & \text{sonst} \end{cases} \tag{2.25}$$

Dann wird $\lambda_j$ als unterrepräsentiertester Parameter in $S = \{S^1, ..., S^r\}$ bezeichnet, falls

$$\sum_{i=1}^{r} I(S^i, j) = \min_{k=1,...,2n+1} \sum_{i=1}^{r} I(S^i, k) \tag{2.26}$$

und als überrepräsentiertester Parameter, falls

$$\sum_{i=1}^{r} I(S^i, j) = \max_{k=1,...,2n+1} \sum_{i=1}^{r} I(S^i, k) \tag{2.27}$$

Analog dazu ist $\psi_j$ der unterrepräsentierteste Parameter, falls

$$\sum_{i=1}^{r} I(S^i, j + (n+1)) = \min_{k=1,...,2n+1} \sum_{i=1}^{r} I(S^i, k) \tag{2.28}$$

und der überrepräsentierteste Parameter, falls

$$\sum_{i=1}^{r} I(S^i, j + (n+1)) = \max_{k=1,...,2n+1} \sum_{i=1}^{r} I(S^i, k) \tag{2.29}$$

Da das Minimum und Maximum jeweils über $k = 1, ..., 2n+1$ gebildet wird, gibt es für $\lambda$ und $\psi$ nicht jeweils einen über- und unterrepräsentiertesten Parameter, sondern nur insgesamt unter allen $\lambda_i$ und $\psi_i$ einen solchen.

Falls $\sum_{i=1}^{r} I(S^i, j) \approx 0$ für alle $j = 1, ..., 2n+1$ gilt, wird $L_D(\lambda, \delta, \psi)$ als nahe bei $\max_{\lambda,\psi} L_D(\lambda, \delta, \psi)$ angesehen und die Wahl der Parameter $\lambda$ und $\psi$ wird akzeptiert. Ist dies nicht der Fall, so wird von dem überrepräsentiertesten Parameter eins abgezogen und zum unterrepräsentiertesten addiert. Mit diesen neuen Werten für $\lambda$ und $\psi$ werden wiederum künstliche Mengen von Ereignismengen erzeugt und überprüft, ob $\sum_{i=1}^{r} I(S^i, j) \approx 0$. Dies wird solange wiederholt, bis $L_D(\lambda, \delta, \psi)$ als nahe bei $\max_{\lambda,\psi} L_D(\lambda, \delta, \psi)$ angesehen und zum 6. Schritt des Algorithmus' übergegangen werden kann. Um den Kalibrierungsschritt etwas zu beschleunigen wird anfangs z.B. mehr als eins von den zwei entsprechenden Parametern subtrahiert bzw. addiert.

Um ein Netzwerk Modell wie oben beschrieben schätzen zu können, werden $n^2+n+1$ Parameter benötigt. Um die Anzahl der Parameter noch weiter zu reduzieren, stellen Hjelm et al. (2006) zusätzlich ein Netzwerk Modell basierend auf Modulen vor (MNAM, engl.: module NAM). Ereignisse, die Abhängigkeiten miteinander teilen, d.h. die gleichen Abhängigkeiten zu anderen Ereignissen aufweisen, können zu einem Modul zusammengefasst werden. Für die genaue Definition geteilter Abhängigkeiten und die Anpassungen des Algorithmus' zum Schätzen des Modells siehe Hjelm et al. (2006). Des Weiteren können in einem NAM bzw. MNAM Kreise auch vermieden werden (siehe hierzu ebenfalls Hjelm et al. (2006)).

## 2.9 Verbindende Bayes Netze

Als eine Verallgemeinerung der onkogenetischen Bäume von Desper et al. (1999) verwenden Beerenwinkel et al. (2007) verbindende Bayes Netze (engl.: conjunctive Bayesian networks, CBNs). Ein CBN ist eine Spezialisierung eines Bayes Netzes mit dem Unterschied, dass ein Ereignis nur dann auftreten kann, wenn alle Eltern-Ereignisse aufgetreten sind. Im Gegensatz zu onkogenetischen Baummodellen, bei denen ein Ereignis höchstens von einem vorher aufgetretenen Ereignis abhängen kann, erlauben CBNs, dass mehrere Ereignisse ein Folgeereignis beeinflussen. Somit kann eine größere Anzahl von Problemen modelliert werden.

Ein CBN als wahrscheinlichkeitstheoretisches Modell besteht aus einer endlichen Menge von binären Zufallsvariablen, den Ereignissen, und einer partiellen Ordnung, die die Abhängigkeiten der Ereignisse angibt. Eine partielle Ordnung erfüllt dabei die Bedingungen der Transitivität, Reflexivität und Antisymmetrie. Man kann sich diese partielle Ordnung als einen gerichteten azyklischen Graphen (DAG) vorstellen, dessen Kanten die Ordnungsbeziehung wiedergeben. Im Vergleich zu allgemeinen Bayes Netzen sind CBNs immer noch sehr eingeschränkt, haben aber den Vorteil, dass eine Maximum Likelihood Schätzung durchgeführt werden kann, was für grafische Modelle eher unüblich ist. Außerdem ist die Anzahl der Parameter eines CBNs nicht abhängig von der Graphenstruktur. Des Weiteren besitzen CBNs wünschenswerte algebraische, statistische und kombinatorische Eigenschaften (siehe Beerenwinkel et al. (2007), Kapitel 5). Sie können effizient gelernt werden, Rauschen in den Daten berücksichtigen und sind auf den untersuchten zwei Datensätzen besser als onkogenetische Bäume.

Formal lässt sich ein CBN Modell charakterisieren als ein 3-Tupel $(\mathcal{E}, \leq, \theta)$. Dabei ist $\mathcal{E}$ die Menge genetischer Ereignisse, $\leq$ die partielle Ordnung und $\theta = (\theta_1, ..., \theta_n)$ ein Parametervektor, der für jedes Ereignis $e \in \{1, ..., n\}$ die bedingte Wahrscheinlichkeit angibt, dass $e$ eingetreten ist, gegeben, dass alle Vorgänger-Ereignisse eingetreten sind. Die partielle Ordnung $e_1 < e_2$ für zwei Ereignisse bedeutet, dass Ereignis $e_1$ vor Ereignis $e_2$ eintritt. Der Zustandsraum des CBN Modells ist ein Gitter $\mathcal{G} = \mathcal{J}(\mathcal{E})$ von Ordnungsidealen in $\mathcal{E}$. Ein *Ordnungsideal* ist dabei eine Teilmenge $g \subseteq \mathcal{E}$, so dass wenn $e_2 \in g$ und $e_1 < e_2$ auch $e_1 \in g$. Die Elemente von $\mathcal{G}$ werden *Genotypen* genannt. Ein Genotyp ist also eine Teilmenge von $\mathcal{E}$ bzw. ein Bitstring, der für jedes Ereignis angibt, ob es aufgetreten ist, oder nicht.

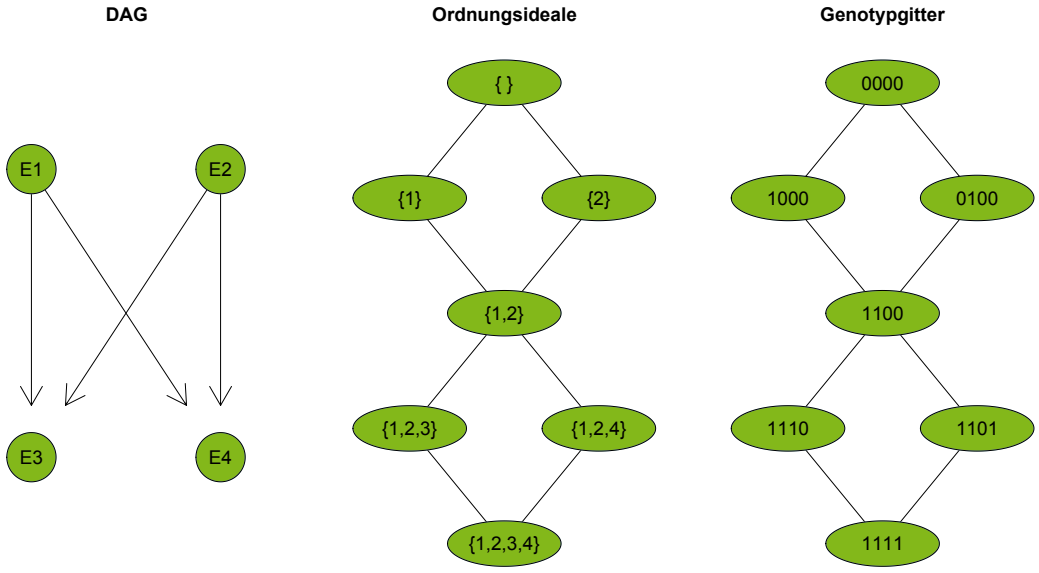



Abbildung 2.7: Beispiel für ein verbindendes Bayes-Netz mit $n = 4$ Ereignissen

Ein Beispiel für ein CBN mit $n = 4$ Ereignissen ist in Abbildung 2.7 gegeben. Zusätzlich zum gerichteten azyklischen Graphen, der das CBN darstellt, sind auch die Ordnungsideale und das Genotypgitter aufgeführt. Sowohl das dritte als auch das vierte Ereignis können erst eintreten, wenn die ersten beiden Ereignisse eingetreten sind. Bestimmte Ordnungsideale bzw. Genotypen können somit überhaupt nicht eintreten (z.B. die Genotypen $0010, 1010, 0111$). Aus den $16 = 2^4$ möglichen Genotypen können für das gegebene Modell nur 7 realisiert werden. Alle anderen Genotypen besitzen die Wahrscheinlichkeit Null.

Sei $\min(g^c)$ die Menge der kleinsten Elemente des Komplements $g^c = \mathcal{E}\backslash\{g\}$, d.h. die Menge der Ereignisse, die in $g$ noch nicht aufgetreten sind, aber als nächstes eintreten könnten. Dann lässt sich analog zu (2.1) die Wahrscheinlichkeit, einen Genotyp $g \in \mathcal{G}$ des zugrunde liegenden CBN Modells zu beobachten, wie folgt berechnen:

$$P_g(\theta) = \prod_{e \in g} \theta_e \cdot \prod_{e \in \min(g^c)} (1 - \theta_e) \tag{2.30}$$

Das CBN Modell für eine Menge $\mathcal{E}$ (oft auch der Einfachheit halber direkt als $\mathcal{E}$ bezeichnet) ist also ein gerichtetes grafisches Modell für binäre Zufallsvariablen $(X_e)_{e \in \mathcal{E}}$. Der zugehörige Graph enthält Kanten $e \to f$, falls $e < f$ für zwei Ereignisse $e$ und $f$ aus $\mathcal{E}$ erfüllt ist. Die bedingten Wahrscheinlichkeiten lassen sich in folgender Matrixschreibweise angeben, wobei mit $pa(e)$ die Menge der Eltern von $e$ bezeichnet ist, also die Ereignisse, die $e$ direkt voraus gehen:

$$[P(X_e = b \mid X_{pa(e)} = a)]_{a \in \{0,1\}^{pa(e)},\ b \in \{0,1\}} = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 - \theta_e & \theta_e \end{bmatrix}. \tag{2.31}$$

### Maximum Likelihood Schätzung

Sei nun bekannt, dass $n$ genetische Ereignisse auftreten können. Sei $\mathcal{E}$ die Menge dieser Ereignisse und $\mathcal{G}$ das zugehörige Gitter. Die zugrunde liegenden Daten dieses CBN Modells haben die Form einer Funktion $u : \mathcal{G} \to \mathbb{N}$ mit $g \mapsto u_g$, wobei $u_g$ angibt, wie oft der Genotyp $g$ beobachtet worden ist. Ist solch ein Datensatz $u \in \mathbb{N}^{\mathcal{G}}$ gegeben, kann der Maximum Likelihood Schätzer $\hat{\theta}_e$ von $\theta_e$ für jedes Ereignis $e \in \mathcal{E}$ wie folgt berechnet werden:

$$\hat{\theta}_e = \frac{\sum_{g:e\in g} u_g}{\sum_{g:\text{below}(e)\subseteq g} u_g} \qquad \text{für alle } e \in \mathcal{E}. \tag{2.32}$$

Der Schätzer ergibt sich also als relativer Anteil der Genotypen, die das Ereignis $e$ enthalten, zu den Genotypen, die die Eltern-Ereignisse $pa(e)$ enthalten.

Um nun einen Algorithmus angeben zu können, der zu einem gegebenen Datensatz $u$ ein CBN Modell liefert, müssen noch einige Begriffe erläutert werden. Die Elemente von $\mathcal{G}$ lassen sich als Bitstrings in $\{0,1\}^n$ darstellen. Jedem dieser Elemente kann eine Häufigkeit zugewiesen werden. Sei supp($u$) die Menge der Genotypen, die im Datensatz auftreten. Man kann dann sagen, dass $u$ *die Ereignisse trennt*, wenn für zwei beliebige Ereignisse $e, f$ ein Genotyp $g \in \text{supp}(u)$ existiert, so dass $g \cap \{e, f\}$ entweder $\{e\}$ oder $\{f\}$ ergibt. Ist dies nicht der Fall, können die Ereignisse $e$ und $f$ als ein gemeinsames Ereignis $\{e, f\}$ aufgefasst werden und die Anzahl der Ereignisse kann auf $n-1$ reduziert werden. Ein beliebiger Genotyp $g$ ist *kompatibel* mit dem Modell $(\mathcal{E}, \leq, \theta)$, wenn $g \in \mathcal{G}$ oder äquivalent, wenn $P_g(\theta)$ nicht Null ist. Ein Datensatz $u$ heißt *kompatibel* mit $\mathcal{E}$, wenn alle $g \in \text{supp}(u)$ kompatibel mit $\mathcal{E}$ sind. Beerenwinkel et al. (2007) beweisen dann folgendes Theorem: Sei $u$ ein Datensatz mit zugehöriger Verteilung, der die Ereignisse trennt. Dann gibt es eine eindeutige größte Menge $\mathcal{E}_u$ mit entsprechender partieller Ordnung, so dass $u$ kompatibel mit $\mathcal{E}_u$ ist. Dieses $\mathcal{E}_u$ ist dann das eindeutige Maximum Likelihood CBN Modell für $u$.

Der Algorithmus, der für einem gegebenen Datensatz $u$ das CBN Modell $\mathcal{E}_u$ und den Maximum Likelihood Parameter $\hat{\theta}$ liefert, lautet schließlich wie folgt. Gegeben ist der Datensatz $u$ mit zugehöriger Verteilung bzw. Häufigkeitsangabe für alle Genotypen aus $\{0,1\}^n$. In einem ersten Schritt wird überprüft, ob $u$ die $n$ Ereignisse trennt. Falls nicht, werden die entsprechenden Ereignisse vereint und $u$ sowie die zugehörigen Häufigkeiten entsprechend angepasst. Im zweiten Schritt kann die Menge $\mathcal{E}_u$ mit partieller Ordnung wie folgt definiert werden. Für zwei beliebige Ereignisse $e, f$ wird $e < f$ genau dann in das Modell aufgenommen, wenn $g \cap \{e, f\} \neq \{f\}$ für alle $g \in \text{supp}(u)$. Für jedes Ereignis $e$ kann dann im dritten Schritt $\hat{\theta}_e$ anhand von (2.32) berechnet werden.

Im Allgemeinen ist es schwierig, den obigen Algorithmus auf reale Datensätze anzuwenden, da diese verrauscht sein können. Eine Relation $e < f$ kann nur dann in das Modell aufgenommen werden, wenn kein Genotyp beobachtet wurde, der $f$ enthält, aber nicht $e$. So kann der Algorithmus z.B. eine starke Relation fälschlicherweise nicht aufnehmen, nur weil die Daten verrauscht sind. Eine Lösung hierzu liefert der Ansatz, eine Auswahl von CBN Modellen mit unterschiedlicher Fehlertoleranz zu bilden.

Sei dazu $\mathcal{E}_\epsilon$ so, dass alle Relationen $e < f$ enthalten sind, bei denen höchstens ein Anteil $\epsilon$ der Daten dagegen spricht. Für $\epsilon = 0$ würde man also $\mathcal{E}_u$ von oben erhalten. Im Allgemeinen werden einige $g \in \text{supp}(u)$ nicht kompatibel mit $\mathcal{E}_\epsilon$ sein. Bevor daher $\theta$ geschätzt wird, werden diese Genotypen entfernt. Um dennoch alle Genotypen, kompatibel oder inkompatibel, im Modell zu berücksichtigen, wird folgendes Mischungsmodell verwendet. Dazu sei $\mathcal{G}_\epsilon = \mathcal{J}(\mathcal{E}_\epsilon)$ der Raum möglicher Genotypen für das Modell $\mathcal{E}_\epsilon$. Für die Genotypen $g \notin \mathcal{G}_\epsilon$, die inkompatibel mit $\mathcal{E}_\epsilon$ sind, wird angenommen, dass sie mit einer Wahrscheinlichkeit von $1/(2^n - |\mathcal{G}_\epsilon|)$ auftreten. Das Mischungsmodell $\mathcal{E}'_\epsilon$ lässt sich dann mit den Ereigniswahrscheinlichkeiten $\theta_e$ und einem Mischungsparameter $\lambda$ für alle Beobachtungen $g \in \{0,1\}^n$ wie folgt angeben:

$$P'_g(\theta, \lambda) = \begin{cases} \lambda P_g(\theta), & \text{wenn } g \in \mathcal{G}_\epsilon \\ (1-\lambda)(2^n - |\mathcal{G}_\epsilon|)^{-1}, & \text{wenn } g \notin \mathcal{G}_\epsilon \end{cases} \tag{2.33}$$

Über die log-Likelihood Funktion der Daten $u : \{0,1\}^n \to \mathbb{N}$, die im einfachen Modell die Gestalt

$$\ell_u(\theta) = \sum_{g \in \mathcal{G}} u_g \cdot \left( \sum_{e \in g} \log \theta_e + \sum_{e \in \min(g^c)} \log(1 - \theta_e) \right) \tag{2.34}$$

hatte, und sich nun zu

$$\ell'_u(\theta) = \sum_{g \in \mathcal{G}_\epsilon} u_g [\log \lambda + \log P_g(\theta)] + \sum_{g \notin \mathcal{G}_\epsilon} u_g [\log(1-\lambda) - \log(2^n - |\mathcal{G}_\epsilon|)] \tag{2.35}$$

ändert, können die Parameter $\theta$ und $\lambda$ geschätzt werden. Für $\theta$ lassen sich die Schätzer analog zum einfachen CBN Modell über (2.32) berechnen. Der Schätzer für $\lambda$ ergibt sich aus dem Anteil der Daten, der mit dem Modell $\mathcal{E}_\epsilon$ kompatibel ist:

$$\hat{\lambda} = \frac{\sum_{g \in \mathcal{G}_\epsilon} u_g}{\sum_g u_g} \tag{2.36}$$

Für gegebene Daten $u$ kann für verschiedene Werte von $\epsilon$ das Modell $\mathcal{E}_\epsilon$ gebildet werden. Als CBN Modell wird das Modell gewählt, das die log-Likelihood $\ell'_u$ maximiert. Um die Signifikanz des Unterschiedes zwischen verschiedenen Modellen oder auch zu onkogenetischen Bäumen zu überprüfen, können Bootstrap Stichproben behilflich sein.

## 2.10 Weitere Modellklassen

Neben den oben beschriebenen Modellklassen gibt es weitere zusätzliche Progressionsmodelle, die als Erweiterung und Ergänzung entwickelt wurden. Diese sollen im Folgenden kurz beschrieben werden.

Basierend auf der Idee der verbindenden Bayes Netze (CBNs) sind CT-CBNs (engl.: continuous time conjunctive Bayesian networks) (Beerenwinkel und Sullivant, 2009) und H-CBNs (engl.: hidden conjunctive Bayesian networks) (Gerstung et al., 2009) entstanden. Im Vergleich zu CBNs kann bei CT-CBNs das Fortschreiten der Krankheit mit Hilfe einer stetigen Modellierung

beschrieben werden, da als Grundlage kein diskreter, sondern ein stetiger Markovprozess verwendet wird. Die Dauer bis zum Auftreten eines Ereignisses, gegeben, dass alle Elternereignisse eingetreten sind, wird hier als exponentialverteilte Wartezeit modelliert. Diese Wartezeit beinhaltet sowohl das Entstehen der Mutation als auch deren benötigte Zeit, sich in der Population durchzusetzen. Damit kann ein expliziter Zeitablauf der genetischen Ereignisse angegeben werden, anhand dessen quantitative Vorhersagen über die Geschwindigkeit der Krebsentstehung getroffen werden können. H-CBNs können darüber hinaus latente Variablen berücksichtigen. Ein zusätzliches Fehlermodell kann dabei Beobachtungsfehler beschreiben, die das Erfassen des wahren Krankheitsprozesses erschweren. Dadurch können Beobachtungen erklärt werden, die nicht zum geschätzten Modell passen.

Eine weitere Modellklasse, die ebenfalls Beobachtungsfehler berücksichtigt, sind die HOTs bzw. HOT-mixtures (engl.: hidden-variable oncogenetic trees) (Tofigh, 2009; Tofigh et al., 2011). Jeder Knoten des gerichteten Baummodells ist dabei mit einer latenten und einer beobachteten Variable assoziiert. Der Wert der latenten Variable gibt an, ob die Tumorprogression diesen bestimmten Knoten erreicht hat, während der Wert der beobachteten Variable angibt, ob diese Mutation auch wirklich erkannt wurde. Wie bei onkogenetischen Bäumen und deren Mischungsmodellen kann also die Wahrscheinlichkeit berechnet werden, dass bestimmte Mutationen eintreten. Zusätzlich gibt jedoch ein weiterer Wahrscheinlichkeitsprozess an, ob diese Mutationen auch beobachtet wurden, wenn sie im dahinter liegenden latenten Prozess eingetreten sind.

Während für die drei oben beschriebenen Modelle EM-Algorithmen zum Schätzen entwickelt wurden, werden für die Progressions-Netzwerke (engl.: Progression Networks) (Shahrabi Farahani und Lagergren, 2013) Methoden des Mixed Integer Linear Programming (MILP) verwendet, um ein spezielles Bayes Netz zur Modellierung der Krankheitsprogression zu schätzen. Um eine Reihenfolge der genetischen Ereignisse abzubilden, wird jeder Kante des Bayes Netzes eine obere Schranke für die Wahrscheinlichkeit zugewiesen, dass das Nachfolge-Ereignis eintritt ohne dass alle Elternereignisse beobachtet wurden. Im Vergleich zum H-CBN erzielen die Progressions-Netzwerke auf einem ausgewählten Datensatz übereinstimmende Ereignispfade. Insgesamt sind sie auch in der Lage eine größere Anzahl an Variablen zum Schätzen eines Modells zu berücksichtigen.

Der RESIC Algorithmus (engl.: retracing the evolutionary steps in cancer) (Attolini et al., 2010; Cheng et al., 2012) soll ebenfalls dazu beitragen, die zeitliche Abfolge genetischer Ereignisse während der Tumorentwicklung abzuleiten. Zunächst werden dabei nur Ereignispaare betrachtet und berechnet, welches dieser beiden Ereignisse am wahrscheinlichsten zuerst eintritt. In einem weiteren Schritt wird der RESIC-Algorithmus dazu verwendet, wirkliche Ereignis-Pfade anzugeben, die sich ähnlich eines CBNs darstellen lassen.

Das CAPRESE Framework (engl.: Cancer progression extraction with single edges) (Loohuis et al., 2014) stellt eine neue Methodik dar, um Progressionsbäume aus Querschnittsdaten abzuleiten. Wo bisher eine Kombination aus Korrelation und Häufigkeiten verwendet wurde, um eine Ereignisreihenfolge zu schätzen, wird hier der Ansatz der wahrscheinlichkeitstheoretischen Kausalität nach Suppes (1970) genutzt. Um eine gewisse Robustheit gegenüber Rauschen herzustellen, wird außerdem ein Shrinkage ähnlicher Schätzer verwendet, der die Kausalität

zwischen zwei Ereignissen misst. Ein Ereignis A verursacht demnach ein Ereignis B, wenn A zeitlich vor B eintritt und außerdem das Auftreten von A die Wahrscheinlichkeit erhöht, dass auch B beobachtet wird. Im Vergleich zu onkogenetischen Bäumen und verbindenden Bayes Netzen schneidet CAPRESE auf synthetischen Datensätzen besser ab und zeigt auf relativ kleinen Datensätzen eine nennenswerte Effizienz.

Eine weitere Möglichkeit, ein grafisches Modell einer Krankheitsprogression zu schätzen, stellt der CAPRI Algorithmus (engl.: cancer progression inference) (Ramazzotti et al., 2015) dar. Dieser Algorithmus basiert auf einer wahrscheinlichkeitstheoretischen Scoring Methode, gekoppelt mit Bootstrap und Maximum Likelihood Schätzung. Das Vorhandensein von Rauschen in den Daten kann berücksichtigt werden und auch mit nur kleinen Datensätzen kann der Algorithmus gut umgehen und zuverlässige Ergebnisse liefern.

# Kapitel 3

# Vergleich der Modelle

In diesem Kapitel sollen nun verschiedene Progressionsmodelle miteinander verglichen werden, um Vor- und Nachteile der jeweiligen Modellklassen herauszustellen. Ebenfalls soll die Frage beantwortet werden, ob es ein bestes Modell gibt, das man immer oder auch nur in bestimmten Datensituationen verwenden sollte. Dazu wird eine Simulationsstudie durchgeführt, in der alle Modelle jeweils an Daten aus einem gewählten bekannten Modell angepasst werden. Dabei kann überprüft werden, wie gut die anpassende Modellklasse das wahre Modell abbilden kann. Bevor jedoch die Vergleiche konkret durchgeführt werden, sollen die dabei verwendeten Methoden beschrieben werden.

## 3.1 Methoden

### 3.1.1 Wilcoxon-Mann-Whitney Test

Der Wilcoxon-Mann-Whitney Test ist ein nichtparametrischer Test, der auf der Verwendung von Rängen beruht. Mit Hilfe dieses Tests kann überprüft werden, ob die Verteilungen von zwei Stichproben identisch sind, oder eine der Zufallsvariablen größere bzw. kleinere Werte liefert. Das parametrische Gegenstück des Wilcoxon-Mann-Whitney Tests ist unter der Normalverteilungsannahme der $t$-Test.

Ein intuitiver Ansatz zur Beantwortung der Fragestellung des Tests besteht darin, beide Stichproben zu vereinen und jedem Wert einen Rang in der Gesamtstichprobe zuzuweisen. Ergibt die Summe der Ränge der einen Stichprobe einen zu kleinen oder zu großen Wert, spricht das dafür, dass die zugehörige Zufallsvariable allgemein kleinere bzw. größere Werte annimmt. Das Verwenden von Rängen ist dann von Vorteil, wenn die Verteilung der Zufallsvariablen nicht bekannt ist, da die Verteilung von Teststatistiken, die auf Rängen basieren, unter recht allgemeinen Annahmen unabhängig von der Verteilung der Daten ist.

Im Folgenden werden die Voraussetzungen, die Teststatistik sowie die entsprechenden Hypothesen des Wilcoxon-Mann-Whitney Tests vorgestellt, wie sie in Conover (1999) beschrieben sind.

Gegeben seien zwei Stichproben $X_1, ..., X_n$ und $Y_1, ..., Y_m$ vom Umfang $n$ bzw. $m$. Sei $N = n + m$ die Größe der vereinigten Stichprobe. Den Beobachtungen werden die Ränge 1 bis $N$

von der kleinsten bis zur größten zugewiesen. Dabei bezeichne $R(X_i)$ bzw. $R(Y_j)$ den Rang von $X_i$ bzw. $Y_j$ für alle $i$ und $j$. Sollten Bindungen auftreten, das heißt, zwei oder mehrere gleiche Beobachtungswerte, werden Durchschnittsränge vergeben. Dies ist eine mögliche Vorgehensweise beim Auftreten von Bindungen. Alternativ können z.B. die entsprechenden Rangwerte zufällig den gleichen Beobachtungen zugeordnet werden um Dezimalzahlen bei den Rängen zu vermeiden. Bei der Vergabe von Durchschnittsrängen ist zu beachten, dass der Test dadurch konservativer wird und das Niveau $\alpha$ nicht ausschöpft.

Folgende Voraussetzungen müssen erfüllt sein, damit der Wilcoxon-Mann-Whitney Test anwendbar ist:

1. Beide Stichproben sind zufällig aus ihrer entsprechenden Verteilung gezogen worden.
2. Unabhängigkeit liegt sowohl innerhalb als auch zwischen den beiden Stichproben vor.
3. Die Daten sind mindestens ordinal skaliert.

Für den Fall, dass keine oder nur wenige Bindungen vorliegen, wird als Teststatistik die Summe der Ränge aus der ersten Stichprobe verwendet:

$$T = \sum_{i=1}^{n} R(X_i) \tag{3.1}$$

Liegen viele Bindungen vor, so wird die Teststatistik $T$ angepasst, indem das arithmetische Mittel von $T$ subtrahiert und durch die Standardabweichung dividiert wird:

$$T_1 = \frac{T - n\frac{N+1}{2}}{\sqrt{\frac{nm}{N(N-1)} \sum_{i=1}^{N} R_i^2 - \frac{nm(N+1)^2}{4(N-1)}}} \tag{3.2}$$

Dabei ist $\sum_{i=1}^{N} R_i^2$ die Summe der quadrierten Ränge oder Durchschnittsränge von allen Beobachtungen aus beiden Stichproben.

Der Wilcoxon-Mann-Whitney Test wird in dieser Arbeit dazu verwendet die beiden folgenden Hypothesen bezüglich der Zufallsvariablen $X$ und $Y$ zu überprüfen. Zum einen soll getestet werden, ob die Zufallsvariable $X$ kleinere Werte liefert als $Y$ und zum anderen, ob die beiden Zufallsvariablen die gleiche Verteilung besitzen. Das Vorgehen zum dritten Test-Szenario, dass die Zufallsvariable $X$ größere Werte liefert als $Y$, ist in Conover (1999) auf S.275 nachzulesen.

Für den einseitigen Test, dessen Alternativhypothese den Gedanken „X neigt dazu, kleiner zu sein als Y“ ausdrückt, lässt sich das Testproblem wie folgt formulieren. $F$ bzw. $G$ sind dabei die zu $X$ bzw. $Y$ gehörenden Verteilungsfunktionen.

$$\begin{aligned} H_0 &: F(x) = G(x) \qquad \forall x \in \mathbb{R} \\ H_1 &: F(x) > G(x) \qquad \forall x \in \mathbb{R} \end{aligned}$$

Die Nullhypothese wird zum Niveau $\alpha$ abgelehnt, falls die Teststatistik $T$ kleiner als das entsprechende $\alpha$-Quantil $w_\alpha$ ist. Für $n, m \leq 20$ sind diese Quantile in Conover (1999) vertafelt.

Bei größeren Stichprobenumfängen werden approximative Quantile verwendet, die sich wie folgt aus der Standardnormalverteilung ergeben:

$$w_\alpha \approxeq \frac{n(N+1)}{2} + z_\alpha \sqrt{\frac{nm(N+1)}{12}}, \tag{3.3}$$

wobei $z_\alpha$ das $\alpha$-Quantil der Standardnormalverteilung ist. Wird bei vorliegenden Bindungen die Teststatistik $T_1$ benutzt, wird die Nullhypothese zum Niveau $\alpha$ abgelehnt, falls $T_1 < z_\alpha$. Der $p$-Wert lässt sich für große Stichprobenumfänge approximativ bestimmen durch

$$p \approxeq P\left(Z \leq \frac{T + \frac{1}{2} - n\frac{N+1}{2}}{\sqrt{\frac{nm(N+1)}{12}}}\right), \tag{3.4}$$

wobei $Z$ eine standardnormalverteilte Zufallsvariable ist. Für $T_1$ gilt direkt $p \approxeq P(Z \leq T_1)$.

Das Testproblem des zweiseitigen Tests, der überprüft, ob zwei Zufallsvariablen die gleiche Verteilung zugrunde liegt, lautet:

$$\begin{aligned} H_0 &: F(x) = G(x) \qquad \text{für alle } x \in \mathbb{R} \\ H_1 &: F(x) \neq G(x) \qquad \text{für mind. ein } x \in \mathbb{R} \end{aligned}$$

In diesem Fall wird die Nullhypothese zum Niveau $\alpha$ verworfen, falls die Teststatistik $T$ kleiner ist als das $\frac{\alpha}{2}$-Quantil oder größer als das $(1 - \frac{\alpha}{2})$-Quantil. Das Berechnen bzw. Ablesen der Quantile erfolgt genauso wie beim einseitigen Testproblem. Analoges gilt für den Fall, dass die Teststatistik $T_1$ verwendet wird. Der approximative $p$-Wert für die Teststatistik $T_1$ berechnet sich dann zu

$$p \approxeq 2 \cdot min\{P(Z \leq T_1), P(Z \geq T_1)\}. \tag{3.5}$$

Für den zweiseitigen Wilcoxon-Mann-Whitney Test ist zu bemerken, dass die Power des Tests nur gut ist, wenn die Alternativhypothese besagt, dass eine der beiden Zufallsvariablen größere Werte liefert als die andere. Für andere Alternativen, die sich z.B. auf einen Unterschied in der Streuung beziehen, hat der Test wenig Power. In einem solchen Fall kann auf den zweiseitigen Kolmogorov-Smirnov Test zurückgegriffen werden, der gute Power gegen mehr und allgemeinere Alternativhypothesen besitzt. Für die hier vorgestellte Alternative, dass eine Zufallsvariable größere Werte liefert als die andere, besitzt der KS-Test jedoch weniger Power.

## 3.2 Simulationsstudie

### 3.2.1 Fragestellung und Ziel

In diesem Abschnitt sollen die vorgestellten Progressionsmodelle miteinander verglichen werden. Jedes Modell hat andere Möglichkeiten, den Krankheitsverlauf darzustellen. Daher besitzt jedes Modell eine andere Komplexität. Eine Anordnung von einfach zu komplex lässt sich jedoch nicht angeben, da nicht alle Modelle ineinander geschachtelt sind. Jedes Modell stellt

bestimmte Annahmen an die Krankheitsprogression und kann somit den Verlauf der Krankheit in einer bestimmten Weise darstellen. Ein Beispiel hierfür ist, dass es in manchen Modellen nur genau ein Vorgänger-Ereignis gibt, in anderen aber mehrere Ereignisse eintreten müssen, um das Auftreten eines weiteren Ereignisses zu ermöglichen. Mit Hilfe einer Simulationsstudie soll geklärt werden, ob komplexe Modelle generell immer besser sind oder ob einfache Modelle auch in komplexeren Situationen sinnvolle Ergebnisse liefern und die Krankheitspfade ausreichend darstellen können. Ebenso interessant ist es zu erfahren, wie gut ein Modell abschneidet, wenn im Voraus schon bekannt ist, dass die Daten eine Struktur aufweisen, die das Modell nicht darstellen kann. Wird dennoch ein Teil der Krankheitsverläufe sinnvoll wiedergegeben oder sind die Ergebnisse komplett nutzlos?

Die grundlegende Idee der Simulationsstudie ist daher, sich ein wahres Modell vorzugeben und aus diesem Modell Daten zu ziehen. An diese Daten soll nun wieder ein Modell angepasst werden. Für diese Anpassung werden verschiedene Modellklassen verwendet. Anschließend wird verglichen, wie gut das geschätzte und das wahre Modell zueinander passen. Welche Abstandsmaße und Kennzahlen für diesen Vergleich in Frage kommen, wird in Abschnitt 3.2.2 beschrieben. Eine ausführliche Beschreibung des Aufbaus der Simulationsstudie erfolgt in Abschnitt 3.2.3. Weiterhin werden einige Analyseschritte für die Auswertung der Ergebnisse in Abschnitt 3.2.4 vorgestellt.

### 3.2.2 Abstandsmaße und Kennzahlen

Für den oben beschriebenen Ansatz der Modellanpassung soll nun bestimmt werden, wie gut die anpassende Modellklasse das wahre Modell erkennen kann. Der Abstand zwischen wahrem und angepasstem Modell soll dabei möglichst gering sein. Um aber überhaupt einen Abstand angeben zu können, werden entsprechende Abstandsmaße benötigt. Es gibt zwei verschiedene Ansätze dafür. Zum einen kann ein Progressionsmodell über die zugrunde liegende Wahrscheinlichkeitsverteilung charakterisiert werden. Für jede Ereigniskombination $x \in \{0,1\}^n$ kann mit Hilfe des Modells die Wahrscheinlichkeit ausgerechnet werden, dass genau diese Ereignisse eingetreten sind. Ein anderer Ansatz betrachtet die Topologie der Modelle, d.h. die zugehörige Baumstruktur der Knoten und Kanten. Die Kantengewichte sowie die Gewichte einzelner Baumkomponenten werden hierbei außer Acht gelassen. Im Folgenden werden diese beiden Ansätze und mögliche zugehörige Abstandsmaße genauer definiert.

#### Abstand der induzierten Wahrscheinlichkeitsverteilung

Betrachtet wird zunächst die induzierte Wahrscheinlichkeitsverteilung. Sowohl das wahre als auch das angepasste Modell liefern Wahrscheinlichkeiten für jede Ereigniskombination. Für jedes Modell liegen insgesamt $2^n$ Werte vor. Diese werden in einem Vektor dargestellt. Sei $d_w$ der Vektor der Ereigniswahrscheinlichkeiten des wahren Modells und $d_a$ der des angepassten. Dabei ist natürlich in jedem Eintrag $i$ der Vektoren die Wahrscheinlichkeit für jeweils die gleiche Ereigniskombination gespeichert. Optimal wäre es, wenn die Wahrscheinlichkeiten des wahren

und angepassten Modells für die Ereigniskombinationen übereinstimmen, also $d_w = d_a$ ist. Man hofft also auf einen kleinen Abstand der Vektoren. Für diese Berechnungen gibt es unterschiedliche Abstandsmaße. In diesem Fall sollen wie in Beerenwinkel et al. (2005) der $L_1$-, der $L_2$- und der Cosinus-Abstand betrachtet werden. Für zwei Vektoren $x$ und $y$ der gleichen Länge $l$ sind diese wie folgt definiert:

$$d_{L_1}(x, y) = \sum_{i=1}^{l} |x_i - y_i| \tag{3.6}$$

$$d_{L_2}(x, y) = \sqrt{\sum_{i=1}^{l} (x_i - y_i)^2} \tag{3.7}$$

$$d_{Cos}(x, y) = 1 - \cos \sphericalangle(x, y) = 1 - \frac{< x, y >}{||x||||y||} = 1 - \frac{\sum_{i=1}^{l} x_i \cdot y_i}{\sqrt{(\sum_{i=1}^{l} x_i^2) \cdot (\sum_{i=1}^{l} y_i^2)}} \tag{3.8}$$

Der Cosinus-Abstand berechnet sich also über den aufgespannten Winkel der Wahrscheinlichkeitsvektoren. Sind diese sehr ähnlich, ist der Winkel und damit auch der Abstand klein, ist der Winkel groß, ist der Abstand ebenfalls groß, z.B. wenn die Vektoren orthogonal aufeinander stehen, d.h. die Wahrscheinlichkeiten sehr unterschiedlich sind.

Andere mögliche Abstandsmaße, die die Abweichung zwischen diskreten Wahrscheinlichkeitsverteilungen betrachten, sind die Kullback-Leibler Divergenz (Kullback und Leibler, 1951), die totale Variation (Rachev, 1991) und die Chernoff Distanzen (Chernoff, 1952). Die Kullback-Leibler Divergenz ist jedoch hier nicht anwendbar, da manche Wahrscheinlichkeiten den Wert 0 annehmen und die totale Variation ist ein Vielfaches der $L_1$-Distanz.

Die zuvor beschriebenen Methoden, den Abstand der induzierten Wahrscheinlichkeitsverteilungen zu betrachten, berücksichtigen die im Baum enthaltenen Kanten nur indirekt. Natürlich werden die Kanten zur Berechnung der Wahrscheinlichkeiten benötigt, aber durch diesen Wert wird nicht deutlich, zwischen welchen Knoten genau die Kanten verlaufen und zwischen welchen nicht. Auch Modelle, die nicht exakt die gleichen Kanten aufweisen, können einen kleinen Abstand der Wahrscheinlichkeitsverteilungen besitzen. Darüber hinaus sind auch die einzelnen Pfade im Baummodell von großem Interesse. Es ist wichtig, in welcher Reihenfolge bestimmte Ereignisse eintreten können. Die dahinter liegenden Wahrscheinlichkeiten können dabei sogar zweitrangig sein. Daher werden zusätzliche Abstandsmaße definiert und verwendet, die nur die im Baum enthaltenen Kanten berücksichtigen und die zugehörigen Wahrscheinlichkeiten vernachlässigen.

#### Zusätzliche und fehlende Kanten

Zunächst ist es sinnvoll, sich anzuschauen, welche Kanten im wahren und angepassten Modell enthalten sind. Anschließend kann man vergleichen und berechnen, wie viele Kanten jeweils gleich sind, wie viele Kanten im angepassten Modell fehlen und wie viele Kanten zusätzlich enthalten sind, die das wahre Modell nicht aufweist. Dabei muss bedacht werden, dass wahres und angepasstes Modell aus unterschiedlichen Modellklassen stammen können und somit auch durchaus Kanten aus einem onkogenetischen Baum oder verbindenden Bayes Netz mit Kanten

aus einem Baum-Mischungs-Modell verglichen werden müssen. Wie in einer solchen Situation vorgegangen werden kann, wird im Folgenden beschrieben. In dieser Simulation werden in der Klasse der Baum-Mischungs-Modelle nur solche betrachtet, die eine Sternkomponente besitzen. Beim Kantenvergleich kann diese Sternkomponente jedoch vernachlässigt werden, da sie keine Informationen über die Reihenfolge von Ereignissen enthält, da diese alle unabhängig voneinander eintreten können. Die Struktur der Sterne ist damit immer festgelegt. Bei einem Baum-Mischungs-Modell mit 2 Komponenten muss daher nur mit den Kanten der zweiten Komponente verglichen werden. Bei 3 Komponenten werden die Kanten aus der zweiten und dritten Komponente zusammengefasst, wobei Kanten, die in jeder Baumkomponente vorkommen, nur einmal gezählt werden. Insgesamt können dadurch natürlich mehr als $n$ Kanten auftreten. Dementsprechend erhöht sich aber einfach die Zahl der fehlenden oder zusätzlichen Kanten. Ein Vergleich ist trotzdem möglich.

#### Prozentsatz wiedergefundener Kanten

Ein einfaches Maß für die Güte der Anpassung ist der Prozentsatz der wiedergefundenen Kanten, welches von Tofigh et al. (2011) verwendet wird. Hierbei wird lediglich berechnet, wie viele der wahren Kanten, d.h. der Kanten, die das wahre Modell enthält, auch im angepassten Modell enthalten sind. Sternkomponenten bleiben hierbei wieder unberücksichtigt. Bei dieser Berechnung kann es allerdings auftreten, dass der Prozentsatz bei 100% liegt, das angepasste Modell aber dennoch nicht dem wahren entspricht, weil durch weitere Baumkomponenten zusätzliche Kanten auftreten können. Dieses Maß gibt also nur wieder, ob alle im wahren Modell enthaltenen Pfade aufgedeckt werden können. Ob noch zusätzliche und somit eigentlich fehlerhafte Informationen im angepassten Modell enthalten sind, wird nicht deutlich.

#### Recovery, Precision, Dissimilarity

Ein anderer Ansatz, der von Yin et al. (2006) im Rahmen der Modellwahl bei Baum-Mischungs-Modellen entwickelt wurde (siehe auch Kapitel 4), betrachtet beides. Zum einen gibt es ein Maß, das wie der Prozentsatz der wiedergefundenen Kanten von Tofigh et al. (2011) überprüft, wie gut das wahre Modell wiedergefunden wurde, und zum anderen ein Maß, welches kontrolliert, wie gut das angepasste Modell zum wahren passt. Ein drittes Maß kombiniert die Aspekte der vorangegangenen. Die Maße heißen *recovery*, *precision* und *dissimilarity* und sind wie folgt definiert:

$$recov = \frac{1}{K} \sum_{k} \max_{l} S_{kl} \tag{3.9}$$

$$prec = \frac{1}{\hat{K}} \sum_{l} \max_{k} S_{kl} \tag{3.10}$$

$$dissim = \sum_{(k,l) \in A} (1 - S_{kl}) + |K - \hat{K}| \tag{3.11}$$

Dabei gibt $S_{kl}$ die Ähnlichkeit zwischen zwei Baumkomponenten $T_k$ und $T_l$ an:

$$S_{kl} = S_{lk} = 1 - \frac{||A_k - A_l||_\infty}{n} \quad \in [0, 1]. \tag{3.12}$$

Dabei sind $A_k$ und $A_l$ $(n \times n)$-Adjazenzmatrizen der Baumkomponenten $T_k$ und $T_l$, die in den Spalten $j$ kennzeichnen, ob Knoten $i$ ein Nachfolger des $j$-ten Knotens ist (1 für 'ja' und 0 für 'nein'). Die Unendlichnorm einer Matrix $||A||_\infty = max_i \sum_j |a_{ij}|$ bestimmt das Maximum der absoluten Zeilensummen. Mit $||A_k - A_l||_\infty$ wird also der größte Unterschied an ausgehenden Kanten zwischen den zwei Baumkomponenten bestimmt. Weiterhin ist $K$ die Anzahl der Baumkomponenten im wahren Modell und $\hat{K}$ die im angepassten. Der Index $k$ läuft über die Baumkomponenten im wahren Modell und der Index $l$ über die im angepassten. Beim dritten Maß ist die Menge $A$ das maximale $S_{ij}$-gewichtete bipartite Matching zwischen den Baumkomponenten der entsprechenden Modelle. Das bedeutet folgendes: Man versucht, die beste Zuordnung im Sinne der größten Ähnlichkeit von Baumkomponenten des wahren und angepassten Modells zu finden. Dabei darf im Gegensatz zu den ersten beiden Maßen keine Baumkomponente doppelt vergeben werden. Es darf aber nicht einfach für die erste wahre Baumkomponente die am besten passende aus dem angepassten Modell gewählt und dann genauso für die weiteren wahren Baumkomponenten verfahren werden. Es kann nämlich vorkommen, dass die zweite wahre Baumkomponente viel besser zu derjenigen im angepassten Modell passt, die zuvor schon an die erste wahre Baumkomponente vergeben wurde und somit als Matchingpartner nicht mehr zur Verfügung steht. Daher muss zunächst für alle Kombinationen von wahren und angepassten Baumkomponenten die Summe der Ähnlichkeiten berechnet werden. Bei einer ungleichen Anzahl von Baumkomponenten im wahren und angepassten Modell bleiben einige Baumkomponenten übrig. Die Kombination, die den größten Wert bei der Summe der Ähnlichkeiten erzielt, bildet die Menge $A$. In $A$ sind also Paare $(k, l)$ enthalten, wobei jeder Wert von $k = 1, ..., K$ und $l = 1, ..., \hat{K}$ höchstens einmal vorkommen darf. Um den eventuellen Unterschied in der Anzahl der Baumkomponenten zwischen wahrem und angepasstem Modell zu berücksichtigen, wird der Strafterm $|K - \hat{K}|$ hinzuaddiert.

Das erste Maß *recov* summiert also die Ähnlichkeiten der wahren Baumkomponente(n) zu der am besten passenden Baumkomponente des geschätzten Modells. Damit wird gemessen, wie gut die wahre Baumstruktur erkannt wird, d.h. wie gut die jeweiligen wahren Baumkomponenten im angepassten Modell wiedergegeben werden. Das zweite Maß *prec* summiert die Ähnlichkeiten der angepassten Baumkomponente(n) zu der am besten passenden Baumkomponente des wahren Modells. Hiermit wird ausgedrückt, wie präzise die wahre Struktur aufgedeckt wird, d.h. wie gut die angepassten Baumkomponenten jeweils zu den wahren passen. Der Wertebereich von *recovery* und *precision* liegt jeweils zwischen 0 und 1. Ergebnisse nahe bzw. gleich 1 sind simultan für beide Maße erstrebenswert. Ein hoher Wert der Trefferquote, also *recov* nahe 1, besagt, dass das wahre Modell gut durch das angepasste beschrieben werden kann. Ist allerdings gleichzeitig die Genauigkeit, also *prec*, deutlich kleiner als 1, bedeutet das, dass das angepasste Modell zwar die Informationen des wahren wiedergeben kann, gleichzeitig aber noch zusätzliche und damit fehlerhafte Informationen enthält. Das dritte Maß *dissim* kombiniert die beiden vorangegangenen Aspekte und misst die Unähnlichkeit. Hier sind kleine Werte nahe 0 erstrebenswert, da so die Ungleichheit am geringsten ist. Bei der Berechnung der ersten

beiden Maße wird üblicherweise die Sternkomponente, die bei Baum-Mischungs-Modellen auftreten kann, nicht mit berücksichtigt. Lediglich beim dritten Maß *dissim* wird sie beim Strafterm $|K - \hat{K}|$ miteinbezogen. Eine Variante für die Berechnung der Unähnlichkeit ist, die Sternkomponente im Strafterm nur dann zu berücksichtigen, falls ihr Gewicht größer oder gleich 0.1 ist. Das heißt nur dann, wenn mehr als 10% der Daten durch diese Baumkomponente erklärt werden, wird sie in den Strafterm aufgenommen. Das Nichtberücksichtigen einer Sternkomponente aus dem angepassten Modell ist allerdings nur dann sinnvoll, wenn das wahre Modell aus nur einer Baumkomponente besteht und das anpassende Modell aus mehreren. Besitzt das wahre Modell mehr Baumkomponenten als das angepasste, ist es unsinnig aus dem angepassten eine weitere Baumkomponente zu entfernen, da das den Abstand nur weiter vergrößert. Besitzen beide Modelle mehr als eine Baumkomponente stimmen die Sterne zumindest schon mal überein, so dass das Nichtberücksichtigen der einen Sternkomponente sinnlos wäre, da dadurch Ähnlichkeit verloren geht.

Ein bei diesem Ansatz noch zu optimierender Aspekt ist die Definition der Ähnlichkeit, die in (3.12) vorgestellt wurde. Diese berücksichtigt nur den größten Unterschied an ausgehenden Kanten zwischen den Baumkomponenten. Tritt dieser maximale Unterschied häufiger auf, fällt dies nicht ins Gewicht. Allerdings ist intuitiv eine Komponente, die diesen maximalen Unterschied nur einmal aufweist, trotzdem noch ähnlicher als eine Komponente, bei der dies häufiger auftritt. Ein solcher Unterschied kann aber durch das beschriebene Ähnlichkeitsmaß nicht dargestellt werden. Daher ist zu überlegen, ob z.B. die Unendlichnorm durch ein anderes Abstandsmaß ersetzt werden sollte, das alle auftretenden Unterschiede berücksichtigt.

#### Graph edit distance

Ein weiteres, allgemeines Abstandsmaß für Graphen ist die *graph edit distance*, auch *ged* (siehe z.B. Bunke und Shearer (1998)), die sich aus den zusätzlichen und fehlenden Kanten ableitet. Der Abstand zwischen zwei Graphen ist definiert als kleinste Anzahl von Operationen, die notwendig sind, um den einen Graphen in den anderen zu überführen. Solche Operationen sind Hinzufügen und Entfernen von Kanten und Knoten. Den einzelnen Operationen können zusätzlich Kosten zugewiesen werden, so dass sich in einem solchen Fall der Abstand zwischen zwei Graphen über die Abfolge von Operationen ergibt, die in der Summe die geringsten Kosten aufweisen. Bei den vorliegenden Progressionsmodellen werden jedoch keine Kosten definiert, bzw. das Hinzufügen und Entfernen von Kanten kostet jeweils genau eine Einheit. Operationen, die die Knoten betreffen, werden nicht benötigt, da die Knotenmenge von wahrem und anpassendem Modell immer gleich ist. Um die *graph edit distance* zu berechnen, muss also gezählt werden, wie viele Kanten mindestens hinzugefügt und entfernt werden müssen, um den Graphen vom angepassten Modell so zu verändern, dass er den vom wahren Modell darstellt. Eine möglichst geringe Zahl an Kantenoperationen ist erstrebenswert. Auch hier stellt sich wieder die Frage, wie bei einem Vergleich vorgegangen werden soll, wenn mehrere bzw. eine unterschiedliche Zahl von Baumkomponenten vorliegen. Dazu wird wie beim *dissimilarity*-Score auf ein bipartites Matching zurückgegriffen. Man bestimmt zunächst die Kombination der Baumkomponenten aus wahrem und angepasstem Modell, die zu der kleinsten Summe von

*graph edit distances* führt und addiert im Fall einer ungleichen Anzahl von Baumkomponenten $n \cdot |K - \hat{K}|$ hinzu. Dieser Summand ergibt sich aus der Überlegung, dass aus den zusätzlichen Baumkomponenten alle Kanten entfernt werden müssen, damit ein leerer Baum ohne Information entsteht, bzw. bei fehlenden Baumkomponenten alle Kanten hinzugefügt werden müssen, damit aus einem leeren Baum ohne Kanten der gewünschte entsteht. Beim Vergleich der einzelnen Baumkomponenten verschiedener Modelle wird die Sternkomponente außen vor gelassen. Sie wird jedoch im Strafterm $n \cdot |K - \hat{K}|$ berücksichtigt. Auch hier kann wie im Abschnitt zuvor die Variante eingeführt werden, die Sternkomponente des angepassten Modells nur dann zu bestrafen, falls sie ein Gewicht größer oder gleich 0.1 aufweist. Aus den gleichen Gründen wie oben ist dies nur dann angebracht, wenn das wahre Modell aus einer und das anpassende Modell aus mehreren Baumkomponenten besteht.

Einen Nachteil dieses Ansatzes soll folgendes Beispiel verdeutlichen. Gegeben seien die beiden Bäume in Abbildung 3.1 als wahres und angepasstes Modell. Auf Kantenwahrscheinlichkeiten kann hier verzichtet werden, da sie bei diesem Abstandsmaß ohne Bedeutung sind. Das wahre Modell ist also ein Pfad mit der Ereignis-Reihenfolge 1,2,3,4,5. Das angepasste Modell liefert ebenfalls einen Pfad, allerdings in etwas anderer Reihenfolge: 4,5,1,2,3. Der Abstand des Graphen, der sich mit Hilfe der *graph edit distance* berechnet, ist gleich 2: die Kante vom fünften zum ersten Ereignis muss entfernt und die Kante vom dritten zum vierten Ereignis hinzugefügt werden. Es sind damit zwei Kantenoperationen notwendig, um das angepasste Modell ins wahre zu überführen. Der Abstand über die *ged* ist damit sehr gering. Allerdings ist die Interpretation der beiden Baummodelle sehr unterschiedlich. Im wahren Modell kann das vierte Ereignis erst eintreten, nachdem 3 andere Ereignisse zuvor beobachtet worden sind. Das anpassende Modell sieht Ereignis 4 jedoch als erstes Ereignis an, ohne das die anderen Ereignisse gar nicht eintreten können. Trotz des kleinen Abstandes auf der Ebene der Kantenoperationen ist die Interpretation der Modelle grundverschieden. Ein solches Phänomen kann die *graph edit distance* nicht aufdecken und somit zu falschen Schlussfolgerungen führen. Alternativ sind Abstandsmaße denkbar, die die Tiefe im Baum mit berücksichtigen, die also die Anzahl der Vorgänger mit einbeziehen.

Eine andere, etwas von der Definition abweichende Berechnungsweise der *graph edit distance* ist Folgende: Wie zu Beginn beschrieben, wird ausgerechnet, wie viele Kanten im angepassten Modell zusätzlich vorhanden sind und wie viele fehlen. Ein Addieren dieser beiden Werte besagt, wie viele Kanten hinzugefügt bzw. entfernt werden müssen, um das angepasste Modell ins wahre zu überführen, also genau das, was die *graph edit distance* berechnen soll. Allerdings werden beim Zählen der zusätzlichen und fehlenden Kanten doppelte Kanten, die in mehreren Baumkomponenten vorkommen, nicht berücksichtigt, da sie die gleiche Information über die Ereignisreihenfolge enthalten. Damit ist die Anzahl der Kantenoperationen höchstens so groß wie die Anzahl, die sich mit der ersten beschriebenen Berechnungsmethode für die *ged* ergibt. Man muss sich also überlegen, ob eine zusätzliche Baumkomponente als Ganzes bestraft werden soll, d.h. alle ihre Kanten als zusätzliche Kanten entfernt werden müssen (Variante 1), oder ob Kanten, die in einer anderen Baumkomponente enthalten sind, also die gleiche und damit keine falsche Information tragen, erlaubt sind (Variante 2). Enthält das angepasste Modell weniger Baumkomponenten als das wahre und damit auch zu wenige Kanten, würde die

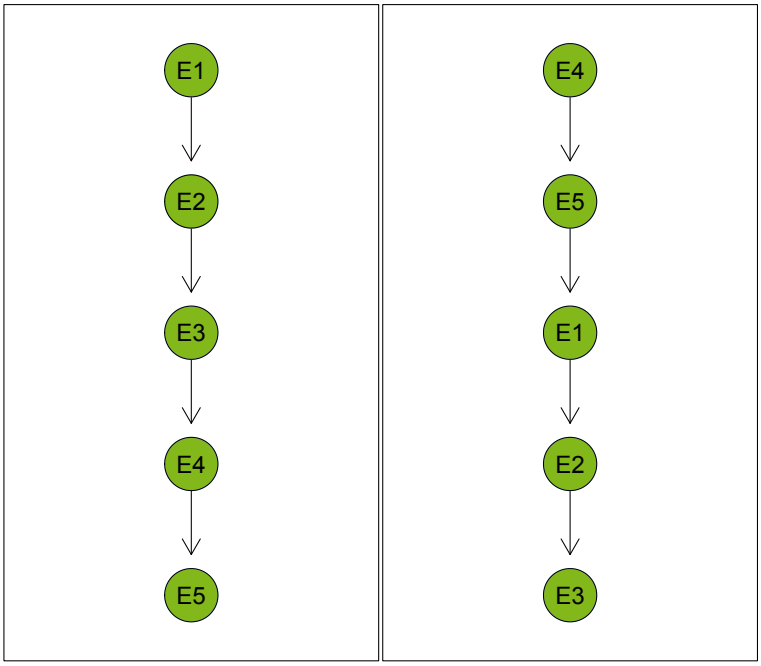



Abbildung 3.1: Zwei mit der *graph edit distance* zu vergleichende Baummodelle

erste Variante immer vollständige Komponenten hinzufügen, auch wenn darin manche Kanten identisch zu welchen in vorangehenden Baumkomponenten sind. Bei der zweiten Variante würde einfach nur die Zahl der fehlenden Kanten ausschlaggebend sein, ohne dass bekannt ist, in wie viele Baumkomponenten sie sich aufteilen. Es muss dabei auch nicht immer eine vollständige Baumkomponente entstehen, d.h. eine mit $n$ Kanten. Die erste Variante legt also Wert auf die vollständige Rekonstruktion der Baumkomponenten des wahren Modells, während bei der zweiten Variante die Information über die Reihenfolgebeziehung von Ereignispaaren im Vordergrund steht.

### 3.2.3 Aufbau der Simulation

Um nun den Vergleich der Progressionsmodelle durchzuführen, wird wie folgt vorgegangen:

1. Wähle ein zugrunde liegendes wahres Baummodell $T$ mit $n$ Ereignissen.
2. Ziehe $N$ Beobachtungen aus $T$ und erhalte die Datenmatrix $X \in \mathbb{B}^{N \times n}$.
3. Wähle eine Modellklasse und passe ein Progressionsmodell $T^*$ an $X$ an.
4. Vergleiche $T$ und $T^*$.

Aufgrund geeigneter zur Verfügung stehender Software beschränkt sich der Vergleich der Progressionsmodelle auf folgende Auswahl. In Klammern sind dabei jeweils die Abkürzungen genannt, mit denen die Modelle im Folgenden bezeichnet werden.

- onkogenetische Bäume (OT)
- onkogenetische Baum-Mischungs-Modelle (OTM)
  - mit 2 Komponenten
    - mit gleichen Kantengewichten bei der Sternkomponente (OTM2e)
    - mit ungleichen Kantengewichten bei der Sternkomponente (OTM2ne)
  - mit 3 Komponenten
    - mit gleichen Kantengewichten bei der Sternkomponente (OTM3e)
    - mit ungleichen Kantengewichten bei der Sternkomponente (OTM3ne)
- Distanzbäume (DIST)
- verbindende Bayes Netze (CBN)

Für die Modellklasse der onkogenetischen Bäume und Baum-Mischungs-Modelle gibt es ein R-Paket (R Core Team, 2017) `Rtreemix` (Bogojeska et al., 2008; Bogojeska, 2016). Ebenso gibt es ein R-Paket `oncomodel` (von Heydebreck, 2003) für Distanzbäume. Zum Schätzen der verbindenden Bayes Netze wird ein Perl-Skript verwendet (Beerenwinkel et al., 2007).

Um einen Überblick zu bekommen, wie gut die einzelnen Modellklassen geeignet sind, wird als wahres Modell (Schritt 1) jeweils eines aus jeder Klasse gewählt. Das bedeutet, dass es sieben verschiedene wahre Modelle gibt. Anschließend werden alle sieben Modellklassen verwendet, um an Daten aus diesen sieben wahren Modellen wiederum ein Modell anzupassen (Schritt 3). Mit Hilfe der Abstandsmaße soll dann untersucht werden, welche Modellklasse sich am besten eignet, um Daten aus einem bestimmten Modell anzupassen (Schritt 4).

Im Folgenden wird nun noch beschrieben, wie Daten aus einem gegebenen Modell gezogen werden (Schritt 2). Wie in Abschnitt 2.3 schon erwähnt, sind die Ereignisse ‘Kante $e$ ist im Baum enthalten’ unabhängig voneinander. Das zugrunde liegende Modell mit diesen entsprechenden Kanten ist fest vorgegeben und unveränderlich. Um jedoch eine Beobachtung bzw. einen Datenpunkt aus dem Modell zu generieren, werden zunächst alle Kanten des Baumes entfernt und nach bestimmten Kriterien wieder eingefügt. Die Wahrscheinlichkeit, dass eine Kante im Baum enthalten ist, wird über die Kantengewichte angegeben, d.h. über die bedingten Wahrscheinlichkeiten. Bei einem onkogenetischen Baum würde man wie folgt vorgehen: Um eine Beobachtung aus dem Modell zu generieren, wird zunächst unabhängig für jede Kante eine Zufallszahl zwischen 0 und 1 gezogen. Ist die Zufallszahl einer Kante kleiner oder gleich ihrem Kantengewicht, d.h. ihrer bedingten Wahrscheinlichkeit, wird angenommen, dass die Kante für die aktuelle Beobachtung im Modell enthalten ist, d.h. beschritten werden kann. Man zeichnet also in den Baum, an dem man die Beobachtung ablesen möchte, alle Kanten ein, die nach obigem Vorgehen enthalten sind. Es kann dabei vorkommen, dass manche Knoten von der Wurzel nicht über einen Pfad erreicht werden können. Diese Knoten bleiben unberücksichtigt. Als aufgetretene Ereignisse werden nur die Knoten verwendet, die man über einen Pfad ausgehend von der Wurzel erreichen kann. Man kann die Beobachtung entweder als Menge der eingetretenen Ereignisse angeben, oder als Bitstring der Länge $n$, wobei $n$ die Anzahl der möglichen Ereignisse ist und die eingetretenen Ereignisse mit 1 und

die restlichen mit 0 codiert werden. Diese Vorgehensweise wird nun so oft wiederholt, wie man Beobachtungen in seinem Datensatz haben möchte. Üblicherweise wird ein Datensatz mit $N$ Beobachtungen als $(N \times n)$ Matrix dargestellt, wobei die Zeilen die $N$ Patienten bzw. Beobachtungen und die Spalten die $n$ Ereignisse repräsentieren. Dabei wird die Codierung über den Bitstring verwendet.

Für onkogenetische Baum-Mischungs-Modelle und verbindende Bayes Netze sieht das Verfahren ganz ähnlich aus. Bei Baum-Mischungs-Modellen muss zuvor noch eine weitere Zufallszahl zwischen 0 und 1 gezogen werden, die bestimmt, aus welcher Baumkomponente die Beobachtung generiert werden soll. Das weitere Vorgehen ist dann analog zu onkogenetischen Bäumen. Bei verbindenden Bayes Netzen stehen die bedingten Wahrscheinlichkeiten nicht an den Kanten, sondern an den Knoten. Man kann mit ihnen aber wie mit Kantenwahrscheinlichkeiten arbeiten. Für jede Wahrscheinlichkeit wird eine Zufallszahl gezogen, die darüber entscheidet, ob das Ereignis theoretisch eintreten kann oder nicht. Ist dies bekannt, wird überprüft, ob das Ereignis auch praktisch eintritt, d.h. ob alle notwendigen Vorgänger auch eingetreten sind. Im Prinzip kontrolliert man also wiederum, ob man von der Wurzel das entsprechende Ereignis über einen Pfad erreichen kann. Nur muss bei einem Ereignis, dessen Zufallszahl kleiner oder gleich der zugehörigen bedingten Wahrscheinlichkeit ist, nicht nur eine, sondern ggf. mehrere Kanten eingetragen werden, nämlich alle, die zu diesem Knoten hinführen.

Wie oben schon erwähnt, wird im ersten Schritt der Simulationsstudie ein zugrunde liegendes wahres Modell aus jeder Modellklasse ausgewählt. Diese ausgewählten Modelle basieren teilweise auf Beispielen, die in der entsprechenden Literatur angegeben wurden. Für die sieben betrachteten Modellklassen sind die Beispielbäume und damit die wahren Modelle im Anhang A.1.1 dargestellt. Die onkogenetischen Bäume und Baum-Mischungs-Modelle besitzen jeweils fünf Ereignisse, während die verbindenden Bayes Netze vier und die Distanzbäume drei Ereignisse aufweisen. Bei den Baum-Mischungs-Modellen wird zwischen gleichen und ungleichen Kantengewichten bei der Sternkomponente unterschieden, um zu überprüfen, ob dies einen Unterschied in der Güte der Anpassung mit sich bringt.

Ein weiterer festzulegender Parameter ist die Anzahl $N$ an Beobachtungen in einem Datensatz. Gibt es nur wenige Beobachtungen, sind nicht alle Ereigniskombinationen, die theoretisch eintreten können, auch zu beobachten. Der Datensatz spiegelt also nur einen Teil der vorhandenen Struktur wider, während dies bei vielen Beobachtungen eher unwahrscheinlich ist. Um also zu überprüfen, wie gut die einzelnen Modellklassen mit einer unterschiedlichen Zahl an Beobachtungen umgehen können, bzw. wie viele Beobachtungen notwendig sind, um das wahre Modell gut zu lernen, wird die Zahl der Beobachtungen $N \in \{50, 100, 200, 500, 1000\}$ in dieser Simulationsstudie variiert.

Da der erzeugte Datensatz im zweiten Schritt der Simulationsstudie abhängig von den dabei verwendeten Zufallszahlen ist, wird pro wahrem Modell nicht nur ein Datensatz generiert, sondern $M = 100$ Datensätze. Somit werden aus jedem der sieben wahren Modelle und für jeden der fünf Werte für die Beobachtungsanzahl jeweils 100 Datensätze generiert, an die jede Modellklasse ein Progressionsmodell anpassen soll.

### 3.2.4 Auswertung der Simulationsstudie

Neben der Beschreibung des Aufbaus der Simulationsstudie soll nun abschließend auf einzelne Analyseschritte in der Auswertung eingegangen werden. Im vierten und letzten Schritt der Simulationsstudie sollen das wahre und das angepasste Modell miteinander verglichen werden. Ziel ist es, eine Reihenfolge der sieben Modellklassen für die Güte der Anpassung angeben zu können.

Ein Vergleich zwischen wahrem und angepasstem Modell erfolgt anhand der in Abschnitt 3.2.2 vorgestellten Abstandsmaße und Kennzahlen. Für diese Simulationsstudie beschränken wir uns auf den Abstand zwischen den Wahrscheinlichkeitsverteilungen, der mit Hilfe der $L_1$-, $L_2$- bzw. Cosinus-Distanz angegeben werden kann, und auf einen graphentopologischen Abstand, der mit der *graph edit distance* (Variante 1) gemessen wird.

Da für jede Parameterkombination 100 Datensätze generiert werden, ergeben sich auch 100 Werte für den Abstand zwischen einem wahren und einem angepassten Modell für eine feste Beobachtungsanzahl. Diese lassen sich mit Hilfe von Boxplots veranschaulichen und gegenüberstellen, wobei kleine Distanzen jeweils für eine gute Anpassung der Modellklasse sprechen.

Für das Aufstellen eines Rankings der Modellklassen bezüglich der Fähigkeit, das wahre Modell wiederzugeben, reichen Boxplots jedoch nicht aus. Eine Modellklasse soll gegenüber einer anderen als besser eingestuft werden, wenn sie stochastisch kleinere Abstände zum wahren Modell aufweist. Dies kann mit Hilfe des in Abschnitt 3.1.1 beschriebenen Wilcoxon-Mann-Whitney Tests überprüft werden. Um für eine gegebene wahre Modellklasse und eine feste Anzahl an Beobachtungen eine Reihenfolge dieser paarweisen Testergebnisse abzuleiten, wird gezählt, wie oft andere Modellklassen signifikant besser bzw. schlechter abgeschnitten haben. Im Fall von Bindungen werden Durchschnittsränge verwendet. Ein detailliertes Beispiel dazu wird in Abschnitt 3.3.1 vorgestellt.

Die Wilcoxon-Mann-Whitney Tests werden für jede Kombination aus wahrem und angepasstem Modell, jedes Abstandsmaß und jede Beobachtungsanzahl durchgeführt. Das führt zu einer sehr großen Menge an Rankings. Um diese Ergebnisse über die fünf verschiedenen Werte für $N$ zusammenzufassen, werden verschiedene Ansätze verfolgt.

Als erstes können die Ränge für die unterschiedlichen Beobachtungsanzahlen einfach gemittelt und dann wieder als Rangordnung angegeben werden. Eine zweite Möglichkeit ist, die $p$-Werte des Wilcoxon-Mann-Whitney Tests über die inverse Normalmethode (Hartung et al., 2008) zusammenzufassen. Dazu werden die fünf verschiedenen $p$-Werte $p_1, ..., p_5$ der Tests für jedes $N$ und die Inverse $\Phi^{-1}$ der Verteilungsfunktion der Standardnormalverteilung verwendet, um

$$Z = \sum_{i=1}^{5} \frac{\Phi^{-1}(p_i)}{\sqrt{5}} \tag{3.13}$$

zu berechnen. Unter der Nullhypothese der Gleichheit der Verteilungen der Abstandsmaße ist $Z$ approximativ[1] standardnormalverteilt und kann somit mit den Quantilen der Standard-

[1] Sind die $p_i$ Re[0,1]-verteilt, gilt die exakte Verteilung.

normalverteilung verglichen werden, um für alle fünf Tests eine einzige Testentscheidung zu erhalten.

Weiterhin (als dritte und vierte Variante) können auch direkt die mittleren bzw. medianen Abstände selbst verwendet werden, um ein Ranking unabhängig von $N$ zu erhalten, indem diese über alle Werte von $N$ gemittelt werden. Die Reihenfolge ergibt sich dann aus den Rangplätzen der geordneten Mittelwerte. Hierbei ist allerdings zu beachten, dass bei nur wenigen Beobachtungen die Modellanpassung wahrscheinlich schlechter ausfallen wird als bei vielen Beobachtungen. Die Distanzwerte für kleine $N$ sind somit größer und können den Mittelwert damit stärker beeinflussen. Eine Standardisierung der Daten kann jedoch größere Nachteile haben. Manchmal, besonders im Fall der *graph edit distance*, haben alle angepassten Modelle den gleichen Abstand zum wahren Modell. Somit sind Standardabweichung und MAD beide 0 und eine Standardisierung ist unmöglich. Ein weiterer nachteiliger Aspekt ist, dass sich alle Werte ändern würden, wenn eine zusätzliche Modellklasse in den Vergleich aufgenommen werden soll. Daher werden im Folgenden die ursprünglichen Distanzwerte betrachtet und über die $N$ gemittelt. Für die hier durchgeführte Simulationsstudie sind die Ergebnisse mit und ohne Standardisierung ohnehin sehr ähnlich.

Für den Fall der Abstände zwischen den Wahrscheinlichkeitsverteilungen bietet es sich zusätzlich an, die empirische Verteilungsfunktion der Daten zu betrachten. Diese kann als Maßstab für eine gute Modellanpassung herangezogen werden. Eine Modellklasse kann dabei als gut angenommen werden, wenn die Abstände zum wahren Modell kleiner sind als der Abstand zwischen wahrem Modell und empirischer Verteilungsfunktion. Man berechnet den Quotienten aus mittleren bzw. medianen Distanzen zwischen angepasstem und wahrem Modell und den Abständen zwischen der empirischen Verteilungsfunktion und dem wahren Modell. Um unabhängig vom Zähler zu sein, wird der natürliche Logarithmus dieses Quotienten betrachtet. Anschließend können die so berechneten Werte wieder über die verschiedenen $N$ gemittelt und daraus eine Reihenfolge bestimmt werden.

Insgesamt stehen sechs (für die *graph edit distance* nur vier) verschiedene Ansätze zur Verfügung, ein Ranking unabhängig von der Anzahl an Beobachtungen zu erhalten:

- Mitteln der Ränge, die sich aus den Tests ergeben haben
- Anwenden der inversen Normalmethode, um $p$-Werte zusammenzufassen
- Mitteln der mittleren bzw. medianen Abstände
- Vergleich der mittleren bzw. medianen Abstände mit der empirischen Verteilungsfunktion

Weichen die Ergebnisse dieser sechs (bzw. vier) Methoden nicht stark voneinander ab, deutet das darauf hin, dass man eine Reihenfolge unabhängig von der Beobachtungsanzahl angeben kann.

Für die Darstellung der Ergebnisse sollten neben der resultierenden Reihenfolge natürlich auch die mittleren bzw. medianen Abstände sowie der Vergleich zur empirischen Verteilungsfunktion angeben werden. Anhand dessen erhält man einen Eindruck über den Unterschied zwischen verschiedenen Rangplätzen.

## 3.3 Simulationsergebnisse

Im Folgenden werden die Ergebnisse der Simulationsstudie vorgestellt. Da zwei unterschiedliche Arten von Abstandsmaßen zwischen Baummodellen verwendet werden, erfolgt die Darstellung der Ergebnisse in zwei Abschnitten. In Abschnitt 3.3.1 werden die einzelnen Analyseschritte für die Auswertung basierend auf der Wahrscheinlichkeitsverteilung detailliert beschrieben, während in Abschnitt 3.3.2 für die *graph edit distance* nur die Ergebnisse erläutert werden.

### 3.3.1 Abstände basierend auf der Wahrscheinlichkeitsverteilung

Anhand folgender Situation soll die Auswertung der Simulationsergebnisse veranschaulicht werden: Das wahre Modell ist ein onkogenetisches Baum-Mischungs-Modell mit zwei Komponenten und gleichen Kantengewichten bei der Sternkomponente (OTM2e), als Abstandsmaß wird die $L_1$-Distanz der Wahrscheinlichkeitsverteilungen verwendet und die Anzahl der gezogenen Beobachtungen aus dem wahren Modell beträgt $N = 200$.

Einen ersten Eindruck über das Abschneiden der unterschiedlichen Modellklassen liefern Boxplots der berechneten $L_1$-Distanzen. Zum Vergleich wird hier auch die empirische Verteilungsfunktion (edf) als Maßstab mit aufgenommen. Für das oben genannte Beispiel sind die Ergebnisse in Abbildung 3.2 dargestellt.

Das Anpassen eines OTM2ne führt zu den kleinsten Abständen zum wahren Modell. Auch die Modellklassen OTM3e und OTM3ne schneiden besser ab als die wahre Modellklasse OTM2e und sogar die empirische Verteilungsfunktion weist etwas kleinere Abstände auf. Das onkogenetische Baummodell (OT) hat die größten $L_1$-Abstände. Dies ist nicht verwunderlich, da diese Modellklasse die wenigsten Möglichkeiten hat, bestimmte Strukturen in den Daten abzubilden. Da mit dem OTM2e als wahrem Modell 35% der Daten durch die Sternkomponente generiert wird, wird von einer Modellklasse, die keine unabhängigen Ereignisse modellieren kann, nicht erwartet, gut abzuschneiden.

Nun sollen die Modellklassen mit Hilfe des Wilcoxon-Mann-Whitney Tests in eine Reihenfolge gebracht werden. In Tabelle 3.1 sind die Ergebnisse für das ausgewählte Beispiel und die resultierende Reihenfolge dargestellt. Die Werte der vorletzten Spalte geben dabei an, wie oft eine Modellklasse signifikant kleinere Distanzen im Vergleich zu den anderen erzielen konnte. Die Werte der letzten Zeile besagen, wie oft keine signifikant größeren Abstände erreicht wurden. Ein Mitteln dieser Werte führt zur endgültigen Rangliste der Modellklassen, die in der letzten Spalte angegeben ist.

In diesem Beispiel gibt es eine eindeutige Reihenfolge in dem Sinne, dass die Überlegenheit einer Modellklasse zur nächst besten im Sinne von kleineren $L_1$-Distanzen immer statistisch signifikant ist. Diese Reihenfolge für das Anpassen des wahren OTM2e lautet wie folgt:

OTM2ne (11), OTM3ne (16.79), OTM3e (12.85), edf, OTM2e (7), CBN (5), DIST (5), OT (5).

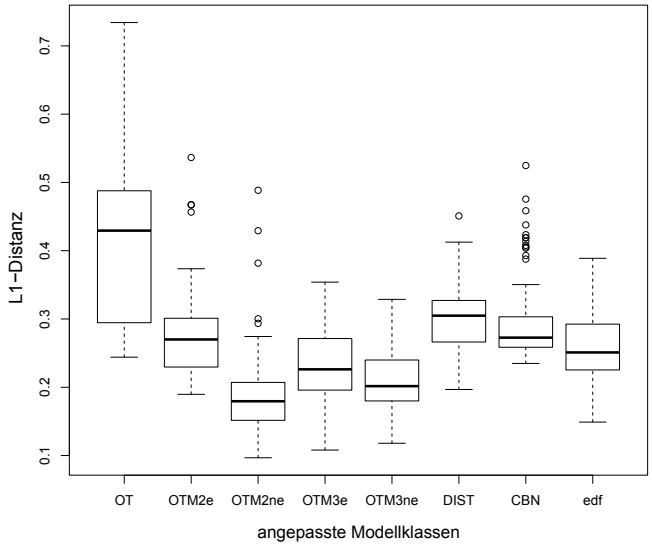



Abbildung 3.2: Boxplots der $L_1$-Distanzen zum wahren OTM2e Modell (Stichprobenumfang $N = 200$)

.

Tabelle 3.1: Ergebnisse des Wilcoxon-Mann-Whitney Tests (T: $p$-Wert $< 0.05$, F: $p$-Wert $\geq 0.05$) und die resultierende Reihenfolge vom besten (1) zum schlechtesten (8) Modell beim Anpassen eines OTM2e, unter Verwendung der $L_1$-Distanz bei $N = 200$ Beobachtungen.

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | # T's | Rangliste |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 OT | - | F | F | F | F | F | F | F | 0 | 8 |
| 2 OTM2e | T | - | F | F | F | T | T | F | 3 | 5 |
| 3 OTM2ne | T | T | - | T | T | T | T | T | 7 | 1 |
| 4 OTM3e | T | T | F | - | F | T | T | T | 5 | 3 |
| 5 OTM3ne | T | T | F | T | - | T | T | T | 6 | 2 |
| 6 DIST | T | F | F | F | F | - | F | F | 1 | 7 |
| 7 CBN | T | F | F | F | F | T | - | F | 2 | 6 |
| 8 edf | T | T | F | F | F | T | T | - | 4 | 4 |
| # F's | 0 | 3 | 7 | 5 | 6 | 1 | 2 | 4 | | |

In Klammern ist dabei die mittlere Anzahl an Parametern über die $M = 100$ angepassten Modelle für jede Parametereinstellung angegeben. Das wahre OTM2e hat sieben Parameter. Aufgrund des eindeutigen Rankings enthält Tabelle 3.1 redundante Teile. Für andere Fälle ist diese detaillierte Information jedoch notwendig, da die Unterschiede zwischen zwei Modellklassen nicht immer signifikant sein müssen (siehe Tabelle A.1 im Anhang A.1.2).

Die Ergebnisse der Wilcoxon-Mann-Whitney Tests lassen sich auf folgende Weise veranschaulichen. Es sollen Graphen verwendet werden, deren Knoten die anpassenden Modellklassen darstellen und deren Kanten die statistische Signifikanz der Wilcoxon-Mann-Whitney Tests widerspiegeln. Eine Kante von Modellklasse A zu Modellklasse B bedeutet dann, dass Modell A zu signifikant kleineren Abständen zum wahren Modell führt als Modell B. Als Folge hat

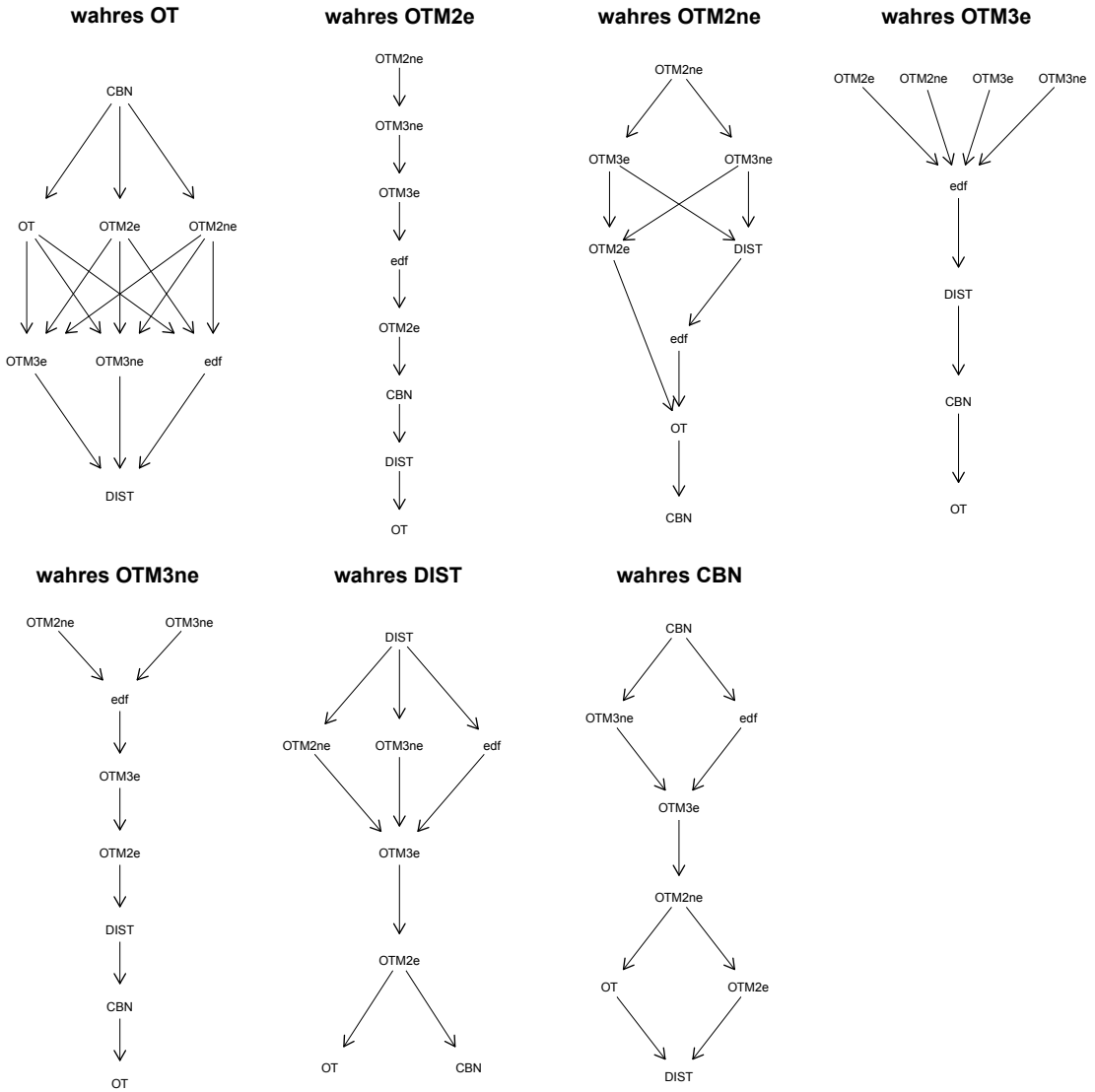



Abbildung 3.3: Grafische Darstellung der Ergebnisse des Wilcoxon-Mann-Whitney Tests für $N = 200$, die $L_1$-Distanz und alle sieben wahren Modelle. Die Pfeile zeigen in Richtung eines Modells mit signifikant größeren Abständen zum wahren Modell.

Modell A auch signifikant kleinere Abstände zu allen Modellklassen, die Nachfolger von A sind. Für die mittlere Anzahl an Beobachtungen ($N = 200$) und für alle sieben wahren Modelle sind diese Graphen in Abbildung 3.3 dargestellt. Es ist zu erkennen, dass eine eindeutige Reihenfolge nur für das wahre OTM2e existiert (zweiter Graph der ersten Reihe). Eine detaillierte Auswertung und Interpretation dieser Reihenfolgen für jede wahre Modellklasse erfolgt später.

Bisher sind die Simulationsergebnisse nur für eine feste Anzahl an Beobachtungen betrachtet worden. Im Folgenden soll daher untersucht werden, wie sich die mittleren $L_1$-Distanzen beim Anpassen eines Modells in Abhängigkeit von unterschiedlichen Beobachtungsanzahlen verhalten. In Abbildung 3.4 ist dieser Verlauf für das Anpassen eines wahren OTM2e grafisch dargestellt. In einigen wenigen der 100 Simulationsdurchläufe pro Parametereinstel-

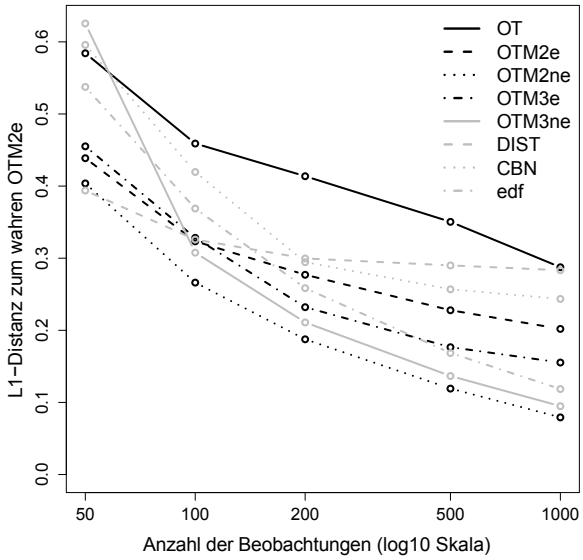



Abbildung 3.4: Mittelwerte der $L_1$-Distanzen beim Anpassen eines OTM2e mit einer unterschiedlichen Anzahl an Beobachtungen.

lung konnte die Modellanpassung aufgrund zu weniger Beobachtungen nicht durchgeführt werden. Dies war der Fall beim Anpassen eines OTM3ne an das wahre OTM2e mit $N = 50$ Beobachtungen bzw. an das wahre OTM2ne mit $N = 50$ und $N = 100$ Beobachtungen. In diesen Fällen wird als Bestrafung der Abstand zwischen wahrem und angepasstem Modell auf einen 5% höheren Wert gesetzt als der größte auftretende Abstand bei der entsprechenden Parametereinstellung.

Betrachtet man nun den Verlauf der Mittelwerte über die Anzahl an Beobachtungen hinweg, fällt sofort auf, dass die Abstände der Wahrscheinlichkeitsverteilungen zwischen wahrem und angepasstem Modell mit wachsender Beobachtungsanzahl kleiner werden. Dies ist unabhängig von der anpassenden Modellklasse und der Güte der Anpassung zu beobachten. Die medianen Abstände zeigen das gleiche Verhalten. Die empirische Verteilungsfunktion ist für viele Beobachtungen eine gute Approximation an die wahre Verteilung, bei nur wenigen vorliegenden Beobachtungen sind die Ergebnisse aber vergleichsweise schlecht. Da in realen Datensätzen üblicherweise eher wenige Beobachtungen vorliegen, wird hier eine geeignete Modellklasse benötigt, die gute Ergebnisse liefert.

Ein weiterer interessanter Aspekt, der aus dieser Grafik hervorgeht, ist, dass nicht die OTM2e als wahre Modellklasse die besten Ergebnisse bei der Modellanpassung liefern. Das OTM2ne weist über alle Beobachtungsanzahlen hinweg den kleinsten Abstand auf, gefolgt von den Baum-Mischungs-Modellen mit drei Komponenten (eine Ausnahme stellt hier $N = 50$ dar). Das bedeutet, dass hier eine kleine Überanpassung vorteilhaft ist.

Unabhängig von der Anzahl an Beobachtungen soll nun eine Rangfolge der Modellklassen angegeben werden. In Abschnitt 3.2.4 wurden dazu bereits verschiedene Varianten beschrieben. Die Ergebnisse dieser sechs Ansätze sind für die Anpassung eines wahren OTM2e in Tabelle 3.2 dargestellt.

Tabelle 3.2: Rangfolge der anpassenden Modellklassen unter Verwendung verschiedener Ansätze, die Ergebnisse für unterschiedliche Beobachtungsanzahlen zusammenzufassen. Das wahre Modell ist dabei das OTM2e und als Abstandsmaß wird die $L_1$-Distanz verwendet.

| Methode | Test Ergebnisse | $p$-Werte | Mittelwert | Median | mittlere edf | mediane edf | global |
|---|---|---|---|---|---|---|---|
| OT | 8 | 8 | 8 | 8 | 8 | 8 | 8 |
| OTM2e | 4.5 | 4 | 5 | 4 | 5 | 5 | 5 |
| OTM2ne | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| OTM3e | 3 | 2.5 | 2 | 2 | 3 | 3 | 2 |
| OTM3ne | 2 | 5.5 | 3 | 3 | 2 | 2 | 3 |
| DIST | 6 | 7 | 6 | 6 | 6 | 6 | 6 |
| CBN | 7 | 5.5 | 7 | 7 | 7 | 7 | 7 |
| edf | 4.5 | 2.5 | 4 | 5 | 4 | 4 | 4 |

Es ist zu erkennen, dass sich die Ergebnisse der sechs Methoden nicht nennenswert voneinander unterscheiden. Nur selten gibt es eine etwas größere Diskrepanz. Das deutet darauf hin, dass eine globale Reihenfolge für die Anpassungsgüte der verschiedenen Modellklassen unabhängig von der Anzahl an Beobachtungen angegeben werden kann. Um den Unterschied zwischen einzelnen Rangplätzen besser beurteilen zu können, können zusätzlich die mittleren und medianen Distanzen der 100 Simulationsdurchläufe sowie ein Vergleich zur empirischen Verteilungsfunktion angegeben werden. Für die hier betrachtete Anpassung an ein wahres OTM2e finden sich diese Angaben in der zweiten Spalte von Tabelle 3.3.

Insgesamt sind in den Tabellen 3.3 bis 3.5 die Ergebnisse der Modellanpassung für alle sieben wahren Modelle und alle drei verschiedenen Abstandsmaße dargestellt. Da die Ergebnisse für die drei Distanzmaße aber sehr ähnlich sind, wird auf eine detaillierte Diskussion für jedes einzelne Maß verzichtet. Stattdessen wird in Tabelle 3.6 eine globale Reihenfolge für jedes wahre Modell angegeben, die sowohl unabhängig von der Beobachtungsanzahl als auch unabhängig vom gewählten Distanzmaß für den Abstand zur wahren Wahrscheinlichkeitsverteilung ist.

Wenn das wahre Modell ein onkogenetischer Baum (OT) ist, liefern die CBNs die besten Ergebnisse im Sinne des kleinsten Abstandes zur wahren Wahrscheinlichkeitsverteilung. Das ist nicht überraschend, da die CBNs als direkte Verallgemeinerung von OTs alle benötigten Strukturen modellieren können. In einem CBN können zwar auch Knoten dargestellt werden, die mehr als einen Elternknoten besitzen, falls diese Struktur jedoch nicht in den Daten vorkommt, kann ein CBN genauso aussehen wie ein OT. CBNs sind besonders bei wenigen Beobachtungen ($N = 50, 100, 200$) den OTs überlegen. In diesen Situationen schätzt das onkogenetische Baummodell oft eine falsche Baumstruktur. Das könnte darin begründet sein, dass der Algorithmus zum Schätzen eines Baumes kein Maximum-Likelihood Verfahren ist wie bei den CBNs. Bei vielen Beobachtungen ($N = 500, 1000$) schneiden CBNs und OTs gleich gut ab.

Beim Anpassen eines onkogenetischen Baumes schneiden OTs und Baum-Mischungs-Modelle mit 2 Komponenten ähnlich gut ab. Die Sternkomponenten bei den Mischungs-Modellen

Tabelle 3.3: Rangfolge der Modellklassen bei Verwendung der $L_1$-Distanz unabhängig von der Anzahl an Beobachtungen und zusammengefasst über die sechs verschiedenen Ranking-Methoden. Zusätzlich sind die mittleren und medianen Distanzen zum wahren Modell angegeben sowie der Vergleich zur empirischen Verteilungsfunktion.

| angepasstes | | wahres Modell | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Modell | | OT | OTM2e | OTM2ne | OTM3e | OTM3ne | DIST | CBN |
| OT | Ranking | 2 | 8 | 7 | 8 | 8 | 7 | 6 |
| | Mittelwert | 0.1650 | 0.4190 | 0.4316 | 0.5452 | 0.6220 | 0.5859 | 0.3039 |
| | Median | 0.1482 | 0.4031 | 0.4286 | 0.5355 | 0.6205 | 0.5541 | 0.3001 |
| | mittlere edf | 0.0761 | -0.4779 | -0.4826 | -0.8762 | -0.9822 | -0.4826 | -1.0701 |
| | mediane edf | 0.1481 | -0.4374 | -0.4828 | -0.8696 | -0.9908 | -0.4378 | -1.0903 |
| OTM2e | Ranking | 4 | 5 | 3 | 2.5 | 5 | 6 | 7 |
| | Mittelwert | 0.1649 | 0.2938 | 0.2877 | 0.2164 | 0.2959 | 0.4300 | 0.3056 |
| | Median | 0.1486 | 0.2843 | 0.2796 | 0.2113 | 0.2916 | 0.4269 | 0.3034 |
| | mittlere edf | 0.0703 | -0.1137 | -0.0520 | 0.0940 | -0.2222 | -0.1720 | -1.0772 |
| | mediane edf | 0.1376 | -0.0893 | -0.0316 | 0.1059 | -0.2195 | -0.1750 | -1.1026 |
| OTM2ne | Ranking | 3 | 1 | 1 | 1 | 2 | 3 | 5 |
| | Mittelwert | 0.1551 | 0.2112 | 0.2128 | 0.2151 | 0.2426 | 0.3692 | 0.2404 |
| | Median | 0.1518 | 0.2050 | 0.2099 | 0.2123 | 0.2390 | 0.3582 | 0.2380 |
| | mittlere edf | 0.0990 | 0.3363 | 0.3691 | 0.1401 | 0.0140 | -0.0168 | -0.8381 |
| | mediane edf | 0.0915 | 0.3538 | 0.3751 | 0.1421 | 0.0166 | 0.0023 | -0.8612 |
| OTM3e | Ranking | 7 | 2 | 2 | 4 | 4 | 5 | 4 |
| | Mittelwert | 0.1787 | 0.2695 | 0.2468 | 0.2238 | 0.2733 | 0.4022 | 0.1331 |
| | Median | 0.1674 | 0.2620 | 0.2423 | 0.2221 | 0.2725 | 0.3995 | 0.1284 |
| | mittlere edf | -0.0435 | 0.0147 | 0.1771 | 0.0805 | -0.1210 | -0.1043 | -0.1110 |
| | mediane edf | -0.0144 | 0.0260 | 0.2011 | 0.0746 | -0.1288 | -0.1095 | -0.1016 |
| OTM3ne | Ranking | 5 | 3 | 6 | 2.5 | 1 | 3 | 2 |
| | Mittelwert | 0.1667 | 0.2752 | 0.3574 | 0.2252 | 0.2404 | 0.3636 | 0.1195 |
| | Median | 0.1631 | 0.2663 | 0.3427 | 0.2223 | 0.2370 | 0.3592 | 0.1162 |
| | mittlere edf | 0.0111 | 0.1332 | -0.0018 | 0.1071 | 0.0509 | -0.0003 | 0.0107 |
| | mediane edf | 0.0019 | 0.1461 | 0.0246 | 0.1075 | 0.0571 | 0.0001 | 0.0090 |
| DIST | Ranking | 8 | 6 | 4.5 | 6 | 6 | 1 | 8 |
| | Mittelwert | 0.3017 | 0.3186 | 0.2848 | 0.4168 | 0.4512 | 0.1770 | 0.3435 |
| | Median | 0.3215 | 0.3205 | 0.2849 | 0.4153 | 0.4515 | 0.1664 | 0.3282 |
| | mittlere edf | -0.7306 | -0.2252 | -0.0565 | -0.6061 | -0.6613 | 0.8385 | -1.1966 |
| | mediane edf | -0.8076 | -0.2417 | -0.0640 | -0.6139 | -0.6728 | 0.8960 | -1.1775 |
| CBN | Ranking | 1 | 7 | 8 | 7 | 7 | 8 | 1 |
| | Mittelwert | 0.1253 | 0.3621 | 0.5809 | 0.5222 | 0.5762 | 0.5873 | 0.1021 |
| | Median | 0.1217 | 0.3355 | 0.5590 | 0.5153 | 0.5534 | 0.5557 | 0.0955 |
| | mittlere edf | 0.2874 | -0.3006 | -0.7810 | -0.8322 | -0.9054 | -0.4852 | 0.1664 |
| | mediane edf | 0.2834 | -0.2432 | -0.7487 | -0.8288 | -0.8771 | -0.4383 | 0.1956 |
| edf | Ranking | 6 | 4 | 4.5 | 5 | 3 | 3 | 3 |
| | Mittelwert | 0.1673 | 0.2904 | 0.3038 | 0.2593 | 0.2656 | 0.3635 | 0.1206 |
| | Median | 0.1633 | 0.2884 | 0.3031 | 0.2557 | 0.2635 | 0.3592 | 0.1170 |
| | mittlere edf | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| | mediane edf | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |

Tabelle 3.4: Rangfolge der Modellklassen bei Verwendung der $L_2$-Distanz unabhängig von der Anzahl an Beobachtungen und zusammengefasst über die sechs verschiedenen Ranking-Methoden. Zusätzlich sind die mittleren und medianen Distanzen zum wahren Modell angegeben sowie der Vergleich zur empirischen Verteilungsfunktion.

| angepasstes | | wahres Modell | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Modell | | OT | OTM2e | OTM2ne | OTM3e | OTM3ne | DIST | CBN |
| OT | Ranking | 4 | 8 | 7 | 8 | 8 | 8 | 7 |
| | Mittelwert | 0.0703 | 0.1229 | 0.1133 | 0.2166 | 0.2152 | 0.3235 | 0.1236 |
| | Median | 0.0589 | 0.1189 | 0.1110 | 0.2151 | 0.2205 | 0.3087 | 0.1213 |
| | mittlere edf | -0.0288 | -0.6411 | -0.5638 | -1.3000 | -1.2703 | -0.3613 | -0.9001 |
| | mediane edf | 0.1184 | -0.6179 | -0.5565 | -1.3314 | -1.3298 | -0.3302 | -0.9267 |
| OTM2e | Ranking | 2 | 5 | 4 | 3 | 5 | 6 | 6 |
| | Mittelwert | 0.0639 | 0.0756 | 0.0712 | 0.0644 | 0.0817 | 0.2669 | 0.1136 |
| | Median | 0.0584 | 0.0713 | 0.0688 | 0.0615 | 0.0795 | 0.2660 | 0.1114 |
| | mittlere edf | 0.0374 | -0.1276 | -0.0680 | -0.0125 | -0.2725 | -0.1714 | -0.8159 |
| | mediane edf | 0.1188 | -0.0841 | -0.0509 | -0.0069 | -0.2796 | -0.1882 | -0.8418 |
| OTM2ne | Ranking | 3 | 1 | 1 | 1 | 3 | 4 | 5 |
| | Mittelwert | 0.0611 | 0.0587 | 0.0554 | 0.0613 | 0.0675 | 0.2288 | 0.1101 |
| | Median | 0.0591 | 0.0550 | 0.0531 | 0.0580 | 0.0655 | 0.2219 | 0.1076 |
| | mittlere edf | 0.0653 | 0.2432 | 0.2930 | 0.0818 | -0.0382 | -0.0171 | -0.7878 |
| | mediane edf | 0.0732 | 0.2917 | 0.3247 | 0.0990 | -0.0431 | -0.0069 | -0.8108 |
| OTM3e | Ranking | 7 | 2 | 2 | 5 | 4 | 5 | 4 |
| | Mittelwert | 0.0678 | 0.0693 | 0.0626 | 0.0660 | 0.0762 | 0.2467 | 0.0604 |
| | Median | 0.0648 | 0.0684 | 0.0606 | 0.0642 | 0.0738 | 0.2436 | 0.0588 |
| | mittlere edf | -0.0424 | 0.0090 | 0.1370 | -0.0220 | -0.1843 | -0.0926 | -0.0507 |
| | mediane edf | -0.0168 | 0.0179 | 0.1611 | -0.0318 | -0.1910 | -0.1003 | -0.0725 |
| OTM3ne | Ranking | 5 | 3 | 5 | 2 | 1 | 3 | 2 |
| | Mittelwert | 0.0640 | 0.0755 | 0.0900 | 0.0642 | 0.0669 | 0.2250 | 0.0578 |
| | Median | 0.0630 | 0.0719 | 0.0861 | 0.0622 | 0.0638 | 0.2206 | 0.0555 |
| | mittlere edf | 0.0054 | 0.0581 | -0.0503 | 0.0431 | -0.0110 | -0.0003 | 0.0055 |
| | mediane edf | 0.0039 | 0.0879 | -0.0200 | 0.0350 | 0.0003 | -0.0009 | 0.0071 |
| DIST | Ranking | 8 | 7 | 6 | 6 | 7 | 1 | 8 |
| | Mittelwert | 0.2015 | 0.1169 | 0.0888 | 0.1556 | 0.1701 | 0.0909 | 0.1515 |
| | Median | 0.2013 | 0.1155 | 0.0886 | 0.1500 | 0.1650 | 0.0827 | 0.1384 |
| | mittlere edf | -1.2776 | -0.6091 | -0.3178 | -0.9602 | -1.0325 | 1.0234 | -1.1090 |
| | mediane edf | -1.1627 | -0.6055 | -0.3262 | -0.9654 | -1.0375 | 1.0989 | -1.0465 |
| CBN | Ranking | 1 | 6 | 8 | 7 | 6 | 7 | 1 |
| | Mittelwert | 0.0514 | 0.0951 | 0.1413 | 0.1589 | 0.1686 | 0.3041 | 0.0492 |
| | Median | 0.0495 | 0.0896 | 0.1365 | 0.1539 | 0.1608 | 0.2866 | 0.0459 |
| | mittlere edf | 0.2204 | -0.3351 | -0.7817 | -0.9861 | -1.0254 | -0.3002 | 0.1629 |
| | mediane edf | 0.2425 | -0.2916 | -0.7610 | -0.9930 | -1.0133 | -0.2575 | 0.1891 |
| edf | Ranking | 6 | 4 | 3 | 4 | 2 | 2 | 3 |
| | Mittelwert | 0.0640 | 0.0733 | 0.0734 | 0.0680 | 0.0690 | 0.2250 | 0.0581 |
| | Median | 0.0631 | 0.0730 | 0.0729 | 0.0652 | 0.0666 | 0.2204 | 0.0559 |
| | mittlere edf | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| | mediane edf | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |

Tabelle 3.5: Rangfolge der Modellklassen bei Verwendung der Cosinus-Distanz unabhängig von der Anzahl an Beobachtungen und zusammengefasst über die sechs verschiedenen Ranking-Methoden. Zusätzlich sind die mittleren und medianen Distanzen zum wahren Modell angegeben sowie der Vergleich zur empirischen Verteilungsfunktion.

| angepasstes | | wahres Modell | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Modell | | OT | OTM2e | OTM2ne | OTM3e | OTM3ne | DIST | CBN |
| OT | Ranking | 4 | 7 | 6 | 6 | 6 | 8 | 7 |
| | Mittelwert | 0.0112 | 0.0578 | 0.0571 | 0.0185 | 0.0355 | 0.0603 | 0.0246 |
| | Median | 0.0084 | 0.0501 | 0.0504 | 0.0164 | 0.0334 | 0.0549 | 0.0230 |
| | mittlere edf | -0.0434 | -0.7866 | -0.6578 | -0.8837 | -1.2231 | -0.8527 | -1.8184 |
| | mediane edf | 0.2483 | -0.7288 | -0.6213 | -0.9109 | -1.2965 | -0.7976 | -1.9802 |
| OTM2e | Ranking | 3 | 5 | 3 | 2 | 5 | 6 | 6 |
| | Mittelwert | 0.0102 | 0.0384 | 0.0391 | 0.0087 | 0.0175 | 0.0359 | 0.0207 |
| | Median | 0.0078 | 0.0326 | 0.0335 | 0.0070 | 0.0154 | 0.0352 | 0.0195 |
| | mittlere edf | 0.0214 | -0.3027 | -0.1807 | 0.1452 | -0.4201 | -0.3296 | -1.6475 |
| | mediane edf | 0.2476 | -0.2313 | -0.1385 | 0.1922 | -0.4530 | -0.3460 | -1.8163 |
| OTM2ne | Ranking | 2 | 1 | 1 | 1 | 2 | 4 | 5 |
| | Mittelwert | 0.0087 | 0.0268 | 0.0281 | 0.0095 | 0.0147 | 0.0280 | 0.0202 |
| | Median | 0.0075 | 0.0216 | 0.0232 | 0.0077 | 0.0121 | 0.0259 | 0.0179 |
| | mittlere edf | 0.1515 | 0.4472 | 0.5345 | 0.1846 | -0.0846 | -0.0742 | -1.6339 |
| | mediane edf | 0.1638 | 0.5695 | 0.6342 | 0.2548 | -0.0643 | -0.0309 | -1.7413 |
| OTM3e | Ranking | 7 | 2 | 2 | 3 | 4 | 5 | 4 |
| | Mittelwert | 0.0110 | 0.0338 | 0.0327 | 0.0097 | 0.0164 | 0.0303 | 0.0071 |
| | Median | 0.0098 | 0.0298 | 0.0296 | 0.0083 | 0.0148 | 0.0295 | 0.0059 |
| | mittlere edf | -0.1227 | -0.0029 | 0.2378 | 0.0972 | -0.2853 | -0.1580 | -0.1164 |
| | mediane edf | -0.0299 | 0.0305 | 0.2884 | 0.0979 | -0.3400 | -0.1720 | -0.1446 |
| OTM3ne | Ranking | 5 | 3 | 5 | 4 | 1 | 3 | 2 |
| | Mittelwert | 0.0095 | 0.0372 | 0.0514 | 0.0106 | 0.0150 | 0.0261 | 0.0066 |
| | Median | 0.0087 | 0.0320 | 0.0469 | 0.0089 | 0.0123 | 0.0251 | 0.0054 |
| | mittlere edf | 0.0149 | 0.1871 | 0.0897 | 0.0959 | -0.0333 | -0.0016 | 0.0103 |
| | mediane edf | 0.0319 | 0.2461 | 0.1285 | 0.1084 | 0.0258 | -0.0028 | 0.0217 |
| DIST | Ranking | 8 | 8 | 7 | 7 | 8 | 2 | 8 |
| | Mittelwert | 0.1433 | 0.0908 | 0.0587 | 0.0285 | 0.0530 | 0.0399 | 0.0401 |
| | Median | 0.0861 | 0.0835 | 0.0538 | 0.0169 | 0.0371 | 0.0213 | 0.0303 |
| | mittlere edf | -3.1368 | -1.3311 | -0.6951 | -1.0752 | -1.5161 | 0.0083 | -2.3262 |
| | mediane edf | -2.1729 | -1.3162 | -0.6641 | -0.8446 | -1.3740 | 0.5876 | -2.2064 |
| CBN | Ranking | 1 | 6 | 8 | 8 | 7 | 7 | 1 |
| | Mittelwert | 0.0060 | 0.0409 | 0.0770 | 0.0206 | 0.0386 | 0.0561 | 0.0048 |
| | Median | 0.0050 | 0.0363 | 0.0711 | 0.0175 | 0.0336 | 0.0475 | 0.0038 |
| | mittlere edf | 0.4791 | -0.3364 | -0.9472 | -0.9023 | -1.2707 | -0.7798 | 0.3215 |
| | mediane edf | 0.5338 | -0.3038 | -0.9392 | -0.9153 | -1.2645 | -0.6502 | 0.3570 |
| edf | Ranking | 6 | 4 | 4 | 5 | 3 | 1 | 3 |
| | Mittelwert | 0.0094 | 0.0386 | 0.0440 | 0.0123 | 0.0169 | 0.0261 | 0.0066 |
| | Median | 0.0088 | 0.0356 | 0.0407 | 0.0106 | 0.0143 | 0.0250 | 0.0055 |
| | mittlere edf | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| | mediane edf | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |

Tabelle 3.6: Globale Reihenfolge der Modellklassen unabhängig von der Beobachtungsanzahl und dem gewählten Distanzmaß für den Abstand zur wahren Wahrscheinlichkeitsverteilung.

| angepasstes | wahres Modell | | | | | | |
|---|---|---|---|---|---|---|---|
| Modell | OT | OTM2e | OTM2ne | OTM3e | OTM3ne | DIST | CBN |
| OT | 4 | 8 | 7 | 7.5 | 8 | 8 | 7 |
| OTM2e | 3 | 5 | 3 | 2 | 5 | 6 | 6 |
| OTM2ne | 2 | 1 | 1 | 1 | 2 | 4 | 5 |
| OTM3e | 7 | 2 | 2 | 4 | 4 | 5 | 4 |
| OTM3ne | 5 | 3 | 5 | 3 | 1 | 3 | 2 |
| DIST | 8 | 7 | 6 | 6 | 7 | 1 | 8 |
| CBN | 1 | 6 | 8 | 7.5 | 6 | 7 | 1 |
| edf | 6 | 4 | 4 | 5 | 3 | 2 | 3 |

weisen jeweils nur ein geringes Gewicht auf, da bei einem wahren OT keine unabhängigen Ereignisse modelliert werden müssen. Insgesamt können daher onkogenetische Baummodelle als anpassende Modellklasse vernachlässigt werden, da CBNs alle Fähigkeiten von OTs umschließen und zusätzlich weitere sinnvolle Eigenschaften bereit stellen. Bei allen anderen wahren Modellen liegen die OTs außerdem immer auf einem der letzten beiden Plätze.

Liegen Baum-Mischungs-Modelle als wahre Modelle zugrunde, schneiden alle Modelle, die nur aus einer Baumkomponente bestehen (OT, DIST und CBN), schlecht ab. Dies war vorherzusehen, da diese Modellklassen nicht die Fähigkeit besitzen, Ereignis-Pfade und Unabhängigkeit gleichzeitig abzubilden. Am besten schneiden hier noch die Distanzbäume ab, da dort keine strikte Reihenfolge der Ereignisse vorgegeben wird. Beim Anpassen eines wahren OTs, also nur einer Baumkomponente, schneiden Distanzbäume jedoch am schlechtesten ab. Sie sind sogar noch schlechter als ein Baum-Mischungs-Modell mit drei Komponenten. Nur falls das wahre Modell wirklich ein Distanzbaum ist, sollte diese Modellklasse zum Schätzen eines Baumes verwendet werden.

Überraschenderweise sind OTM2ne für fast alle Baum-Mischungs-Modelle am besten geeignet, unabhängig davon, ob das wahre Modell aus zwei oder drei Baumkomponenten besteht und die Sternkomponente gleiche oder ungleiche Kantengewichte aufweist. Um dieses Verhalten zu Verstehen, wird beispielhaft die Anpassung eines wahren OTM3e genauer betrachtet. Es stellt sich heraus, dass das beobachtete Ranking der Modellklassen nicht durch gute Anpassung des OTM2ne entstanden ist, sondern eher durch das schlechte Abschneiden des OTM3ne. Für letztere sind die beiden geschätzten echten Baumkomponenten oft identisch und damit redundant. Somit reproduzieren die OTM3ne nicht die zwei in den Daten vorliegenden genetischen Prozesse, obwohl die beiden zugehörigen Baumkomponenten des wahren Modells keine zu vernachlässigenden Gewichte aufweisen. Ob diese Beobachtung verallgemeinerbar ist, muss in einer größeren Simulationsstudie überprüft werden.

Weiterhin schneiden Baum-Mischungs-Modelle mit gleichen Kantengewichten bei der Sternkomponente fast immer schlechter ab als ihr Gegenstück mit ungleichen Kantengewichten.

Die einzige Ausnahme tritt hier bei der Anpassung an Baum-Mischungs-Modelle mit zwei Komponenten auf. Dies liegt jedoch am oben beschriebenen schlechten Abschneiden der OTM3ne. Somit sollte die größere Flexibilität der ungleichen Kantengewichte beim Anpassen eines Progressionsmodells bevorzugt werden.

Ein CBN kann am besten mit einem CBN selbst angepasst werden. Das war zu erwarten, da keine andere Modellklasse in der Lage ist, Situationen zu modellieren, in denen das Eintreten eines Ereignisses von mehr als einem anderen Ereignis abhängt. Am zweitbesten schneiden hier jedoch die Baum-Mischungs-Modelle mit drei Komponenten ab. Auch wenn hier keine Vereinigung von Kanten modelliert werden kann, scheint die zweite echte Baumkomponente diese fehlende Fähigkeit teilweise kompensieren zu können.

### 3.3.2 Abstände basierend auf der *graph edit distance*

Bisher wurde als Maß für die Güte einer Modellanpassung nur der Abstand zwischen wahrer und angepasster Wahrscheinlichkeitsverteilung herangezogen. Anhand der Wahrscheinlichkeitsverteilungen kann aber keine Aussage darüber getroffen werden, ob auch die Baumstruktur und damit die Krankheitspfade geeignet wiedergegeben werden. Daher wird nun eine vergleichbare Auswertung anhand der *graph edit distance* (*ged*) durchgeführt, die die grafische Struktur der wahren und angepassten Modelle vergleicht.

Für das Zusammenfassen der Ergebnisse über die unterschiedlichen Beobachtungsanzahlen werden die gleichen Verfahren wie zuvor benutzt, die auf dem Wilcoxon-Mann-Whitney Test und den mittleren und medianen Abständen beruhen. Das daraus resultierende Ranking der Modellklassen ist in Tabelle 3.7 dargestellt. Die Distanzbäume können in dieser Analyse nicht berücksichtigt werden, da sich ihre abweichende Baumstruktur nicht über die *ged* vergleichen lässt. Außerdem gibt es kein Vergleichsmaß wie die empirische Verteilungsfunktion im vorangegangenen Abschnitt, da die Daten selbst keine direkte Information über eine mögliche Baumstruktur hergeben.

Insgesamt sind die Ergebnisse bezüglich der *graph edit distance* sehr ähnlich zu denen der Abstände der Wahrscheinlichkeitsverteilungen. Die Überlegenheit der OTM2s beim Anpassen von wahren OTM3s ist jedoch noch deutlicher. Mit nur einer Ausnahme schneiden OTM2ne und OTM2e immer besser ab als Baum-Mischungs-Modelle mit drei Komponenten. Während die mittleren und medianen Abstände der Wahrscheinlichkeitsverteilungen recht ähnlich waren, tritt hier ein deutlicher Abstand der mittleren und medianen *ged* auf, wenn ein wahres OTM3e angepasst werden soll.

### 3.3.3 Abschließende Bemerkungen

Zusammenfassend lässt sich sagen, dass die Anzahl der zugrunde liegenden Beobachtungen keine große Rolle spielt, wenn die Anpassung verschiedener Modellklassen an einen Datensatz verglichen werden soll. Außerdem ist es relativ gleichwertig, ob wahres und angepasstes Modell

Tabelle 3.7: Globale Reihenfolge der Modellklassen unabhängig von der Beobachtungsanzahl unter Verwendung der *graph edit distance*.

| angepasstes Modell | | wahres Modell | | | | | |
|---|---|---|---|---|---|---|---|
| | | OT | OTM2e | OTM2ne | OTM3e | OTM3ne | CBN |
| OT | Ranking | 2 | 5 | 5 | 5 | 5 | 2 |
| | Mittelwert | 0.600 | 6.484 | 9.264 | 10.580 | 13.348 | 4.116 |
| | Median | 0.400 | 6.200 | 9.000 | 10.400 | 13.400 | 4.000 |
| OTM2e | Ranking | 3 | 1 | 1 | 1 | 2 | 3 |
| | Mittelwert | 5.608 | 1.552 | 4.264 | 5.672 | 8.332 | 8.124 |
| | Median | 5.400 | 1.200 | 4.000 | 5.400 | 8.200 | 8.000 |
| OTM2ne | Ranking | 4 | 2 | 2 | 2 | 3 | 4 |
| | Mittelwert | 5.686 | 1.582 | 4.370 | 5.700 | 8.388 | 8.184 |
| | Median | 5.400 | 1.200 | 4.000 | 5.400 | 8.200 | 8.000 |
| OTM3e | Ranking | 6 | 4 | 3 | 4 | 4 | 5 |
| | Mittelwert | 10.658 | 6.542 | 6.626 | 8.956 | 8.378 | 12.096 |
| | Median | 10.600 | 5.800 | 6.400 | 10.000 | 8.700 | 12.000 |
| OTM3ne | Ranking | 5 | 6 | 4 | 3 | 1 | 6 |
| | Mittelwert | 10.630 | 6.685 | 8.015 | 8.922 | 8.364 | 12.116 |
| | Median | 10.600 | 5.970 | 7.600 | 10.000 | 8.400 | 12.000 |
| CBN | Ranking | 1 | 3 | 6 | 6 | 6 | 1 |
| | Mittelwert | 0.076 | 5.900 | 10.316 | 12.972 | 13.958 | 0.242 |
| | Median | 0.000 | 5.400 | 10.600 | 13.200 | 14.000 | 0.000 |

anhand der Wahrscheinlichkeitsverteilungen oder der grafischen Struktur miteinander verglichen werden. Insgesamt liefert die Simulationsstudie die Erkenntnis, dass alle Eigenschaften und Strukturen onkogenetischer Bäume durch ein CBN abgedeckt werden können. Bezüglich der Anpassungsgüte sind die zwei relevanten Modellklassen die Baum-Mischungs-Modelle und die verbindenden Bayes Netze. Es lässt sich nicht festlegen, welche davon überlegen ist, da OTMs Ereignispfade und Unabhängigkeit gleichzeitig modellieren können, während CBNs die Abhängigkeit von mehreren Ereignissen abbilden können. Beide Modellklassen besitzen damit Eigenschaften, die in bestimmten Situationen notwendig sind und die nicht von der jeweils anderen Modellklasse kompensiert werden können.

# Kapitel 4

# Modellwahl

In diesem Kapitel soll die Frage beantwortet werden, nach welchen Kriterien entschieden werden kann, welches Modell am besten zur Anpassung an die Daten geeignet ist, wenn das wahre Modell unbekannt ist. Dazu wird zunächst der Ansatz von Yin et al. (2006) vorgestellt, der sich mit der Modellwahl bei onkogenetischen Baum-Mischungs-Modellen beschäftigt. Anschließend soll dieser Ansatz auf andere Modelle erweitert oder auch ggf. modifiziert werden.

## 4.1 Modellwahl für onkogenetische Baum-Mischungs-Modelle

In ihrer Arbeit beschäftigen sich Yin et al. (2006) mit der Modellwahl bei onkogenetischen Baum-Mischungs-Modellen. Es geht darum, eine geeignete Zahl $K$ an Baumkomponenten zu bestimmen und dabei weder zu wenige noch zu viele auszuwählen. Bei zu kleinem $K$ ist es nicht möglich alle vorhandenen Pfade in den Daten darzustellen. Ist $K$ zu groß gewählt, treten Probleme wie Overfitting und Redundanz auf. Weiterhin wichtig ist, dass das geschätzte Modell interpretierbar bleibt, d.h. Ziel ist es, das Modell so einfach wie möglich zu wählen und dennoch alle wichtigen Strukturen zu erfassen.

Yin et al. (2006) untersuchen verschiedene bekannte Modellwahlkriterien und entwickeln ein neues Maß um explizit die Redundanz von Baumkomponenten zu berücksichtigen. Die betrachteten Kriterien sind folgende:

- Kreuzvalidierung
- Bayesian model selection
- Empirical Bayes
- AIC und BIC
- modifiziertes BIC

Im Weiteren sollen diese Kriterien vorgestellt und Vor- und Nachteile benannt werden.

Die Idee der **Kreuzvalidierung** beruht darauf, die Daten in eine Trainings- und eine Testmenge zu unterteilen. Bei einer $k$-fachen Kreuzvalidierung werden zufällig $k$ (fast) gleich große Gruppen gebildet. Jede dieser $k$ Gruppen wird einmal als Testmenge verwendet, während die restlichen $k-1$ Gruppen die Trainingsmenge darstellen.

Für die Baummodelle wird die Kreuzvalidierung zur Modellwahl wie folgt verwendet. Zunächst wird das Vorgehen für das Anpassen einer Modellklasse $\mathcal{M}$ beschrieben, wobei $\mathcal{M}$ z.B. ein CBN oder ein OTM2e sein kann. Anhand der Daten wird nun ein konkretes Modell angepasst. Zum Schätzen des Modells werden aber nicht alle Daten verwendet, sondern nur die Daten aus der jeweiligen Trainingsmenge. Als geschätztes Modell erhalten wir $\hat{M} \in \mathcal{M}$. Für die Daten der Testmenge, die nicht zur Modellanpassung verwendet wurde, wird nun die log-Likelihood berechnet:

$$\log\Big(\prod_{x_i \in \text{ Testmenge}} P(x_i|\hat{M})\Big) = \sum_{x_i \in \text{ Testmenge}} \log P(x_i|\hat{M}), \tag{4.1}$$

wobei sich $P(x_i|\hat{M})$ je nach Modellklasse auf unterschiedliche Weise berechnet. Wiederholt man dieses Vorgehen für jede der $k$ Testmengen, erhält man $k$ log-Likelihood Werte, die anschließend gemittelt werden.

Mit allen anderen interessierenden Modellklassen wird genauso verfahren. Letztendlich wird die Modellklasse mit der größten durchschnittlichen Likelihood ausgewählt. Ist die Modellklasse bestimmt, wird aus dieser Klasse ein konkretes Modell auf allen Daten geschätzt.

Eine Variante zur Modellwahl mit Kreuzvalidierung ist die Anwendung der 'one-standard-error-rule' (Hastie et al., 2009). Hier wird nicht das Modell mit der größten durchschnittlichen Likelihood ausgewählt, sondern das Modell mit der geringsten Komplexität, dessen durchschnittliche Likelihood aber noch innerhalb eines Standardfehlers des besten Modells liegt.

Insgesamt ist der Modellwahlansatz über Kreuzvalidierung jedoch sehr rechenintensiv und kann daher nicht auf alle Probleme angewendet werden.

Eine Alternative zur Modellwahl stellen Bayesianische Methoden dar. Ziel ist es, die a posteriori Wahrscheinlichkeit eines Modells gegeben die Daten zu schätzen. Dabei wird der Satz von Bayes ausgenutzt und für die gewünschte a posteriori Wahrscheinlichkeit ergibt sich

$$P(M|D) = \frac{P(D|M)P(M)}{P(D)}, \tag{4.2}$$

wobei $D$ der zugrunde liegende Datensatz ist und $M$ das entsprechende Modell. Mit $\theta$ als zum Modell gehörigem Parametervektor gilt

$$P(M|D) = \frac{P(D|M)P(M)}{P(D)} \propto P(D|M)P(M) = P(M)\int P(D,\theta|M)d\theta. \tag{4.3}$$

Es soll hier nicht nur ein bestimmter Parameterwert, wie z.B. die Maximum-Likelihood Schätzung, betrachtet werden, sondern es wird über alle in Frage kommenden Vektoren integriert. Dieser Integralterm wird als marginale Likelihood bezeichnet. Ziel ist es nun, dieses Integral in geschlossener Form darstellen und damit berechnen zu können. Letztendlich soll das Modell gewählt werden, dessen a posteriori Wahrscheinlichkeit am größten ist.

Zur Berechnung der marginalen Likelihood geht die **Bayesian model selection** dabei wie folgt vor: Sei durch $K$ die Menge aller Baum-Mischungs-Modelle mit $K$ Komponenten bezeichnet und durch $M_K$ ein spezielles dieser Modelle. Mit dem Satz von Bayes soll letztendlich über $P(D|K)$

die a posteriori Wahrscheinlichkeit $P(K|D)$ berechnet werden[1], also die Wahrscheinlichkeit für ein Baum-Mischungs-Modell mit $K$ Komponenten, wenn ein solches Modell an die gegebenen Daten angepasst werden soll. Es gilt

$$P(D|K) = \int P(D|M_K)P(M_K)dM_K, \tag{4.4}$$

da sozusagen über alle Modelle mit $K$ Komponenten gemittelt werden soll. Die Wahrscheinlichkeit $P(D|M_K)$ lässt sich wiederum als marginale Likelihood wie folgt darstellen:

$$P(D|M_K) = \int P(D,\theta|M_K)d\theta = \int P(D|M_K,\theta)P(\theta|M_K)d\theta \tag{4.5}$$

Insgesamt gilt damit

$$P(D|K) = \int\int P(D|M_K,\theta)P(\theta|M_K)P(M_K)d\theta\ dM_K. \tag{4.6}$$

Setzt man nun einen konjugierten Dirichlet Prior auf den Parametervektor $\theta$, erhält man für ein gegebenes Modell $M_K$ eine geschlossene Form für $P(D|M_K)$. Daher kann man die marginale Likelihood durch einen Monte Carlo Ansatz approximieren, der über $P(M_K)$ eine Stichprobe von $n$ Baumtopologien $M_K^{(1)}, ..., M_K^{(n)}$ zieht, so dass man $P(D|K)$ berechnen kann über

$$P(D|K) \approx \frac{1}{n}\sum_{i=1}^{n} P(D|M_K^{(i)}). \tag{4.7}$$

Die Konvergenzgeschwindigkeit ist hier jedoch sehr gering, da man ohne vorheriges Wissen einen uninformativen Prior $P(M_K)$ wählt. Der Beitrag der Baummodelle $M_K$, aus denen die Daten $D$ am wahrscheinlichsten erzeugt werden können, ist für die Summe in Formel (4.7) eher gering, so dass eine Konvergenz zur wahren a posteriori Wahrscheinlichkeit nur langsam erfolgt.

Der Ansatz der Bayesian model selection verwendet zur Bestimmung von $P(D|K)$ alle Modelle $M_K$ mit $K$ Baumkomponenten. Im Gegensatz dazu wird beim **Empirical Bayes** Ansatz (Robert, 2001) nur ein spezielles Modell ausgewählt. Anhand der Daten wird mit Hilfe des ganz normalen Algorithmus' aus Abschnitt 2.7 ein Modell $\hat{M}_K$ geschätzt. Dieses Modell wird als Vorwissen verwendet. Das entspricht einem Prior, der dem Modell $\hat{M}_K$ alle Wahrscheinlichkeitsmasse zuteilt und allen anderen Modellen die Wahrscheinlichkeit Null. Daher kann $P(D|K)$ approximiert werden durch

$$P(D|\hat{M}_K) = \int P(D|\hat{M}_K,\theta)P(\theta|\hat{M}_K)d\theta. \tag{4.8}$$

In Yin et al. (2006) wird beschrieben, wie man einen geeigneten konjugierten Prior wählt und damit $P(D|\hat{M}_K)$ berechnen kann. Die Idee ist über Dirichlet- und Beta-Verteilungen

$$P(x|\hat{M}_K) = f_x^{(\hat{M}_K)}(\tilde{\delta},\tilde{\alpha}) \tag{4.9}$$

zu berechnen, wobei mit den Bezeichnungen aus Formel (2.12) auf Seite 20 gilt

$$f_x^{(M_K)}(\delta,\alpha) := L(x \mid M) = \sum_{k=1}^{K}\delta_k L(x \mid T_k) =: \sum_{k=1}^{K}\delta_k f_x^{(T_k)}(\alpha_k). \tag{4.10}$$

[1] Allerdings kann die Berechnung von $P(D|K)$ auch schon genügen, wenn aufgrund fehlenden Vorwissens $P(K)$ als Gleichverteilung gewählt wird. Genau wie $P(D)$ ist $P(K)$ dann ein Faktor, der für alle Modelle bzw. Modellklassen gleich ist und somit nicht berücksichtigt werden muss.

Die Likelihood wird somit an dem 'Durchschnitts-Parameter' $\tilde{\theta} = (\tilde{\delta}, \tilde{\alpha})$ ausgewertet, wobei $\tilde{\delta}$ und $\tilde{\alpha}$ dabei so gewählt werden, dass alle Ereigniskombinationen, die anhand der Baumtopologie eintreten können, gleichwahrscheinlich sind (siehe Yin et al. (2006) für die genaue Definition).

Letztendlich wird $P(D|\hat{M}_K)$ berechnet über

$$P(D|\hat{M}_K) = \prod_{x \in D} P(x|\hat{M}_K). \tag{4.11}$$

Eine verbreitete Methode zur Approximation der marginalen Likelihood ist auch das **Bayesian Information Criterion** (BIC) (Schwarz, 1978). Hierbei werden Güte der Anpassung und Modellkomplexität gegenübergestellt. Die Anpassungsgüte wird dabei über einen Likelihood Term gemessen, während die Komplexität von der Dimension des Modells abhängt, also z.B. von der Anzahl zu schätzender Parameter. Die Idee dieser Gegenüberstellung ist, dass das gewählte Modell nicht komplexer als notwendig sein soll, um die beobachteten Daten zu beschreiben. Komplexe Modelle, die nur geringfügig mehr Informationen liefern, deren Likelihood also nur ein wenig größer ist, werden somit bestraft.

Ist mit $d$ die Dimension des Modells $M$ gekennzeichnet (s.u. für die Definition), berechnet sich der BIC-Score wie folgt:

$$BIC(M) = LogLikelihood(M) - \frac{d}{2}\log(N), \tag{4.12}$$

wobei $N$ die Anzahl der Beobachtungen ist. Geht $N \to \infty$, so kann mit dem BIC Kriterium das wahre Modell gefunden werden. Das BIC ist somit nur asymptotisch konsistent und daher suboptimal für endliche Datensätze.

Ein dem BIC ähnliches Maß ist **Akaikes Information Criterion** (AIC) (Akaike, 1973), welches wie folgt definiert ist:

$$AIC(M) = LogLikelihood(M) - d \tag{4.13}$$

Beide Modellwahlkriterien (AIC und BIC) beschäftigen sich mit dem Trade-Off zwischen Anpassungsgüte und Modellkomplexität. Sie besitzen aber unterschiedliche Motivationen und beruhen auf unterschiedlichen Voraussetzungen. Diese sind z.B. in Hastie et al. (2009) oder Claeskens und Hjort (2008) nachzulesen.

Um für Baum-Mischungs-Modelle das AIC oder BIC Kriterium berechnen zu können, muss die Dimension $d$ des Modells geeignet definiert werden. Besteht der Baum aus nur einer Komponente ($K = 1$), ist die Dimension des Modells gleich der Anzahl der zu schätzenden Parameter, also der Anzahl der Kantenwahrscheinlichkeiten. Insgesamt werden in dem Modell $n$ Ereignisse dargestellt, was aufgrund der zugrunde liegenden Baumstruktur auch zu $n$ Parametern führt. Für eine Sternkomponente mit gleichen Kantengewichten würde die Dimension $d$ nur gleich 1 sein. Ein Baum-Mischungs-Modell $M_K$ mit $K$ Baumkomponenten besitzt $K + K \cdot n$ Parameter $\theta = (\delta, \alpha)$, wobei $\sum_{k=1}^{K} \delta_k = 1$. Die Dimension eines solchen Modells kann aber deutlich kleiner sein als $K(n+1) - 1$.

Sei wiederum mit den Bezeichnungen aus (2.12) auf Seite 20

$$f_x^{(M_K)}(\delta, \alpha) := L(x \mid M) = \sum_{k=1}^{K} \delta_k L(x \mid T_k) =: \sum_{k=1}^{K} \delta_k f_x^{(T_k)}(\alpha_k), \tag{4.14}$$

dann ist $f$ eine polynomiale Abbildung vom Parameterraum in den Wahrscheinlichkeitsraum des Modells

$$f^{(M_K)} : \mathbb{R}^{K(n+1)-1} \to \mathbb{R}^{2^n-1}, \qquad \theta \to \left(f_x^{(M_K)}(\theta)\right)_{x \in \{0,1\}^n}. \tag{4.15}$$

Somit ist ihr Bild eine algebraische Vielfalt, deren Dimension mit der des statistischen Modells übereinstimmt. Möglichkeiten zur Berechnung dieser werden in Beerenwinkel und Drton (2005) beschrieben. Mit steigender Anzahl an Ereignissen ist diese Methode jedoch computational zu aufwendig. Eine Alternative beruht daher auf der Berechnung der Jacobi Matrix der Abbildung $f$. Falls eine lineare Abbildung $g : \mathbb{R}^m \to \mathbb{R}^n$ vorliegt, so ist die Dimension des Bildes von $g$ gleich dem Rang der Matrix, die $g$ repräsentiert. Dieser kann lokal bei einer glatten Abbildung durch eine lineare Abbildung approximiert werden, die durch die Jacobi Matrix gegeben ist (Spivak, 1979). Für Bayes'sche Netzwerke ist der Rang der Jacobi Matrix mit Wahrscheinlichkeit 1 konstant und entspricht der gesuchten Dimension $d$ (Geiger et al., 1996).

Um die Jacobi Matrix $J(\theta)$ zu berechnen, benötigt man die partiellen Ableitungen nach den Parametern $\theta = (\delta, \alpha)$. Diese werden schließlich ausgewertet an allen Stellen $x \in \{0,1\}^n$, so dass $J$ eine $((K(n+1)) \times 2^n)$-Matrix ist. Für den Parametervektor $\theta$ gilt, dass $\delta = (\delta_1, ..., \delta_K)$ und $\alpha = (\alpha_1, ..., \alpha_K)$, wobei $\alpha_k = (\alpha_{k,1}, ..., \alpha_{k,n})$. Die $\alpha_{k,v}$ geben dabei die bedingte Wahrscheinlichkeit an, dass in der $k$-ten Baumkomponente das Ereignis $v$ eingetreten ist, wenn der Elter von $v$ zuvor schon eingetreten ist ($k = 1, ..., K$; $v = 1, ..., n$). Da $\sum_{k=1}^{K} \delta_k = 1$ gelten muss, kann $\delta_K$ ersetzt werden durch $1 - (\delta_1 + ... + \delta_{K-1})$. Die partiellen Ableitungen ergeben sich damit zu

$$\frac{\partial f_x^{(M_K)}(\delta, \alpha)}{\partial \delta_k} = f_x^{(T_k)}(\alpha_k) - f_x^{(T_K)}(\alpha_K) \tag{4.16}$$

$$\frac{\partial f_x^{(M_K)}(\delta, \alpha)}{\partial \alpha_{k,v}} = \begin{cases} \frac{\delta_k f_x^{(T_k)}(\alpha_k)}{\alpha_{k,v}} & \text{falls } v \in V_k[x] \\ \frac{-\delta_k f_x^{(T_k)}(\alpha_k)}{1-\alpha_{k,v}} & \text{falls } v \notin V_k[x] \quad \text{und} \quad pa(v) \in V_k[x] \\ 0 & \text{sonst} \end{cases} \tag{4.17}$$

Dabei sind $V_k = V[T_k]$ die Knoten der $k$-ten Baumkomponente und $V_k[x] = \{v \in V_k | x_v = 1\}$ die Menge der eingetretenen Ereignisse, die durch die Beobachtung $x$ spezifiziert wird. Bei der partiellen Ableitung nach $\alpha_{k,v}$ muss also danach unterschieden werden, ob $v$ selbst eingetreten ist oder nur der Elternknoten von $v$ in der $k$-ten Baumkomponente oder keins von beidem.

Wird also eine Baumkomponente zu einem Modell hinzugefügt, die sehr ähnlich zu einer bereits bestehenden Baumkomponente ist, wird sich die Dimension nur geringfügig erhöhen. Im Gegensatz dazu wird aber ein deutlicher Anstieg der Redundanz zu beobachten sein. Mit Dimension und Redundanz werden somit unterschiedliche Eigenschaften eines Modells gemessen. Wird in den Modellwahlkriterien aber nur die Dimension berücksichtigt, können

dadurch leicht falsche Schlussfolgerungen bezüglich des besten Modells gezogen werden. Da das Hinzufügen einer ähnlichen Baumkomponente die log-Likelihood nicht reduziert und die Dimension nur wenig erhöht, wird ein Modell mit redundanten Baumkomponenten nicht ausreichend bestraft. Modellwahlkriterien wie AIC und BIC würden demnach unnötig komplexe Modelle auswählen. Daher entwickeln Yin et al. (2006) ein Modellwahlkriterium, das auch die Redundanz von Baumkomponenten berücksichtigen kann.

Neben hohen Dimensionen bestraft das resultierende **modifizierte BIC** ebenfalls eine große Redundanz. Um die Redundanz eines Modells zu bestimmen, muss die Ähnlichkeit $S_{kl}$ zwischen zwei Baumkomponenten $T_k$ und $T_l$ gemessen werden können. Sie wird, wie schon in Abschnitt 3.2.3 angesprochen, wie folgt definiert:

$$S_{kl} = S_{lk} = 1 - \frac{||A_k - A_l||_\infty}{n} \quad \in [0,1]. \tag{4.18}$$

Dabei sind $A_k$ und $A_l$ $(n \times n)$-Adjazenzmatrizen der Baumkomponenten $T_k$ und $T_l$, die in den Spalten $j$ kennzeichnen, ob Knoten $i$ ein Nachfolger des $j$-ten Knotens ist (1 für 'ja' und 0 für 'nein'). Die unendlich Norm einer Matrix $||A||_\infty = max_i \sum_j |a_{ij}|$ bestimmt das Maximum der absoluten Zeilensummen. Mit $||A_k - A_l||_\infty$ wird also der größte Unterschied an ausgehenden Kanten zwischen den zwei Baumkomponenten bestimmt. Die Redundanz ergibt sich dann als maximale Ähnlichkeit zwischen den zum Modell gehörigen Baumkomponenten:

$$R = \max_{\substack{k \neq l \\ k,l=1,\ldots,K}} S_{kl} \tag{4.19}$$

Ein Modellwahlkriterium, das neben der Dimension die Redundanz berücksichtigt, kann wie folgt definiert werden:

$$BIC_R(M) = LogLikelihood(M) - (1 + R)\,\frac{d}{2}\log(N) \tag{4.20}$$

Durch eine identische Baumkomponente in einem Modell ($R = 1$), wird also der Bestrafungsterm im Gegensatz zum Standard BIC-Kriterium verdoppelt. Als neu definiertes Modellwahlkriterium verwenden Yin et al. (2006) ein gewichtetes Mittel zwischen $BIC$ und $BIC_R$:

$$BIC_w(M) = w \cdot BIC(M) + (1 - w) \cdot BIC_R(M) \tag{4.21}$$

mit $w = \min\{\frac{1}{n+1}\max\{d_K - d_{K-1}, 0\}, 1\}$. Dabei ist $d_k$ die Dimension des Modells, wenn $k$ Baumkomponenten vorliegen. Durch diese Gewichtung wird gesichert, dass bei einem nur kleinen Zuwachs der Dimension nach Hinzufügen einer $K$-ten Baumkomponente ($d_K - d_{K-1}$ klein) großes Gewicht auf $BIC_R$ liegt, da die vorliegende Redundanz auch berücksichtigt werden soll. Ist jedoch der Zuwachs der Dimension, also $d_K - d_{K-1}$ groß, deutet dies darauf hin, dass durch die neue Baumkomponente auch wirklich neue Strukturen in den Daten erkannt werden können. In diesem Fall soll die Bestrafung hauptsächlich über die Dimension erfolgen, wie auch bei dem Standard BIC-Kriterium.

Mit Hilfe einer Simulationsstudie bewerten Yin et al. (2006) nun das Abschneiden der unterschiedlichen Modellwahlkriterien. Zum einen wird überprüft, wie oft welche Anzahl an Baumkomponenten ausgewählt wurde. Zum anderen wird auch noch untersucht, wie gut das

ausgewählte Modell zum wahren Modell passt. Dazu werden die drei Scores *recovery*, *precision* und *dissimilarity* herangezogen, die schon in Abschnitt 3.2.3 definiert wurden. Zusätzlich zur Simulationsstudie wird auch noch eine Untersuchung auf echten Datensätzen durchgeführt. Insgesamt zeigt sich, dass einzig die Kreuzvalidierung und das modifizierte BIC Kriterium zufriedenstellende Ergebnisse liefern, d.h. die wahre Modellstruktur am besten aufdecken können. Da die Kreuzvalidierung sehr viel Rechenzeit benötigt, ist das von Yin et al. (2006) entwickelte Modellwahlkriterium am besten geeignet, die richtige Anzahl an Baumkomponenten zu bestimmen.

## 4.2 Simulationsstudie

Basierend auf der in Kapitel 3 durchgeführten Simulationsstudie soll nun ebenfalls untersucht werden, welches Modellwahlkriterium am besten für die Modellwahl bei Krankheitsprogressionsmodellen geeignet ist. In Ergänzung zu dem Paper von Yin et al. (2006) werden hier aber nicht nur onkogenetische Baum-Mischungs-Modelle betrachtet, sondern auch onkogenetische Bäume, verbindende Bayes Netze und Distanzbäume.

Mit Hilfe der Simulationsstudie soll geklärt werden, ob ein Modellwahlkriterium immer die Modellklasse am besten bewertet, aus der auch das wahre Modell stammt. Oder, um es etwas allgemeiner zu formulieren, welche Modellklasse in einer gegebenen Situation am besten geeignet ist, die Wahrheit wiederzugeben. Weiterhin geht es darum, überhaupt ein geeignetes Modellwahlkriterium zu identifizieren. Das heißt, es soll überprüft werden, welches Modellwahlkriterium die meisten 'richtigen' Ergebnisse liefert. Ein wahres OTM kann z.B. nicht gut durch einen OT dargestellt werden. Ein gutes Modellwahlkriterium sollte also für den OT keine besseren Werte liefern als für das OTM. Außerdem soll wiederum der Einfluss der Anzahl von Beobachtungen analysiert werden. Sind z.B. bei $N = 50$ die Ergebnisse schlechter, weil nicht so viele Beobachtungen vorliegen und somit die Modelle nicht zuverlässig schätzen können? Wird bei $N = 1000$ viel öfter die wahre Modellklasse von den Modellwahlkriterien auf Platz 1 gesetzt?

### 4.2.1 Aufbau der Simulationsstudie

Für die hier durchgeführte Analyse gelten die gleichen Gegebenheiten wie in Kapitel 3. Lediglich der vierte Schritt muss abgeändert werden:

1. Wähle ein zugrunde liegendes wahres Baummodell $T$ mit $n$ Ereignissen.
2. Ziehe $N$ Beobachtungen aus $T$ und erhalte die Datenmatrix $X \in \mathbb{B}^{N \times n}$.
3. Wähle eine Modellklasse und passe ein Progressionsmodell $T^*$ an $X$ an.
4. Berechne für die erfolgte Anpassung die entsprechenden Werte der Modellwahlkriterien.

Auch hier werden für jede Parametereinstellung $M = 100$ Wiederholungen durchgeführt. Für jedes Modellwahlkriterium ergeben sich damit 100 Werte für jedes der 7 wahren Modelle und

jede der 5 Beobachtungsanzahlen. Insgesamt werden damit 35 verschiedenen Situationen betrachtet.

Die in dieser Simulationsstudie herangezogenen Modellwahlkriterien sind folgende:

- *AIC*
- *BIC*
- $BIC_w$ (modifiziertes BIC von Yin et al. (2006))
- $BIC_{ged}$ (abgewandeltes modifiziertes BIC)

Das modifizierte BIC Kriterium hat im erwähnten Paper sehr vielversprechende Ergebnisse erzielt. Im Vergleich dazu sollen jedoch auch die gebräuchlichen Kriterien *AIC* und *BIC* betrachtet werden. Weiterhin wird in dieser Arbeit ein abgewandeltes modifiziertes BIC-Kriterium analysiert. Für das modifizierte BIC muss, wie in (4.18) beschrieben, die Ähnlichkeit zwischen zwei Baumkomponenten berechnet werden. Wie schon in Abschnitt 3.2.2 beschrieben, könnte sich hier die Wahl der Unendlichnorm negativ auswirken, da beim Vergleich von zwei Baumkomponenten nur der maximale Unterschied an ausgehenden Kanten berücksichtigt wird, aber nicht gezählt wird, wie oft dieser auftritt. Als alternatives Abstandmaß für zwei Baumkomponenten wird daher hier die *graph edit distance* vorgeschlagen, die alle auftretenden Unterschiede berücksichtigt. Die Berechnung des abgewandelten modifizierten BIC Kriteriums erfolgt also bis auf die Berechnung von $S_{kl}$ analog zum modifizierten BIC. Die alternative Berechnung lautet

$$S_{kl} = S_{lk} = 1 - \frac{ged(T_k, T_l)}{n \cdot 2} \quad \in [0, 1], \tag{4.22}$$

wobei $ged(T_k, T_l)$ die *graph edit distance* zwischen den Baumkomponenten $T_k$ und $T_l$ ist und $n$ die Anzahl der im Baum enthaltenen Ereignisse.

### 4.2.2 Ergebnisse der Simulationsstudie

Die oben genannten vier Modellwahlkriterien werden nun für die angepassten Modelle aus der Simulationsstudie berechnet. Für jede Konstellation aus wahrem Modell und Anzahl an Beobachtungen (insgesamt 35) wird wie folgt vorgegangen. Man berechnet die Modellwahlkriterien für alle 100 Durchläufe, fasst diese aber zu einer Zahl zusammen, indem der Mittelwert gebildet wird.

Bei dieser Mittelwertbildung muss beachtet werden, dass die Modellwahlkriterien den Wert $-\infty$ annehmen können. Das ist der Fall, wenn eine Beobachtung im Datensatz vorliegt, die nicht zum geschätzten Modell passt, z.B. also bei einem OT die Pfadreihenfolge verletzt wird. Es ist jedoch nicht so, dass entweder keiner oder alle 100 Werte für eine Parameterkombination gleich $-\infty$ sind. Eine Übersicht über diese Häufigkeitsverteilung ist in Tabelle 4.1 gegeben.

In 38 Situationen nehmen alle 100 Werte der Modellwahlkriterien den Wert $-\infty$ an. In 17 weiteren Situationen liegt der Anteil dieser Werte über 75%. Das sind die 55 Situationen, in denen das zugrunde liegende wahre Modell ein OTM (4 verschiedene), CBN oder DIST ist und das angepasste Modell ein OT oder CBN (natürlich nicht beim wahren CBN selbst). Genau

Tabelle 4.1: Häufigkeit des Wertes $-\infty$ bei den 100 Durchläufen pro Parameterkonstellation

| Anzahl $-\infty$ | 0 | 1 | 2 | 3 | 28 | 40 | 52 | 78 | 85 | 89 | 93 | 94 | 95 | 96 | 97 | 98 | 99 | 100 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Häufigkeit | 182 | 3 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 3 | 2 | 1 | 5 | 38 |

in diesen Fällen ist auch zu erwarten gewesen, dass eine Modellanpassung nicht erfolgreich sein kann, da den OTs und CBNs die Möglichkeiten fehlen, Ereignispfade und unabhängige Ereignisse gleichzeitig zu modellieren.

Es gibt jedoch noch 8 weitere Parameterkonstellationen, in denen teilweise nur sehr wenige bis hin zu 50% der 100 Werte der Modellwahlkriterien den Wert $-\infty$ annehmen. Es handelt sich dabei um folgende Situationen: OT als wahres und angepasstes Modell mit 50 bis 500 Beobachtungen, OT als wahres und CBN als angepasstes Modell mit 50 und 200 Beobachtungen sowie CBN als wahres und angepasstes Modell mit 50 und 100 Beobachtungen. Gerade in Situationen mit wenigen Beobachtungen kann es vorkommen, dass nicht alle möglichen Ereigniskombinationen im gezogenen Datensatz vorkommen bzw. zufällig gehäuft bestimmte Ereigniskombinationen, so dass die Struktur des geschätzten Modells vom Wahren abweicht. Da hier die Häufigkeit des Wertes $-\infty$ jedoch entweder sehr gering ist bzw. bei einer größeren Anzahl dies sehr selten vorkommt, werden die aufgetretenen Werte von $-\infty$ in den oben genannten 8 Situationen bei der Mittelwertbildung der 100 Werte der Modellwahlkriterien nicht berücksichtigt. Der Grund dafür ist, dass sonst der Mittelwert in diesen Fällen immer bei $-\infty$ liegen würde, die Modellklasse somit automatisch auf den letzten Platz fällt, obwohl in den verbleibenden Situationen die Modellanpassung sehr gut sein kann.

Anhand der berechneten Mittelwerte sollen nun die Modellwahlkriterien miteinander verglichen werden. Dabei geht es um die Frage, ob die Modellwahlkriterien nach der Anpassung verschiedener Modelle für die wahre Modellklasse die größten Werte liefern. Dann wäre das entsprechende Kriterium geeignet, auch in unbekannten Datensituationen eine gute Entscheidung zu treffen, mit welcher Modellklasse ein Progressionsmodell angepasst werden sollte. Um dies zu bewerten, wird für jedes Modellwahlkriterium und für jede Beobachtungsanzahl gezählt, in wie vielen der sieben Situationen[2] die wahre Modellklasse am besten abschneidet. Die Ergebnisse sind in Tabelle 4.2 dargestellt.

Insgesamt wird in lediglich 2 bzw. 3 von 7 Fällen die wahre Modellklasse durch die Modellwahlkriterien am besten bewertet. Damit liegen die Anteile 'richtiger' Entscheidungen nur bei etwa einem Drittel. Es ist weder ein deutlicher Unterschied zwischen den Modellwahlkriterien noch zwischen den Beobachtungsanzahlen festzustellen.

Um die Ergebnisse dieser Tabelle jedoch besser interpretieren zu können, müssen die Sachverhalte etwas detaillierter betrachtet werden. Dazu wird wie beim Modellvergleich in Kapitel 3 eine Reihenfolge angegeben, wie gut die einzelnen Modellklassen bei der Anpassung abschneiden. Ein Rang von 1 bedeutet dabei, dass die Modellwahlkriterien für die anpassende Modellklasse den größten Wert geliefert haben, während Rang 7 entsprechend bedeutet, dass

[2] Es gibt sieben zugrunde liegende wahre Modelle.

Tabelle 4.2: Wie oft bewerten die einzelnen Modellwahlkriterien das wahre Modell als beste Anpassung? Für jede der fünf verschiedenen Beobachtungsanzahlen gibt es sieben wahre Modelle, so dass ein Wert von 0 das schlechteste und ein Wert von 7 das beste Ergebnis darstellt.

| Kriterium | $N = 50$ | $N = 100$ | $N = 200$ | $N = 500$ | $N = 1000$ | Anteil über alle $N$ |
|---|---|---|---|---|---|---|
| *AIC* | 2 | 2 | 2 | 3 | 3 | 0.34 |
| *BIC* | 3 | 2 | 2 | 2 | 2 | 0.31 |
| $BIC_w$ | 3 | 2 | 2 | 2 | 2 | 0.31 |
| $BIC_{ged}$ | 3 | 2 | 2 | 2 | 2 | 0.31 |

diese Modellklasse den kleinsten und somit schlechtesten Wert erzielt hat. Diese Reihenfolgen sind in den Tabellen 4.3 bis 4.5 dargestellt.

Ziel ist es nun zu erklären, warum nur in ungefähr einem Drittel der Fälle die wahre Modellklasse von den Modellwahlkriterien als beste identifiziert wird. Außerdem möchte man die Modellwahlkriterien untereinander vergleichen sowie überprüfen, ob die Beobachtungsanzahl insofern einen Einfluss hat, als dass Ergebnisse mit vielen Beobachtungen gegebenenfalls besser sind.

Die Einträge in Tabelle 4.2 resultieren hauptsächlich daraus, dass bei der Anpassung eines CBNs bzw. Distanzbaums die jeweiligen wahren Modellklassen von den Modellwahlkriterien als bestes bewertet werden. Bis auf drei Ausnahmen steht dadurch in jedem Tabellenfeld mindestens eine 2. In den meisten Fällen ist dies aber auch schon der endgültige Eintrag. Aus Tabelle 4.5 lässt sich ergänzend ablesen, dass sich bis auf zwei Ausnahmen beim Distanzbaum alle vier Modellwahlkriterien einig sind, welches die beste Modellklasse ist. Dies wird auch unabhängig von der zugrunde liegenden Anzahl an Beobachtungen zuverlässig erkannt.

Betrachtet man die Ergebnisse der Anpassung an einen wahren onkogenetischen Baum (Tabelle 4.3 linke Spalte), so ist zu erkennen, dass nur bei $N = 50$ auch der onkogenetische Baum als bestes anpassendes Modell eingeordnet wird. Für alle weiteren Beobachtungsanzahlen liegt der OT jedoch auf dem zweiten Platz. Als besseres Modell wird hier das CBN angesehen. Auch beim Vergleich der Modelle (Kapitel 3) hat sich herausgestellt, dass ein CBN beim Anpassen eines OTs bevorzugt wird. Dies ist durchaus nachvollziehbar, da CBNs eine direkte Verallgemeinerung der OTs sind und somit alle Eigenschaften von OTs umfassen. Die Modellwahlkriterien liefern also auch für wahre OTs eine richtige Entscheidung.

Bei der Anpassung an onkogenetische Baum-Mischungs-Modelle sind die Ergebnisse weniger eindeutig. Insgesamt fällt auf, dass hier sehr häufig (in $62,5\%$ der Fälle) die Distanzbäume von den Modellwahlkriterien als beste anpassende Modellklasse gesehen werden. Auf den nachfolgenden Plätzen, deren berechnete Mittelwerte der Modellwahlkriterien sich teilweise kaum von denen der Distanzbäume unterscheiden, liegen jedoch wieder OTMs. Wie beim Vergleich der Modelle ist auch hier zu beobachten, dass tendenziell sowohl für wahre OTM2s als auch OTM3s Modelle mit nur zwei Baumkomponenten besser abschneiden. Weiterhin sind auch Modelle mit ungleichen Kantengewichten wieder besser platziert als ihre Gegenstücke

Tabelle 4.3: Reihenfolge der angepassten Modellklassen für jedes Modellwahlkriterium und jede Beobachtungsanzahl (Teil 1)

| | | wahres Modell: OT | | | | wahres Modell: OTM2e | | | | wahres Modell: OTM2ne | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *AIC* | *BIC* | $BIC_w$ | $BIC_{ged}$ | *AIC* | *BIC* | $BIC_w$ | $BIC_{ged}$ | *AIC* | *BIC* | $BIC_w$ | $BIC_{ged}$ |
| | OT | 1 | 1 | 1 | 1 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 |
| | OTM2e | 4 | 3 | 4 | 4 | 3 | 1 | 2 | 2 | 3 | 1 | 2 | 2 |
| | OTM2ne | 5 | 5 | 5 | 5 | 2 | 3 | 3 | 3 | 2 | 3 | 3 | 3 |
| $N = 50$ | OTM3e | 6 | 6 | 6 | 6 | 5 | 4 | 4 | 4 | 4 | 4 | 4 | 4 |
| | OTM3ne | 7 | 7 | 7 | 7 | 4 | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| | DIST | 3 | 4 | 3 | 3 | 1 | 2 | 1 | 1 | 1 | 2 | 1 | 1 |
| | CBN | 2 | 2 | 2 | 2 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 |
| | OT | 2 | 2 | 2 | 2 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 |
| | OTM2e | 4 | 4 | 4 | 4 | 4 | 2 | 2 | 2 | 5 | 2 | 2 | 2 |
| | OTM2ne | 5 | 5 | 5 | 5 | 1 | 3 | 3 | 3 | 2 | 3 | 3 | 3 |
| $N = 100$ | OTM3e | 6 | 6 | 6 | 6 | 5 | 4 | 4 | 4 | 4 | 4 | 4 | 4 |
| | OTM3ne | 7 | 7 | 7 | 7 | 3 | 5 | 5 | 5 | 3 | 5 | 5 | 5 |
| | DIST | 3 | 3 | 3 | 3 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | CBN | 1 | 1 | 1 | 1 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 |
| | OT | 2 | 2 | 2 | 2 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 |
| | OTM2e | 6 | 4 | 4 | 4 | 5 | 3 | 2 | 3 | 5 | 3 | 3 | 3 |
| | OTM2ne | 4 | 5 | 5 | 5 | 1 | 2 | 3 | 2 | 2 | 2 | 2 | 2 |
| $N = 200$ | OTM3e | 5 | 6 | 6 | 6 | 4 | 5 | 4 | 4 | 4 | 4 | 4 | 4 |
| | OTM3ne | 7 | 7 | 7 | 7 | 3 | 4 | 5 | 5 | 3 | 5 | 5 | 5 |
| | DIST | 3 | 3 | 3 | 3 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | CBN | 1 | 1 | 1 | 1 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 |
| | OT | 2 | 2 | 2 | 2 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 |
| | OTM2e | 4 | 3 | 4 | 4 | 5 | 4 | 3 | 3 | 5 | 5 | 3 | 3 |
| | OTM2ne | 5 | 5 | 5 | 5 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 |
| $N = 500$ | OTM3e | 6 | 6 | 6 | 6 | 4 | 5 | 5 | 5 | 4 | 4 | 4 | 4 |
| | OTM3ne | 7 | 7 | 7 | 7 | 2 | 3 | 4 | 4 | 3 | 3 | 5 | 5 |
| | DIST | 3 | 4 | 3 | 3 | 3 | 2 | 1 | 1 | 1 | 1 | 1 | 1 |
| | CBN | 1 | 1 | 1 | 1 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 |
| | OT | 2 | 2 | 2 | 2 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 |
| | OTM2e | 3 | 3 | 4 | 4 | 5 | 4 | 3 | 3 | 5 | 5 | 5 | 5 |
| | OTM2ne | 5 | 5 | 5 | 5 | 1 | 1 | 2 | 2 | 1 | 2 | 2 | 2 |
| $N = 1000$ | OTM3e | 6 | 6 | 6 | 6 | 4 | 5 | 5 | 5 | 4 | 4 | 4 | 4 |
| | OTM3ne | 7 | 7 | 7 | 7 | 2 | 3 | 4 | 4 | 3 | 3 | 3 | 3 |
| | DIST | 4 | 4 | 3 | 3 | 3 | 2 | 1 | 1 | 2 | 1 | 1 | 1 |
| | CBN | 1 | 1 | 1 | 1 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 |

mit gleichen Kantengewichten bei der Sternkomponente. Diese Beobachtung gilt allerdings nur für Beobachtungszahlen ab 200. Onkogenetische Bäume und CBNs liegen bei der Anpassung von Baum-Mischungs-Modellen immer auf dem geteilten letzten Platz. Da diese Modellklassen nicht in der Lage sind, abhängige und unabhängige Ereignisse gleichzeitig zu modellieren, liegt die Likelihood immer bei $-\infty$.

Insgesamt ist aber bei der Anpassung von Baum-Mischungs-Modellen nicht zu beobachten, dass die vier Modellwahlkriterien immer übereinstimmen. Beispielsweise bewertet das AIC-Kriterium die Anpassung eines wahren OTM2ne durch ein OTM2e immer schlechter als es die anderen Kriterien tun. Erst bei $N = 1000$ Beobachtungen wird das OTM2e übereinstimmend auf Platz 5 gesetzt. Für die BIC-Varianten lag diese Modellklasse für 50 und 100 Beobachtungen aber noch auf dem zweiten Platz. Somit kann auch keine allgemeine Aussage bezüglich der Beobachtungsanzahlen getroffen werden. Für manche Modellklassen bleibt die Platzierung

Tabelle 4.4: Reihenfolge der angepassten Modellklassen für jedes Modellwahlkriterium und jede Beobachtungsanzahl (Teil 2)

| | | wahres Modell: OTM3e | | | | wahres Modell: OTM3ne | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | *AIC* | *BIC* | $BIC_w$ | $BIC_{ged}$ | *AIC* | *BIC* | $BIC_w$ | $BIC_{ged}$ |
| | OT | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 |
| | OTM2e | 1 | 1 | 2 | 1 | 2 | 1 | 2 | 2 |
| | OTM2ne | 2 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| $N = 50$ | OTM3e | 3 | 4 | 4 | 4 | 4 | 4 | 4 | 4 |
| | OTM3ne | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| | DIST | 4 | 2 | 1 | 2 | 1 | 2 | 1 | 1 |
| | CBN | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 |
| | OT | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 |
| | OTM2e | 2 | 1 | 2 | 1 | 4 | 1 | 2 | 2 |
| | OTM2ne | 1 | 2 | 3 | 3 | 2 | 3 | 3 | 3 |
| $N = 100$ | OTM3e | 4 | 4 | 4 | 4 | 5 | 4 | 4 | 4 |
| | OTM3ne | 3 | 5 | 5 | 5 | 3 | 5 | 5 | 5 |
| | DIST | 5 | 3 | 1 | 2 | 1 | 2 | 1 | 1 |
| | CBN | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 |
| | OT | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 |
| | OTM2e | 3 | 1 | 2 | 1 | 5 | 3 | 2 | 2 |
| | OTM2ne | 1 | 2 | 3 | 3 | 2 | 2 | 3 | 3 |
| $N = 200$ | OTM3e | 4 | 3 | 4 | 4 | 4 | 5 | 4 | 4 |
| | OTM3ne | 2 | 5 | 5 | 5 | 3 | 4 | 5 | 5 |
| | DIST | 5 | 4 | 1 | 2 | 1 | 1 | 1 | 1 |
| | CBN | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 |
| | OT | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 |
| | OTM2e | 4 | 2 | 1 | 1 | 5 | 4 | 3 | 3 |
| | OTM2ne | 2 | 1 | 3 | 2 | 2 | 2 | 2 | 2 |
| $N = 500$ | OTM3e | 3 | 4 | 4 | 4 | 4 | 5 | 5 | 5 |
| | OTM3ne | 1 | 3 | 5 | 5 | 1 | 3 | 4 | 4 |
| | DIST | 5 | 5 | 2 | 3 | 3 | 1 | 1 | 1 |
| | CBN | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 |
| | OT | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 |
| | OTM2e | 4 | 3 | 2 | 2 | 5 | 5 | 4 | 4 |
| | OTM2ne | 2 | 1 | 1 | 1 | 3 | 2 | 2 | 2 |
| $N = 1000$ | OTM3e | 3 | 4 | 4 | 4 | 4 | 4 | 5 | 5 |
| | OTM3ne | 1 | 2 | 5 | 3 | 1 | 3 | 3 | 3 |
| | DIST | 5 | 5 | 3 | 5 | 2 | 1 | 1 | 1 |
| | CBN | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 |

mit wachsender Beobachtungszahl gleich, für manche ändert sie sich jedoch. Ein allgemeines Muster ist jedoch nicht zu erkennen.

Eine große Übereinstimmung besteht jedoch zwischen den zwei BIC-Varianten $BIC_w$ und $BIC_{ged}$. Die zweite Modifikation des BIC-Kriteriums wurde vorgenommen, um dem vermeintlich weniger geeigneten Abstandsmaß zwischen zwei Baumkomponenten entgegenzuwirken. Im entsprechenden Berechnungsschritt wurde die Unendlichnorm durch die *graph edit distance*

Tabelle 4.5: Reihenfolge der angepassten Modellklassen für jedes Modellwahlkriterium und jede Beobachtungsanzahl (Teil 3)

| | | wahres Modell: DIST | | | | wahres Modell: CBN | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | *AIC* | *BIC* | $BIC_w$ | $BIC_{ged}$ | *AIC* | *BIC* | $BIC_w$ | $BIC_{ged}$ |
| | OT | 6.5 | 6.5 | 6.5 | 6.5 | 7 | 7 | 7 | 7 |
| | OTM2e | 3 | 1 | 2 | 2 | 6 | 5 | 3 | 4 |
| | OTM2ne | 5 | 5 | 3 | 3 | 4 | 3 | 5 | 5 |
| $N = 50$ | OTM3e | 4 | 4 | 4 | 5 | 2 | 2 | 4 | 3 |
| | OTM3ne | 1 | 3 | 5 | 4 | 3 | 4 | 6 | 6 |
| | DIST | 2 | 2 | 1 | 1 | 5 | 6 | 2 | 2 |
| | CBN | 6.5 | 6.5 | 6.5 | 6.5 | 1 | 1 | 1 | 1 |
| | OT | 6.5 | 6.5 | 6.5 | 6.5 | 7 | 7 | 7 | 7 |
| | OTM2e | 5 | 2 | 2 | 2 | 6 | 6 | 5 | 5 |
| | OTM2ne | 3 | 4 | 3 | 3 | 4 | 4 | 4 | 4 |
| $N = 100$ | OTM3e | 4 | 5 | 5 | 5 | 2 | 2 | 3 | 3 |
| | OTM3ne | 2 | 3 | 4 | 4 | 3 | 3 | 6 | 6 |
| | DIST | 1 | 1 | 1 | 1 | 5 | 5 | 2 | 2 |
| | CBN | 6.5 | 6.5 | 6.5 | 6.5 | 1 | 1 | 1 | 1 |
| | OT | 6.5 | 6.5 | 6.5 | 6.5 | 7 | 7 | 7 | 7 |
| | OTM2e | 5 | 3 | 2 | 2 | 6 | 6 | 6 | 6 |
| | OTM2ne | 3 | 4 | 4 | 4 | 4 | 4 | 4 | 4 |
| $N = 200$ | OTM3e | 4 | 5 | 5 | 5 | 3 | 2 | 2 | 2 |
| | OTM3ne | 2 | 2 | 3 | 3 | 2 | 3 | 5 | 3 |
| | DIST | 1 | 1 | 1 | 1 | 5 | 5 | 3 | 5 |
| | CBN | 6.5 | 6.5 | 6.5 | 6.5 | 1 | 1 | 1 | 1 |
| | OT | 6.5 | 6.5 | 6.5 | 6.5 | 7 | 7 | 7 | 7 |
| | OTM2e | 5 | 5 | 5 | 5 | 6 | 6 | 6 | 6 |
| | OTM2ne | 4 | 4 | 3 | 4 | 4 | 4 | 4 | 4 |
| $N = 500$ | OTM3e | 3 | 3 | 4 | 3 | 3 | 2 | 2 | 2 |
| | OTM3ne | 2 | 2 | 2 | 2 | 2 | 3 | 3 | 3 |
| | DIST | 1 | 1 | 1 | 1 | 5 | 5 | 5 | 5 |
| | CBN | 6.5 | 6.5 | 6.5 | 6.5 | 1 | 1 | 1 | 1 |
| | OT | 6.5 | 6.5 | 6.5 | 6.5 | 7 | 7 | 7 | 7 |
| | OTM2e | 5 | 5 | 5 | 5 | 6 | 6 | 6 | 6 |
| | OTM2ne | 4 | 4 | 3 | 4 | 4 | 4 | 4 | 4 |
| $N = 1000$ | OTM3e | 3 | 3 | 4 | 3 | 3 | 3 | 2 | 2 |
| | OTM3ne | 1 | 2 | 2 | 2 | 2 | 2 | 3 | 3 |
| | DIST | 2 | 1 | 1 | 1 | 5 | 5 | 5 | 5 |
| | CBN | 6.5 | 6.5 | 6.5 | 6.5 | 1 | 1 | 1 | 1 |

ersetzt. Wie sich jedoch hier herausstellt, hat diese Veränderung nicht zu besseren Ergebnissen geführt.

Zusammenfassend ergibt sich aus dieser Simulationsstudie kein eindeutiges Ergebnis, welches Modellwahlkriterium am besten geeignet ist, eine gute Modellklasse zur Anpassung auszuwählen. Für die Modelle mit nur einer Baumkomponente (OT, DIST und CBN) treffen alle vier Modellwahlkriterien eine qualifizierte Entscheidung, für OTMs liefert keins der Kriterien für

alle Situationen nachvollziehbare belastbare Ergebnisse. Weitere Analyseschritte sind daher notwendig. Zum einen könnten dazu wie im Paper von Yin et al. (2006) die Abstandsmaße *recovery, precision* und *dissimilarity* (siehe Abschnitt 3.2.2) herangezogen werden. Zum anderen ist es auch denkbar, für die geschachtelten Modellklassen der OTs und OTMs einen Likelihood-Quotienten-Test durchzuführen und die Ergebnisse mit denen der Modellwahlkriterien zu vergleichen.

# Kapitel 5

# Variablenselektion für onkogenetische Baummodelle

Krankheitsprogressionsmodelle tragen dazu bei, wichtige Schritte im Verlauf einer Krankheit besser verstehen zu können. Dazu wird für bestimmte Ereignisse eine Reihenfolge angegeben, die in einem Baummodell dargestellt werden kann. Bevor jedoch diese Reihenfolge berechnet werden kann, müssen die Ereignisse identifiziert werden, die für den Krankheitsverlauf eine entscheidende Rolle spielen. Veränderungen an Chromosomen können nämlich auch rein zufällig auftreten und sind nicht ausschließlich auf das Vorhandensein einer bestimmen Krankheit zurückzuführen. In dieser Arbeit werden hauptsächlich die kurzen und langen Arme von Chromosomen betrachtet. Es ist aber mittlerweile sehr verbreitet, Chromosomenbanden oder sogar einzelne Gene zu betrachten, da dies eine genauere Abgrenzung einzelner Ereignisse erlaubt. Verfahren, die teilweise unter hunderten Ereignissen diejenigen identifizieren, die in Zusammenhang mit einer Krankheit und ihrem Fortschreiten stehen, werden daher dringend benötigt. Auch vor dem Hintergrund, dass die in Kapitel 2 beschriebenen Progressionsmodelle teilweise nur auf eine begrenzte Anzahl Ereignisse anwendbar sind bzw. ein Baum mit mehr als 30 Ereignissen unstabil, unübersichtlich und uninterpretierbar wird, werden Kriterien benötigt, die die Anzahl der möglichen Ereignisse reduzieren.

Im Folgenden werden daher zehn verschiedene Variablenselektionskriterien vorgestellt, siehe Abschnitt 5.1. In einer umfassenden Simulationsstudie (Abschnitt 5.2) wird bewertet, welche Kriterien geeignet sind, eine gute Auswahl der Ereignisse zu treffen. Anschließend werden diese Verfahren auf echte Daten angewendet, um ihre Praxistauglichkeit zu überprüfen, siehe Abschnitt 5.3. Einige abschließende Bemerkungen sind in Abschnitt 5.4 aufgeführt.

## 5.1 Variablenselektionsmethoden

In diesem Abschnitt werden zehn Variablenselektionsmethoden vorgestellt, die angewendet werden können, um die für die Krankheitsprogression relevanten Ereignisse herauszufiltern. Ausgangspunkt für diese Variablenselektion ist die binäre Datenmatrix $X = [\mathbf{x_1}, \ldots, \mathbf{x_n}] \in \mathbb{B}^{m \times n}$. Sie beschreibt das Auftreten von $n$ genetischen Ereignissen in $m$ Beobachtungen, d.h. der $m$-dimensionale Vektor $\mathbf{x_i}$ repräsentiert das $i$-te Ereignis.

Tabelle 5.1: Übersicht über die 10 verschiedenen Variablenselektionsmethoden

| Name | Abkürzung | Kurze Beschreibung |
|---|---|---|
| Univariate Häufigkeit | `freq` | Häufigkeit oberhalb eines Grenzwertes |
| Methode von Brodeur | `brod` | Signifikante Häufigkeit im Vergleich zur Gleichverteilung |
| Paarweise Korrelation | `cor` | Ereignispaare mit hoher Korrelation |
| Exakter Test von Fisher | `fisher` | Ereignispaare mit signifikantem Zusammenhang |
| Fishers z-Transformation | `z` | Ereignispaare mit signifikantem Zusammenhang |
| Gewichte von Edmonds | `weight` | Ereignispaare mit großen Kantengewichten |
| Bedingte Wsk. im Baum | `OT` | Große bedingte Wahrscheinlichkeiten im OT |
| Unabhängigkeit im Baum | `single` | Entferne einzelne unabhängige Ereignisse im OT |
| Größte Cliquen | `lcliq` | Zugehörig zum größten Teilgraphen |
| Maximale Cliquen | `mcliq` | Zugehörig zum Teilgraphen mit maximalem Gewicht |

Einen Überblick über alle betrachteten Variablenselektionsmethoden liefert Tabelle 5.1. Die vorgestellten 10 Verfahren lassen sich in 4 Kategorien unterteilen. Die ersten beiden Verfahren basieren auf einer Häufigkeitsauswahl. Es handelt sich hierbei auch um die einzigen zwei Verfahren, die bisher in der Literatur zur Variablenselektion für Krankheitsprogressionsmodelle eingesetzt wurden.

Da die Ereignisse in den Baummodellen eine zeitliche bzw. kausale Abfolge darstellen sollen, sind sie demzufolge in bestimmter Weise voneinander abhängig. Die zweite Gruppe von Variablenselektionsmethoden beschäftigt sich daher damit, auf verschiedene Weisen Abhängigkeiten zwischen Ereignissen zu finden.

Des Weiteren kann man sich zur Variablenselektion auch direkt die Anpassung eines onkogenetischen Baums zunutze machen. Man kann auf den Algorithmus von Edmonds zurückgreifen, der bei der Anpassung eines OTs eine wichtige Rolle spielt, man kann die Kantengewichte eines bereits angepassten OTs betrachten oder man legt den Fokus auf die Ereignisse, die als einzelne unabhängige Knoten von der Wurzel ausgehen. Bei dieser Gruppe von Variablenselektionsmethoden ist zu beachten, dass sie konkret auf die Modellklasse der onkogenetischen Bäume zugeschnitten sind.

Die Idee für einen weiteren Ansatz zur Variablenselektion, der jedoch in der Literatur nicht weiter verfolgt wurde, liefern Desper et al. (1999). Es geht darum, größte oder maximale Cliquen in einem ungerichteten Graphen zu finden, der nur eine Kante zwischen zwei Ereignissen enthält, wenn diese oft genug gleichzeitig aufgetreten sind.

Die expliziten Auswahlregeln der einzelnen Verfahren werden nun im Folgenden beschrieben.

### 5.1.1 Univariate Häufigkeit

Ein einfacher und intuitiver Ansatz zur Variablenselektion ist der, dass die Ereignisse, die am häufigsten beobachtet werden, auch die wichtigsten sind. Es werden demnach die Ereignisse

ausgewählt, deren relative Häufigkeit über einem festgesetzten Schwellenwert $\tau_{\text{freq}} \in (0, 1)$ liegt. Formal gilt, dass ein Ereignis $i \in \{1, \ldots, n\}$ ausgewählt wird, falls

$$\overline{x_i} \geq \tau_{\text{freq}}, \tag{5.1}$$

mit $\overline{x_i} = \frac{1}{m} \sum_{k=1}^{m} x_i^k$ als arithmetischem Mittel und $x_i^k$ als $k$-ter Komponente von $x_i$.

### 5.1.2 Ansatz von Brodeur

Brodeur et al. (1982) schlagen einen ebenfalls häufigkeitsbasierten Ansatz vor, um die nicht zufälligen Ereignisse speziell in Krebsdatensätzen zu identifizieren. Unter der Nullhypothese, dass alle Ereignisse zufällig auftreten, wird angenommen, dass die Ereignisse voneinander unabhängig und mit gleicher Wahrscheinlichkeit auftreten. Darauf basierend kann man nun die beobachtete und die erwartete Verteilung der Ereignisse miteinander vergleichen. Mit Hilfe einer Monte Carlo Simulation werden 10.000 zufällige Datensätze erzeugt, um die Häufigkeiten unter der Nullhypothese zu erhalten. Für jeden dieser 10.000 Durchläufe wird die größte aufgetretene Häufigkeit notiert. Ein Ereignis wird dann als nicht zufällig betrachtet, wenn die beobachtete Häufigkeit das 95%-Quantil der notierten maximalen Häufigkeiten überschreitet, d.h.

$$\overline{x_i} \geq \tau^*_{\text{freq}} \tag{5.2}$$

mit $\tau^*_{\text{freq}}$ als oben benanntes 95%-Quantil.

Dieser Ansatz von Brodeur ist ein häufigkeitsbasierter Ansatz, bei dem der Schwellenwert nicht im Voraus festgelegt werden muss, sondern vom Verfahren selbst berechnet wird.

Soll die Variablenselektion auf Daten stattfinden, bei denen die Ereignisse durch Mutationen an Chromosomenarmen charakterisiert sind, schlagen Brodeur et al. (1982) vor, nicht die Gleichverteilung zugrunde zu legen, sondern eine Verteilung, die die Länge der Chromosomenarme berücksichtigt. Wird diese längenproportionale Nullverteilung gewählt, müssen normalisierte Ereignishäufigkeiten berechnet und mit den normalisierten beobachteten Häufigkeiten verglichen werden. Genaueres dazu kann in Brodeur et al. (1982) oder Huang et al. (2004) nachgelesen werden.

### 5.1.3 Paarweise Korrelation

Die Idee dieser Methode ist, alle diejenigen Ereignisse auszuwählen, die mit mindestens einem weiteren Ereignis ausreichend korreliert sind. Für binäre Ereignisse, wie sie für Krankheitsprogressionsmodelle vorliegen, ist der Korrelationskoeffizient von Pearson äquivalent zum Phi-Koeffizient. Die paarweise Korrelation zwischen zwei Ereignissen $i$ und $j$ ($i, j \in \{1, \ldots, n\}$) ist definiert als

$$r_{ij} := \frac{\sum_{k=1}^{m} (x_i^k - \overline{x_i})(x_j^k - \overline{x_j})}{\sqrt{\sum_{k=1}^{m} (x_i^k - \overline{x_i})^2 \sum_{k=1}^{m} (x_j^k - \overline{x_j})^2}}, \tag{5.3}$$

wobei $x_i^k$ und $x_j^k$ die jeweils $k$-ten Komponenten der entsprechenden Vektoren sind.

Die Definition des Phi-Koeffizienten, der die Assoziation zwischen Ereignis *i* und *j* beschreibt, lautet

$$\phi = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{n_{1\cdot}n_{0\cdot}n_{\cdot 1}n_{\cdot 0}}}, \tag{5.4}$$

wobei $n_{11}, n_{00}, n_{10}$ und $n_{01}$ nicht negative Beobachtungsanzahlen sind, die sich zu $n$ aufsummieren:

| | $j = 1$ | $j = 0$ | total |
|---|---|---|---|
| $i = 1$ | $n_{11}$ | $n_{10}$ | $n_{1\cdot}$ |
| $i = 0$ | $n_{01}$ | $n_{00}$ | $n_{0\cdot}$ |
| total | $n_{\cdot 1}$ | $n_{\cdot 0}$ | $n$ |

Sei ein Schwellenwert $\tau_{\text{cor}} \in (0, 1)$ für die minimale Korrelation gegeben, wird ein Ereignis *i* ausgewählt, falls

$$\exists\, j \in \{1, \ldots, n\} \backslash \{i\} : |r_{ij}| \geq \tau_{\text{cor}}. \tag{5.5}$$

### 5.1.4 Exakter Test von Fisher

Will man den Zusammenhang bzw. die Abhängigkeit zwischen zwei Ereignissen für die Variablenselektion berücksichtigen, kann auch der exakte Test von Fisher (Agresti, 1992) herangezogen werden. Es werden alle $\binom{n}{2}$ p-Werte $p_{ij}$ von Ereignispaaren $(i, j)$ berechnet ($i, j = 1, \ldots, n$, $i < j$) und dann die Ereignisse ausgewählt, deren p-Wert auf eine Abhängigkeit hinweist. Ist ein Schwellenwert $\tau_{\text{fisher}} \in (0, 1)$ gegeben, werden beide Ereignisse *i* und *j* ausgewählt, falls

$$p_{ij} \leq \tau_{\text{fisher}}. \tag{5.6}$$

### 5.1.5 Fishers z-Transformation

Eine weitere Variablenselektionsmethode, die auch auf einem Testverfahren basiert, verwendet Konfidenzintervalle für den Korrelationskoeffizienten von Pearson. Pigott (2012) schlägt vor, Fishers z-Transformation auf den Korrelationskoeffizienten für Ereignispaare anzuwenden, um somit eine approximativ normalverteilte Zufallsvariable zu erhalten. Diese Transformation ist wie folgt definiert:

$$z_{ij} = 0.5 \ln\left(\frac{1 + r_{ij}}{1 - r_{ij}}\right). \tag{5.7}$$

Die asymptotische Varianz von $z_{ij}$ ist gegeben durch $\text{Var}(z_{ij}) = \frac{1}{m-3}$, so dass durch

$$\text{KI} = \left[z_{ij} - u_{1-\frac{\alpha}{2}} \cdot \frac{1}{\sqrt{m-3}}, z_{ij} + u_{1-\frac{\alpha}{2}} \cdot \frac{1}{\sqrt{m-3}}\right] \tag{5.8}$$

ein asymptotisches $(1 - \alpha)$ Konfidenzintervall gegeben ist, mit $u_{1-\frac{\alpha}{2}}$ als $(1 - \frac{\alpha}{2})$-Quantil der Standardnormalverteilung.

Dieses Konfidenzintervall kann nun zur Variablenselektion auf folgende Weise herangezogen werden. Es werden alle paarweisen Korrelationskoeffizienten $r_{ij}$ berechnet. Falls das zugehörige Konfidenzintervall die 0 nicht enthält ($0 \notin KI$), werden beide Ereignisse $i$ und $j$ ausgewählt. Als Schwellenwert $\tau_z$ wird in diesem Fall das Konfidenzniveau verwendet: $\tau_z = 1 - \alpha \in (0, 1)$.

### 5.1.6 Gewichte aus Edmonds' Algorithmus

Der Algorithmus von Edmonds spielt eine große Rolle beim Anpassen eines onkogenetischen Baums, siehe Abschnitt 2.3. Die dafür benötigten Gewichte $w_{ij}$ (siehe (2.3)) können zur Variablenselektion insofern herangezogen werden, als dass man die Ereignisse auswählen möchte, die mit großen Gewichten assoziiert sind. Dafür bestimmt man zunächst das Maximum aus $w_{ij}$ und $w_{ji}$, da ein angepasster Baum eher die Kante mit höherem Gewicht enthalten würde. Sei dies o.B.d.A. $w_{ij}$. Als nächstes werden anhand eines relativen Schwellenwertes $\tau_{\text{weight}} \in (0, 1)$ die $\lceil 100 \cdot \tau_{\text{weight}} \rceil$ Prozent größten Gewichte $w_{ij}$ bestimmt. Alle Ereignisse, die zu mindestens einem dieser Gewichte gehören, werden ausgewählt.

### 5.1.7 Bedingte Wahrscheinlichkeiten im angepassten Baum

Im Gegensatz zu allen bisher vorgestellten Variablenselektionsmethoden wird nun ein onkogenetischer Baum $T = (V, E, r, \alpha)$ an den gesamten Datensatz mit $n$ Ereignissen angepasst. Anschließend werden die Ereignisse ausgewählt, deren angrenzende Kanten eine ausreichend hohe bedingte Wahrscheinlichkeit aufweisen. Diese bedingten Wahrscheinlichkeiten werden oft auch als Kantengewichte bezeichnet. Alle Kanten $(i, j), (j, k) \in E$ werden als angrenzende Kanten zum Knoten bzw. Ereignis $j$ bezeichnet. Sei $\tau_{\text{OT}} \in (0, 1)$ die mindestens gewünschte bedingte Wahrscheinlichkeit. Ein Ereignis $j$ wird dann ausgewählt, falls

$$\max(\alpha(e), \alpha(f) : e = (i, j) \in E, f = (j, k) \in E) \geq \tau_{\text{OT}}. \tag{5.9}$$

Dabei ist zu beachten, dass die Kante $e$ eindeutig definiert ist, da alle Knoten (bis auf den Wurzelknoten $r$) exakt einen Elternknoten aufweisen, während jedoch für die Kante $f$ mehrere Kandidaten in Frage kommen, da jeder Knoten mehr als einen Nachfolger haben kann.

### 5.1.8 Unabhängigkeit im angepassten Baum

Für diese Methode wird wiederum ein onkogenetischer Baum an den gesamten Datensatz angepasst. Ereignisse, die in diesem Baummodell unabhängig von allen anderen sind, gehen direkt von der Wurzel aus und sind als einzelne Knoten ohne Nachfolger dargestellt. Diese unabhängigen Knoten werden aussortiert. Die restlichen Ereignisse bilden die Menge der ausgewählten Variablen.

Beachte, dass diese Variablenselektionsmethode nicht impliziert, dass unabhängige Ereignisse immer unnötig bzw. unwichtig für den Krankheitsverlauf sind.

### 5.1.9 Identifikation von Cliquen

Die letzten beiden Variablenselektionsmethoden basieren auf der Identifikation von Cliquen. Eine Clique $C$ ist ein Teilgraph eines ungerichteten Graphen $G_u = (V, E, w)$, in dem alle Knoten durch eine Kante miteinander verbunden sind. Die Idee, eine Clique mit bestimmten Eigenschaften zur Variablenselektion heranzuziehen, stammt von Desper et al. (1999).

Man beginnt mit einem vollständigen Graphen $G_c = (V, \tilde{E}, w)$, in dem alle $n$ Ereignisse paarweise durch eine Kante miteinander verbunden sind, d.h. $\tilde{E} = \{e = (i, j) : i, j \in 1, \ldots, n, i < j\}$. Als Kantengewichte $w$ werden die Gewichte $w_{ij}$ aus Edmonds' Branching Algorithmus herangezogen, siehe Abschnitt 2.3. Definiere $w : E \to \mathbb{R}_+$ mit $w(e) = w_{ij} + w_{ji}$, $e = (i, j)$. Durch das Aufsummieren der Kantengewichte werden beide Richtungen in den ungerichteten Graphen mit einbezogen. Aus dem vollständigen Graphen $G_c$ werden nun auf bestimmte Weise Kanten entfernt, so dass im resultierenden Graphen $G_u$ Cliquen identifiziert werden können. Desper et al. (1999) sind dabei so vorgegangen, dass jene Kanten $e = (i, j)$ entfernt werden, dessen zugehörige Knoten $i$ und $j$ nicht mindestens 5 mal gemeinsam im zugrunde liegenden Datensatz aufgetreten sind.

Im Gegensatz zu diesem absoluten Schwellenwert soll hier jedoch ein relativer definiert werden, so dass zwei Ereignisse mindestens mit einer bestimmten Wahrscheinlichkeit gemeinsam auftreten müssen, damit die Kante im Graphen erhalten bleibt. Ist $\tau_{\text{clique}} \in (0, 1)$ gegeben, wird eine Kante $e = (i, j)$ aus $G_c$ entfernt, falls

$$\frac{1}{m} \sum_{k=1}^{m} I((x_i^k = 1) \wedge (x_j^k = 1)) < \tau_{\text{clique}}, \tag{5.10}$$

wobei $I$ die Indikatorfunktion beschreibt. Sei $F$ die Menge dieser zu entfernenden Kanten, dann ist $E = \tilde{E} \backslash F$ die resultierende Menge von Kanten im ungerichteten Graphen $G_u$.

Ausgehend von $G_u$ werden nun zwei Variablenselektionsmethoden vorgestellt, die zum einen auf der größten Clique und zum anderen auf der maximalen Clique mit größtem Gewicht basieren. Ein anschauliches Beispiel zur Verdeutlichung des Unterschieds zwischen größten und maximalen Cliquen ist in Abbildung 5.1 gegeben.

Eine Clique $C$ wird größte Clique genannt, falls es keine andere Clique gibt, die mehr Knoten enthält. Die Ereignisse dieser größten Clique werden zur Anpassung des finalen Baummodells ausgewählt. Es ist möglich, dass $C$ nicht eindeutig ist. Es kann mehrere Cliquen mit der gleichen Anzahl an Ereignissen geben. In einem solchen Fall würden alle Ereignisse aus allen größten Cliquen ausgewählt.

Eine Clique $C$ wird als maximale Clique bezeichnet, wenn sie sich nicht zu einer größeren Clique erweitern lässt. Die größte(n) Clique(n) sind immer maximal, aber eine maximale Clique ist nicht notwendigerweise die größte. Zur Variablenselektion werden alle maximalen Cliquen $C_1, \ldots, C_q$ von $G_u$, $C_i = (V_i, E_i, w)$ bestimmt. Die maximale Clique mit größtem Gewicht ist dann

$$C := \arg\max_{C_i} \sum_{e \in E_i} w(e). \tag{5.11}$$

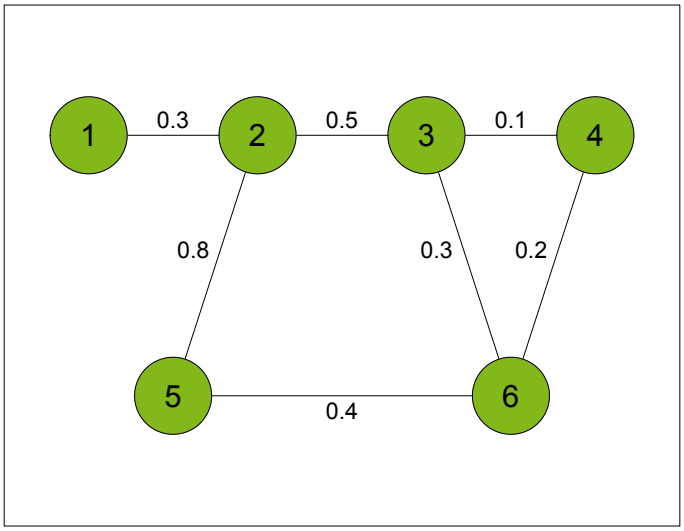



Abbildung 5.1: Anschauliches Beispiel, das den Unterschied zwischen größten und maximalen Cliquen verdeutlicht. Eine *Clique* ist ein Teilgraph, in dem alle Knoten durch eine Kante verbunden sind. In diesem Beispiel gibt es 8 verschiedene Cliquen: 1-2, 2-3, 2-5, 3-4, 3-6, 4-6, 5-6, 3-4-6. Die *größte Clique* ist die Clique mit den meisten Knoten, in diesem Beispiel die Clique 3-4-6, da es keine andere Clique mit mehr als 3 Knoten gibt. Eine *maximale Clique* kann nicht zu einer größeren Clique erweitert werden. In diesem Beispiel gibt es 5 maximale Cliquen: 1-2, 2-3, 2-5, 5-6, 3-4-6. Die Clique 3-4 ist keine maximale Clique, da durch Hinzufügen von Knoten 6 immer noch eine Clique gegeben ist. Die Kantengewichte sind nicht notwendig, um größte oder maximale Cliquen zu identifizieren, sie werden jedoch benötigt, um die *maximale Clique mit größtem Gewicht* zu benennen. Dabei handelt es sich um die maximale Clique mit der größten Summe der Kantengewichte. In diesem Beispiel handelt es sich dabei um die Clique 2-5 mit einem Gewicht von 0.8. Die größte Clique 3-4-6 besitzt nur ein Gewicht von 0.6.

Die Menge $V_i$ von Knoten dieser maximalen Clique mit größtem Gewicht beschreibt dann die Menge der ausgewählten Ereignisse.

## 5.2 Vergleich der Variablenselektionsverfahren anhand einer Simulationsstudie

In diesem Abschnitt sollen die 10 vorgestellten Variablenselektionsverfahren mit Hilfe einer Simulationsstudie ausgewertet werden. Zunächst wird dazu in Abschnitt 5.2.1 der Aufbau der Simulationsstudie beschrieben. Anschließend werden in Abschnitt 5.2.2 und 5.2.3 die Ergebnisse für zwei unterschiedliche Gütekriterien vorgestellt. Für jede Methode wird dabei als erstes der beste Schwellenwert bestimmt, bevor alle Methoden untereinander verglichen werden. In Abschnitt 5.2.4 wird abschließend darauf eingegangen, wie bzw. ob Rauschen in den Originaldaten die Ergebnisse der Simulationsstudie verändert.

### 5.2.1 Aufbau der Simulation

Um die Variablenselektionsverfahren beurteilen zu können, wird folgender Aufbau für die Simulationsstudie gewählt:

1. Ziehe einen zufälligen onkogenetischen Baum $T$ mit $n_1$ Ereignissen.
2. Ziehe $m$ Beobachtungen aus $T$ und erhalte die Datenmatrix $X \in \mathbb{B}^{m \times n_1}$.
3. Ziehe zusätzlich $m$ Beobachtungen aus $Y_i \sim \text{Bin}(1, \pi_i)$, mit $\pi_i \in (0,1)$, $i = 1, \dots, n_2$.
4. Fasse die Daten aus Schritt (2) und (3) zusammen zur Datenmatrix $\tilde{X} \in \mathbb{B}^{m \times (n_1+n_2)}$.
5. Wende ein Variablenselektionsverfahren auf $\tilde{X}$ an und erhalte die Datenmatrix $X^*$, die nur die ausgewählten Ereignisse enthält.
6. Passe einen onkogenetischen Baum $T^*$ an $X^*$ an.
7. Vergleiche $T^*$ und $T$.
8. Vergleiche $X^*$ und $X$.

Der onkogenetische Baum $T$ ist das zugrunde liegende wahre Modell. Dieser Baum wird in Schritt 1 zufällig erzeugt, wobei die Anzahl $n_1$ an Ereignissen und das Intervall $[\alpha_l, \alpha_u]$ $(0 < \alpha_l < \alpha_u < 1)$ für die Kantengewichte bzw. bedingten Wahrscheinlichkeiten fest vorgegeben ist. Im nächsten Schritt wird eine Datenmatrix $X = [\mathbf{x_1}, \dots, \mathbf{x_{n_1}}]$ generiert, die $m$ Beobachtungen aus dem Baum $T$ enthält. Idealerweise werden am Ende diese $n_1$ Ereignisse von unseren Variablenselektionsverfahren wiedergefunden. Um den Auswahlprozess schwieriger und realistischer zu machen, werden zusätzlich Beobachtungen aus einer Binomialverteilung mit Parameter $\pi_i$ für $n_2$ weitere Ereignisse gezogen, siehe Schritt 3. Diese Ereignisse sollen das zufällige Rauschen darstellen, das in einem echten Datensatz neben den wahren Ereignissen immer auftritt. Für jede Rauschvariable wird der Parameter $\pi_i$ unabhängig bestimmt. Als nächstes werden wahre und zufällige Ereignisse zu einer einzigen Datenmatrix $\tilde{X} \in \mathbb{B}^{m \times (n_1+n_2)}$ zusammengefasst. Anschließend wird in Schritt 5 eine Variablenselektionsmethode auf diese Daten angewendet. Jede Methode wählt $p \leq n_1 + n_2$ Spalten von $\tilde{X}$ aus. Diese Auswahl wird durch $X^*$ gekennzeichnet und man kann einen onkogenetischen Baum $T^*$ an diesen Datensatz anpassen.

Um die Güte der Variablenselektionsverfahren zu beurteilen, werden der wahre und der angepasste Baum miteinander verglichen, sowie die wahren und die ausgewählten Ereignisse.

Ein Vergleich von zwei verschiedenen Baummodellen kann mit Hilfe der induzierten Wahrscheinlichkeitsverteilung durchgeführt werden. Dies wurde bereits ausführlich in Abschnitt 3.2.2 vorgestellt, wird hier jedoch kurz wiederholt. Gegeben seien zwei onkogenetische Bäume $T_1$ und $T_2$, jeder mit $n$ Ereignissen. Die zwei Wahrscheinlichkeitsvektoren für die $2^n$ Ereigniskombinationen werden mit $\mathbf{p_1}$ und $\mathbf{p_2} \in [0,1]^{2^n}$ bezeichnet. Abstände zwischen diesen beiden Vektoren und damit zwischen den beiden Baummodellen können dann über die $L_1$-, $L_2$- oder Cosinus-Distanz berechnet werden:

$$d_{L_1}(\mathbf{p_1}, \mathbf{p_2}) = \sum_{i=1}^{2^n} |p_{1_i} - p_{2_i}|, \tag{5.12}$$

$$d_{L_2}(\mathbf{p_1}, \mathbf{p_2}) = \sqrt{\sum_{i=1}^{2^n} (p_{1_i} - p_{2_i})^2} \tag{5.13}$$

$$d_{Cos}(\mathbf{p_1}, \mathbf{p_2}) = 1 - \cos \sphericalangle(\mathbf{p_1}, \mathbf{p_2}) = 1 - \frac{< \mathbf{p_1}, \mathbf{p_2} >}{||\mathbf{p_1}|| \, ||\mathbf{p_2}||} \tag{5.14}$$

$$= 1 - \frac{\sum_{i=1}^{2^n} p_{1_i} \cdot p_{2_i}}{\sqrt{(\sum_{i=1}^{2^n} p_{1_i}^2) \cdot (\sum_{i=1}^{2^n} p_{2_i}^2)}} \tag{5.15}$$

Sollen diese Distanzmaße in oben beschriebener Simulationsstudie angewendet werden, ist zu beachten, dass die Bäume $T$ und $T^*$ aufgrund des Auswahlprozesses unterschiedliche Ereignisse enthalten können. Die Anzahl dieser Ereignisse kann ebenfalls unterschiedlich sein. Daher müssen alle $n_1 + n_2$ Ereignisse beim Berechnen der induzierten Wahrscheinlichkeitsverteilung berücksichtigt werden. Ereigniskombinationen, die dabei ein Ereignis enthalten, das gar nicht im zugehörigen Baummodell vorkommt, erhalten die Wahrscheinlichkeit 0.

Neben dem Vergleich der Baummodelle sollen auch die ausgewählten Ereignisse mit den wahren verglichen werden. Dazu werden zwei verschiedene Kennzahlen betrachtet:

- Sensitivität: Welcher Anteil wahrer Ereignisse wird erkannt und damit ausgewählt?
- Spezifität: Welcher Anteil Störereignisse wird aussortiert?

Ein gutes Variablenselektionsverfahren sollte nur die wahren $n_1$ Ereignisse auswählen.

Im oben beschriebenen Evaluationsprozess müssen noch einige Parameter definiert werden. Dazu gehören die Anzahl $n_1$ an wahren Ereignissen, die Anzahl $n_2$ an zufälligen Stör-Ereignissen, die Anzahl $m$ an Beobachtungen, das Intervall $[\alpha_l, \alpha_u]$ für die Kantengewichte und die Wahrscheinlichkeit $\pi_i$ für den Anteil des Rauschens.

Anhand dieser Parameter können Datensituationen mit unterschiedlichem Schwierigkeitsgrad erzeugt und mit den Variablenselektionsmethoden untersucht werden. Für diese Simulationsstudie werden jeweils zwei Ausprägungen für jeden Parameterwert gewählt (der Parameter $\pi_i$ wird dabei für jede der $n_2$ Rauschvariablen zufällig und unabhängig aus dem entsprechenden Intervall gezogen):

$$\begin{aligned} n_1 &\in \{5, 7\} \\ n_2 &\in \{2, 12\} \\ m &\in \{50, 1000\} \\ [\alpha_l, \alpha_u] &\in \{[0.2, 0.8], [0.5, 0.8]\} \\ \pi &\in \{[0, 0.2], [0.2, 0.4]\} \end{aligned}$$

Es ergeben sich somit 32 unterschiedliche Datensituationen, die in Tabelle 5.2 aufgeführt sind.

Tabelle 5.2: Liste der 32 Parameterkombinationen, die die unterschiedlichen Datensituationen charakterisieren, die von den Variablenselektionsmethoden untersucht werden. (Für $\pi_i$ wird der Einfachheit halber nicht das Intervall, sondern der Erwartungswert angegeben.)

| | $m$ | $n_1$ | $n_2$ | $E(\pi_i)$ | $\alpha_l$ |
|---|---|---|---|---|---|
| 1 | 50 | 5 | 12 | 0.1 | 0.2 |
| 2 | 1000 | 5 | 12 | 0.1 | 0.2 |
| 3 | 50 | 7 | 12 | 0.1 | 0.2 |
| 4 | 1000 | 7 | 12 | 0.1 | 0.2 |
| 5 | 50 | 5 | 12 | 0.3 | 0.2 |
| 6 | 1000 | 5 | 12 | 0.3 | 0.2 |
| 7 | 50 | 7 | 12 | 0.3 | 0.2 |
| 8 | 1000 | 7 | 12 | 0.3 | 0.2 |
| 9 | 50 | 5 | 12 | 0.1 | 0.5 |
| 10 | 1000 | 5 | 12 | 0.1 | 0.5 |
| 11 | 50 | 7 | 12 | 0.1 | 0.5 |
| 12 | 1000 | 7 | 12 | 0.1 | 0.5 |
| 13 | 50 | 5 | 12 | 0.3 | 0.5 |
| 14 | 1000 | 5 | 12 | 0.3 | 0.5 |
| 15 | 50 | 7 | 12 | 0.3 | 0.5 |
| 16 | 1000 | 7 | 12 | 0.3 | 0.5 |
| 17 | 50 | 5 | 2 | 0.1 | 0.2 |
| 18 | 1000 | 5 | 2 | 0.1 | 0.2 |
| 19 | 50 | 7 | 2 | 0.1 | 0.2 |
| 20 | 1000 | 7 | 2 | 0.1 | 0.2 |
| 21 | 50 | 5 | 2 | 0.3 | 0.2 |
| 22 | 1000 | 5 | 2 | 0.3 | 0.2 |
| 23 | 50 | 7 | 2 | 0.3 | 0.2 |
| 24 | 1000 | 7 | 2 | 0.3 | 0.2 |
| 25 | 50 | 5 | 2 | 0.1 | 0.5 |
| 26 | 1000 | 5 | 2 | 0.1 | 0.5 |
| 27 | 50 | 7 | 2 | 0.1 | 0.5 |
| 28 | 1000 | 7 | 2 | 0.1 | 0.5 |
| 29 | 50 | 5 | 2 | 0.3 | 0.5 |
| 30 | 1000 | 5 | 2 | 0.3 | 0.5 |
| 31 | 50 | 7 | 2 | 0.3 | 0.5 |
| 32 | 1000 | 7 | 2 | 0.3 | 0.5 |

Zusätzlich müssen für 8 der 10 Variablenselektionsmethoden passende Schwellenwerte festgelegt werden. Dazu werden für jede Methode vier verschiedene Werte ausgewählt.

$$\begin{aligned}
\tau_{\text{freq}} &\in \{0.05, 0.10, 0.15, 0.20\} \\
\tau_{\text{cor}} &\in \{0.10, 0.20, 0.30, 0.40\} \\
\tau_{\text{fisher}} &\in \{0.01, 0.05, 0.10, 0.15\} \\
\tau_{\text{z}} &\in \{0.50, 0.63, 0.77, 0.90\} \\
\tau_{\text{weight}} &\in \{0.05, 0.10, 0.20, 0.30\} \\
\tau_{\text{OT}} &\in \{0.10, 0.15, 0.20, 0.25\} \\
\tau_{\text{clique}} &\in \{0.05, 0.10, 0.15, 0.20\}
\end{aligned}$$

Für jede Parameterkombination werden $M = 100$ zufällige onkogenetische Bäume mit zugehörigen Datensätzen erzeugt. Es werden zehn verschiedene Variablenselektionsverfahren mit jeweils vier verschiedenen Schwellenwerten angewendet. Ausnahme sind dabei die Methode von Brodeur, bei der der Schwellenwert implizit berechnet wird, und die Methode der Unabhängigkeit im Baum, die keinen Schwellenwert benötigt. Basierend auf diesen Simulationsergebnissen werden die Variablenselektionsverfahren bewertet.

Alle Methoden sowie die Simulationsstudie sind in der statistischen Programmiersprache R in der Version 3.3.3 implementiert (R Core Team, 2017). Es werden die R-Pakete `Rtreemix` (Bogojeska et al., 2008) und `igraph` (Csardi und Nepusz, 2006) verwendet, um onkogenetische Bäume anzupassen bzw. die Cliquen-Berechnungen durchzuführen.

Im folgenden Abschnitt 5.2.2 werden die zehn verschiedenen Variablenselektionsmethoden anhand der $L_1$-Distanz verglichen, die den Unterschied zwischen wahrem und angepasstem Baummodell angibt. Ein weiteres Kriterium für die Güte der Variablenselektionsmethoden ist die Anzahl der richtig bzw. falsch entfernten Ereignisse. Dies wird in Abschnitt 5.2.3 betrachtet.

### 5.2.2 Auswertung basierend auf der $L_1$-Distanz

#### Vorarbeiten

Da für fast jede Variablenselektionsmethode vier verschiedene Schwellenwerte verwendet wurden, besteht der erste Schritt der Auswertung darin, für jede Methode den besten Schwellenwert zu benennen.

Für die Methode der univariaten Häufigkeit ist die Ergebnisgrafik in Abbildung 5.2 (mitte links) dargestellt. Auf der x-Achse sind die 32 Parameterkombinationen aufgeführt, die y-Achse zeigt den Mittelwert der 100 $L_1$-Distanzen zwischen wahrem und angepasstem Baummodell. Die vier unterschiedlichen Linien repräsentieren die vier verschiedenen Schwellenwerte.

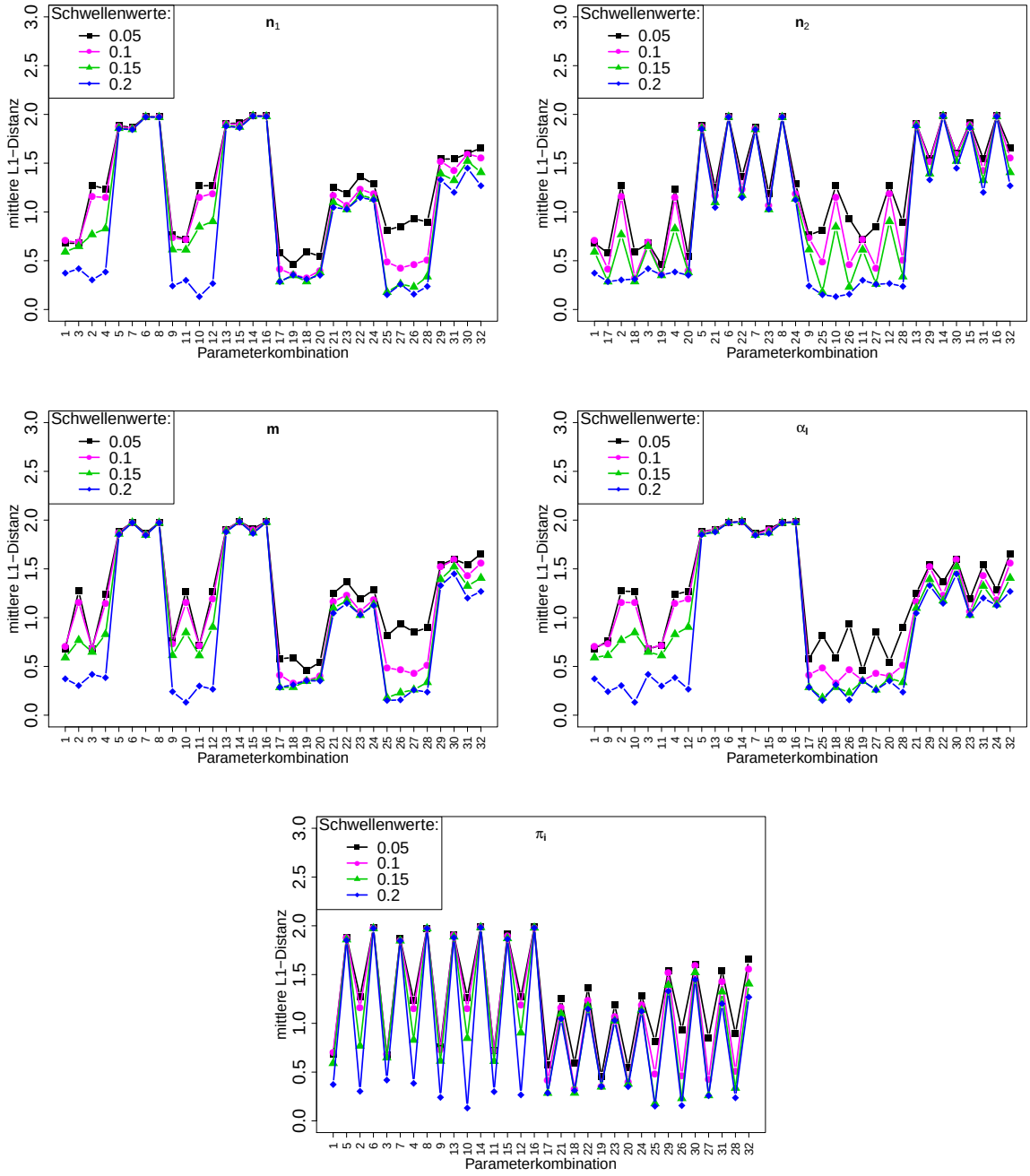



Abbildung 5.2: Ergebnisse der Simulationsstudie für die Methode der univariaten Häufigkeit. Auf der x-Achse sind die 32 Parameterkombinationen, auf der y-Achse die mittleren $L_1$-Distanzen zwischen wahrem und angepasstem Modell dargestellt. Die Unterschiede zwischen den Grafiken bestehen lediglich in der Reihenfolge der Parameterkombinationen auf der x-Achse. Durch diese unterschiedliche Darstellung der gleichen Ergebnisse kann verdeutlicht werden, dass nicht alle fünf Parameterwerte den gleichen Einfluss haben. Für jeden Parameterwert einzeln wurde die Reihenfolge der Parameterkombinationen so angepasst, dass sich zwei aufeinanderfolgende Punkte nur in dem einen speziellen Parameter unterscheiden.

Es fällt sofort ins Auge, dass sich die Ergebnisse mancher Parameterkombinationen sehr ähnlich sind, manchmal aber auch sehr deutliche Unterschiede auftreten. Bevor daher auf die konkrete Auswertung dieser Variablenselektionsmethode eingegangen wird, soll zunächst geklärt werden, ob eventuell einige der Parameterkombinationen immer sehr ähnliche und somit redundante Ergebnisse aufweisen. Ist dies der Fall, können weniger als 32 Datensituationen betrachtet werden.

Um sich einen Überblick darüber zu verschaffen, welche Parameterkombinationen ähnliche Ergebnisse liefern, werden diese so sortiert, dass zwei aufeinanderfolgende Punkte immer in vier der fünf Parameterwerte übereinstimmen. Dies soll für alle fünf Parameter untersucht werden. In Abbildung 5.2 sind daher fünf Grafiken dargestellt. In der ersten Grafik oben links ist es der Parameter $n_1$, der für jeweils 2 aufeinanderfolgende Punkte unterschiedlich ist. Es handelt sich dabei um die Parameterkombinationen mit der Nummer 1 und 3. Das lässt sich auch mit Hilfe der Tabelle 5.2 überprüfen und nachvollziehen. Für den Parameter $n_2$ (Grafik oben rechts) werden als ersten die Kombinationen 1 und 17 miteinander verglichen. Die restlichen drei Grafiken für die Parameter $m$, $\alpha_l$ und $\pi_i$ sind genauso aufgebaut.

Ist es für einen der fünf Parameterwerte also überflüssig, unterschiedliche Werte zu wählen, liefern in der entsprechenden Grafik jeweils zwei aufeinanderfolgende Punkte sehr ähnliche Ergebnisse.

Anhand der Grafik oben links kann demnach überprüft werden, ob es eine Rolle spielt, dass mal 5 und mal 7 wahre Ereignisse zugrunde liegen. In den meisten Fällen ist es sehr deutlich, dass die Ergebnisse für $n_1 = 5$ und $n_1 = 7$ sehr ähnlich sind. Kleine Ausnahmen gibt es z.B. für die Schwellenwerte 0.15 und 0.2 bei den Parameterkombinationen 29 bis 32. Insgesamt lässt sich aber folgern, dass es keinen sehr großen Unterschied macht, ob das wahre Baummodelle 5 oder 7 Ereignisse enthält.

Für die Anzahl $n_2$ der zusätzlichen Rauschvariablen (Abbildung 5.2 oben rechts) sieht das Bild schon ganz anders aus. Hier unterscheiden sich die aufeinanderfolgenden Punkte in den meisten Fällen sehr stark. Das Ergebnis hängt also deutlich davon ab, ob dem wahren Datensatz nur wenige oder viele Störvariablen hinzugefügt werden.

Für den Parameter $m$ (Grafik mitte links), der die Anzahl der Beobachtungen angibt, schwanken die Ergebnisse für die Schwellenwerte 0.05 und 0.1 stärker. Aber auch für die anderen Parameter sind teilweise Unterschiede zu erkennen. Auch bei der Wahl der unteren Grenze $\alpha_l$ (Grafik mitte rechts) unterscheiden sich aufeinanderfolgende Punkte voneinander, meistens jedoch nicht sehr deutlich. Die Grafik, die in Abbildung 5.2 ganz unten dargestellt ist, zeigt sehr deutlich, dass es stark darauf ankommt, mit welcher Wahrscheinlichkeit $\pi_i$ die Störvariablen auftreten.

Insgesamt lässt sich somit folgern, dass der Parameter $n_1$ vernachlässigt werden kann, wohingegen der Parameter $\pi_i$ eine wichtige Rolle dabei spielt, ob es den Variablenselektionsverfahren besser oder schlechter gelingt, das wahre Modell zu identifizieren.

Damit nicht für alle 10 Variablenselektionsverfahren diese fünf Grafiken betrachtet werden müssen, werden die mittleren $L_1$-Distanzen für alle Methoden, Parameterkombinationen und

Schwellenwerte geclustert. Das Ergebnis dieses Cluster Verfahrens mit dem complete-Linkage Ansatz ist durch ein Dendrogramm in Abbildung 5.3 dargestellt.

Es ist sehr gut zu erkennen, dass die erste Vereinigung (roter Kasten) immer zwischen den Parameterkombinationen stattfindet, bei denen sich nur die Anzahl $n_1$ der wahren Ereignisse unterscheidet. Das deckt sich gut mit den vorangegangenen Beobachtungen zur Methode der univariaten Häufigkeit. Ob es 5 oder 7 wahre Ereignisse sind, ändert nicht sehr viel am Ergebnis[1].

Die nachfolgende zweite Vereinigung von Parameterkombinationen erfolgt in den meisten Fällen über $\alpha_I$ (blauer Kasten). In zwei Fällen erfolgt zunächst eine Vereinigung über den Parameter $m$, der Abstand zur Vereinigung über $\alpha_I$ ist jedoch sehr gering (diese Fälle sind durch mittelblaue Kästen gekennzeichnet). Über alle Variablenselektionsmethoden hinweg betrachtet, spielt also die untere Grenze für die bedingte Wahrscheinlichkeit bei den Kantengewichten ebenfalls keine so große Rolle. Bei Betrachtung der Methode der univariaten Häufigkeiten ist das teilweise auch schon deutlich geworden.

Insgesamt lässt sich aus dem Cluster-Dendrogramm also folgern, dass die Parameter $n_1$ und $\alpha_I$ vernachlässigt werden können, da die Unterschiede in den resultierenden $L_1$-Distanzen gering sind. Die Werte werden daher auf

$$n_1 = 5$$
$$\alpha_I = 0.2$$

festgesetzt. Statt 32 Parameterkombinationen werden für die Auswertung und den Vergleich der Variablenselektionsverfahren nur noch insgesamt 8 verschiedene Datensituationen betrachtet, die in Tabelle 5.3 dargestellt sind.

Es wird ebenfalls auf die Betrachtung der $L_2$- und Cosinus-Distanz verzichtet, da wie auch schon in Kapitel 3 die Ergebnisse aller drei Distanzmaße sehr ähnlich sind.

#### Ergebnisse: Was ist der beste Schwellenwert?

Nachdem im vorangegangenen Abschnitt die zu untersuchenden Parameterkombinationen von 32 auf 8 reduziert worden sind, soll nun für jede Methode der beste Schwellenwert gefunden werden. Ausführlich soll dies am Beispiel der univariaten Häufigkeit und der größten Cliquen gezeigt werden. Die Ergebnisse dieser beiden Variablenselektionsverfahren sind in Abbildung 5.4 zu finden. Auf der x-Achse sind die 8 Parameterkombinationen und auf der y-Achse die mittleren $L_1$-Distanzen zwischen wahrem und angepasstem Modell dargestellt. Der Einfachheit halber wird in dieser und allen folgenden Grafiken für den Parameter $\pi_i$ nicht das Intervall, sondern

[1] Insgesamt ist der Unterschied zwischen 5 und 7 natürlich nicht sehr groß, so dass dieses Ergebnis zu erwarten war. Aufgrund von Rechenzeitbeschränkungen beim Berechnen der induzierten Wahrscheinlichkeitsverteilung konnten jedoch nicht mehr als 7 wahre Ereignisse gewählt werden, wenn gleichzeitig 12 Störereignisse simuliert werden sollten. Für die Analyse und Bewertung von Variablenselektionsverfahren ist es jedoch wichtiger, Situationen herzustellen, die eine sehr unterschiedliche Anzahl an Störereignissen aufweisen, als einen großen Unterschied in der Anzahl wahrer Ereignisse.

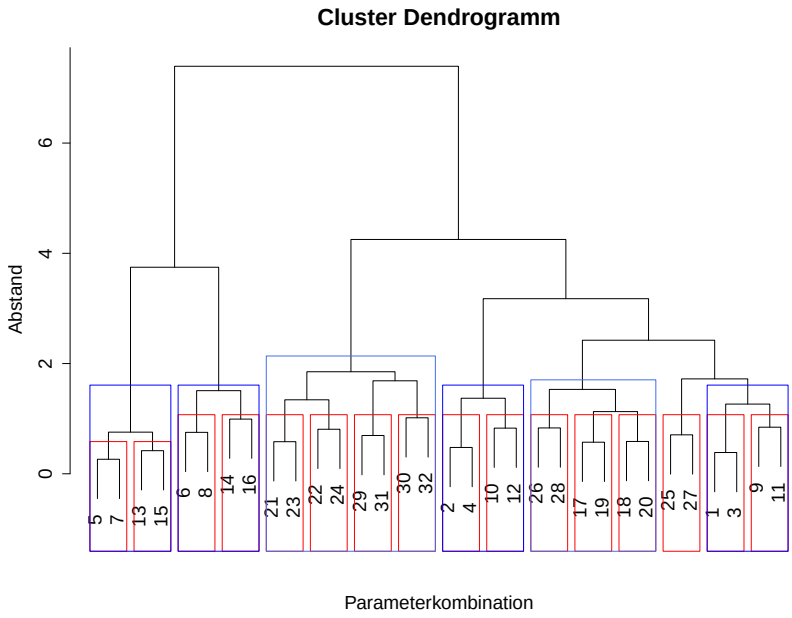



Abbildung 5.3: Cluster Dendrogramm der mittleren $L_1$-Distanzen unter Verwendung des complete-Linkage Verfahrens um die Anzahl der Parameterkombinationen möglicherweise zu reduzieren. Die Parameterkombinationen der ersten Vereinigung (roter Kasten) unterscheiden sich nur in der Anzahl $n_1$ der wahren Ereignisse. Die der zweiten Vereinigung sind meistens die, bei denen $\alpha_l$ eine Rolle spielt (blauer Kasten). Die 32 Parameterkombinationen sind die aus Tabelle 5.2.

Tabelle 5.3: Liste der 8 Parameterkombinationen, die die unterschiedlichen Datensituationen charakterisieren, anhand derer die Variablenselektionsverfahren ausgewertet werden.

| | $m$ | $n_2$ | $E(\pi_i)$ |
|---|---|---|---|
| 1 | 50 | 2 | 0.1 |
| 2 | 1000 | 2 | 0.1 |
| 3 | 50 | 12 | 0.1 |
| 4 | 1000 | 12 | 0.1 |
| 5 | 50 | 2 | 0.3 |
| 6 | 1000 | 2 | 0.3 |
| 7 | 50 | 12 | 0.3 |
| 8 | 1000 | 12 | 0.3 |

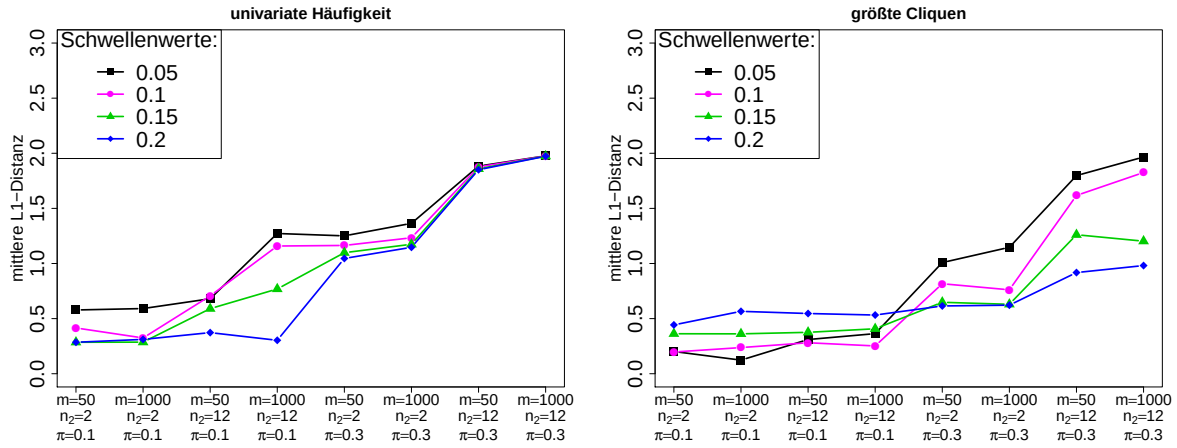



Abbildung 5.4: Ergebnisse der Simulationsstudie für die Variablenselektionsverfahren der univariaten Häufigkeit (links) und größten Cliquen (rechts). Auf der x-Achse sind die 8 Parameterkombinationen, auf der y-Achse die mittleren $L_1$-Distanzen zwischen wahrem und angepasstem Modell dargestellt.

der Erwartungswert angegeben, d.h. für das Intervall $[0, 0.2]$ der Wert 0.1 und für das Intervall $[0.2, 0.4]$ der Wert 0.3. Die vier verschiedenfarbigen Linien stehen für die vier unterschiedlichen Schwellenwerte, die pro Verfahren untersucht wurden.

Im linken Teil der Abbildung sind die Ergebnisse der univariaten Häufigkeit abgebildet. Man kann erkennen, dass für die ersten vier Parameterkombinationen, für die $\pi_i \in [0, 0.2]$ gilt, der Abstand zwischen wahrem und angepasstem Modell kleiner ist als für $\pi_i \in [0.2, 0.4]$. Dies war zu erwarten, da zum einen eine höhere Wahrscheinlichkeit für das Auftreten der Störvariablen eine insgesamt schwierigere Situation darstellt, zum anderen der größte Schwellenwert nur bei $\tau_{\text{freq}} = 0.2$ liegt. Damit liegt der Schwellenwert deutlich unter der Wahrscheinlichkeit des Rauschens, so dass diese Ereignisse durch die Variablenselektion nicht entfernt werden.

Insgesamt liefert die Wahl von $\tau_{\text{freq}} = 0.2$ für alle Parameterkombinationen die besten bzw. sehr gute Ergebnisse. Ein größerer Schwellenwert würde natürlich die Ergebnisse für $\pi_i \in [0.2, 0.4]$ verbessern, ist aber unrealistisch für die meisten Anwendungen. Unabhängig von der Kenntnis der Daten ist somit der Schwellenwert $\tau_{\text{freq}} = 0.2$ die beste Wahl.

Auch für die Methode der größten Cliquen (Abbildung 5.4, rechts) gilt, dass für die Datensituationen mit größerem Rauschen die Abstände zum wahren Modell größer sind. Ein global bester Schwellenwert für alle 8 Datensituationen lässt sich hier jedoch nicht angeben. Betrachtet man die Situationen, in denen der Anteil $\pi_i$ der Rauschvariablen im Intervall $[0, 0.2]$ liegt, liefert in den meisten Fällen der kleinste Schwellenwert das beste und der größte das schlechteste Ergebnis. Für den größeren Anteil der Rauschvariablen ($\pi_i \in [0.2, 0.4]$ bzw. $E(\pi_i) = 0.3$) ist die Reihenfolge exakt umgekehrt und der größte Schwellenwert liefert die besten Ergebnisse. Das bedeutet, dass die Wahl des besten Schwellenwertes abhängig ist von der Wahrscheinlichkeit des Auftretens von Rauschvariablen, die den Krankheitsverlauf nicht beeinflussen.

Für die weiteren 6 Variablenselektionsverfahren sind die Ergebnisse in Abbildung 5.5 dargestellt. Auch hier werden anhand der Grafiken die besten Schwellenwerte identifiziert.

Für die Methode der paarweise Korrelation ist bis auf eine Ausnahme $\tau_{\text{cor}} = 0.3$ am besten bzw. weicht nur minimal vom besten Ergebnis ab. Beim exakten Test von Fisher liefert in fünf der acht Situationen der kleinste Schwellenwert von $\tau_{\text{fisher}} = 0.01$ die besten Ergebnisse. In weiteren zwei Situationen liegen alle Ergebnisse sehr nah beieinander, so dass es gut vertretbar ist, auch hier diesen kleinen Schwellenwert zu verwenden. Für die z-Transformation ist eindeutig $\tau_{\text{z}} = 0.9$ als Schwellenwert zu wählen. Bei den Gewichten von Edmonds gibt es keine eindeutige Wahl. Wie bei den größten Cliquen muss situationsabhängig entschieden werden. Diesmal ist jedoch die Anzahl der Störvariablen ausschlaggebend. In den Fällen von nur wenigen Störvariablen ($n_2 = 2$) liefert der größte Schwellenwert $\tau_{\text{weight}} = 0.3$ die besten Ergebnisse, während für $n_2 = 12$ eher $\tau_{\text{weight}} = 0.05$ gewählt werden sollte. Bei der Methode, die die bedingten Kantenwahrscheinlichkeiten im angepassten Baum berücksichtigt, ist der größte Schwellenwert $\tau_{\text{OT}} = 0.25$ am besten. Und für die maximalen Cliquen ergibt sich ein ähnliches Bild wie bei den größten Cliquen. Es muss nach der Wahrscheinlichkeit des Rauschens unterschieden werden, allerdings liefert hier bei großem Rauschen ($\pi_i \in [0.2, 0.4]$) nicht der größte, sondern zweitgrößte Schwellenwert $\tau_{\text{mcliq}} = 0.15$ die besseren Ergebnisse.

Insgesamt wird also folgende Wahl für die besten Schwellenwerte getroffen:

$$
\begin{aligned}
\tau_{\text{freq}} &= 0.2 \\
\tau_{\text{cor}} &= 0.3 \\
\tau_{\text{fisher}} &= 0.01 \\
\tau_{\text{z}} &= 0.9 \\
\tau_{\text{weight}} &\in \{0.05, 0.30\} \\
\tau_{\text{OT}} &= 0.25 \\
\tau_{\text{lcliq}} &\in \{0.05, 0.20\} \\
\tau_{\text{mcliq}} &\in \{0.05, 0.15\}
\end{aligned}
$$

Für die Methode der Unabhängigkeit im Baum wird generell kein Schwellenwert benötigt. Bei der Methode von Brodeur wird dieser implizit berechnet und ist dann genauso wie der Schwellenwert bei der univariaten Häufigkeit zu verstehen. Für die acht Datensituationen lauten die mittleren Schwellenwerte (und in Klammern die Standardabweichungen) der je 100 Durchläufe wie folgt: $0.37(0.11), 0.25(0.08), 0.30(0.04), 0.18(0.04), 0.44(0.09), 0.33(0.08), 0.48(0.04)$ und $0.33(0.03)$. Diese sind somit bis auf eine Ausnahme immer höher als der Schwellenwert, der bei der univariaten Häufigkeit gewählt wurde. Ob sich dies positiv oder negativ auf das Ergebnis auswirkt, wird im folgenden Abschnitt beantwortet.

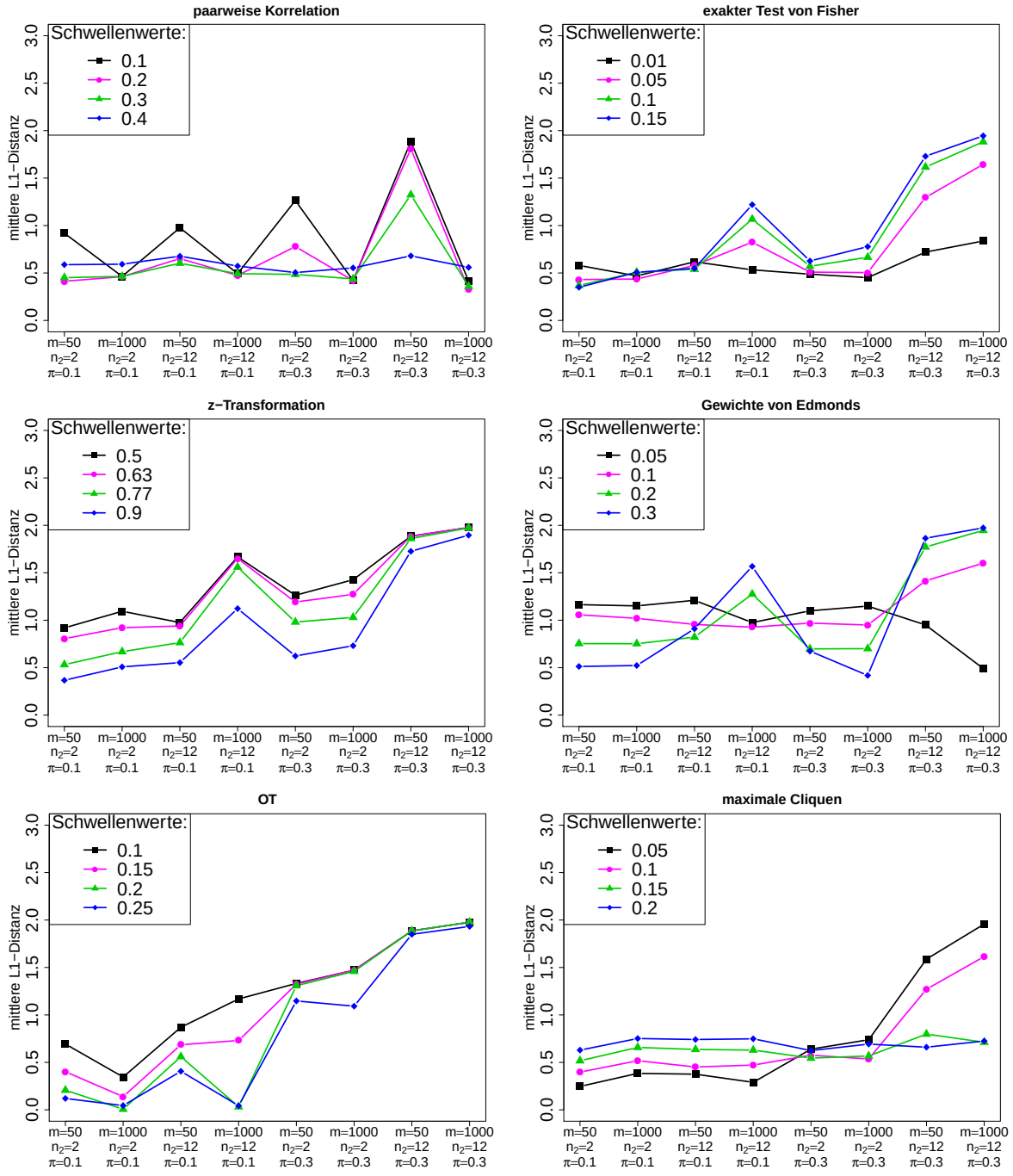



Abbildung 5.5: Ergebnisse der restlichen sechs Variablenselektionsverfahren anhand derer der jeweils beste Schwellenwert identifiziert werden kann.

**Ergebnisse: Was ist die beste Methode?**

Da im vorangegangenen Abschnitt für jede Methode der beste Schwellenwert bestimmt worden ist, können nun die Methoden an sich miteinander verglichen werden. Der Übersichtlichkeit halber werden zunächst die Methoden mit einem eindeutig besten und die Methoden mit einem situationsabhängigen Schwellenwert getrennt voneinander betrachtet, siehe Abbildung 5.6. Der mittlere Standardfehler für die Daten dieser beiden Grafiken beträgt 0.038.

In der linken Grafik ist zu erkennen, dass die Methode der z-Transformation nie die besten Ergebnisse erzielt. Die Methoden der paarweisen Korrelation sowie der Unabhängigkeit im Baum gehören in zwei Datensituationen zwar zu den besten Methoden (direkt gefolgt vom Fisher-Test), sind aber in allen anderen Fällen deutlich schlechter. Diese drei Methoden werden somit nicht weiter betrachtet. Ist die Wahrscheinlichkeit für Rauschen eher klein ($\pi_i \in [0, 0.2]$) sind die besten zwei Methoden die, die die bedingten Wahrscheinlichkeiten im angepassten Baum betrachten (OT) und die univariate Häufigkeit. Für $\pi_i \in [0.2, 0.4]$ sollte jedoch der Fisher-Test zur Variablenselektion herangezogen werden.

Die rechte Grafik in Abbildung 5.6 zeigt die Ergebnisse der beiden Cliquen Methoden und die der Gewichte von Edmonds' Algorithmus. Für $\pi_i \in [0, 0.2]$ ist die Methode der größten Cliquen mit Schwellenwert 0.05 am besten, dicht gefolgt von den maximalen Cliquen (ebenfalls Schwellenwert 0.05). Für die höhere Wahrscheinlichkeit der Störvariablen ($\pi_i \in [0.2, 0.4]$) führt die Methode, die sich auf den Algorithmus von Edmonds bezieht, in zwei Situationen zu den kleinsten $L_1$-Distanzen. Allerdings müsste für diese Methode zusätzlich nach der Anzahl der Störvariablen unterschieden werden. Möchte man dies unberücksichtigt lassen, sind wieder die beiden Cliquenverfahren am besten, wobei hier der jeweils höhere Schwellenwert gewählt werden sollte.

In Abbildung 5.7 sind nun zusammenfassend die jeweils besten Methoden für die Situationen mit geringer und höherer Wahrscheinlichkeit der Störvariablen dargestellt. Für $\pi_i \in [0, 0.2]$ liefern die Methoden der größten Cliquen bzw. die der bedingten Wahrscheinlichkeiten im angepassten Baum (OT) die kleinsten Abstände zum wahren Modell.

Ist die Wahrscheinlichkeit für das Rauschen höher ($\pi_i \in [0.2, 0.4]$), sind auch generell die Abstände zum wahren Modell größer, da es sich um schwierigere Datensituationen handelt. Die beiden Cliquenmethoden sowie der Fisher-Test liefern ähnlich gute Ergebnisse.

Insgesamt sind also die Cliquen-Methoden zur Variablenselektion am besten geeignet. Sie erzielen nicht in allen Einzelfällen das beste Ergebnis, liefern aber für alle betrachteten Datensituationen immer sehr gute und damit geringe Abstände zum wahren Modell. Für $\pi_i \in [0, 0.2]$ ist die Methode der größten Cliquen etwas besser, für $\pi_i \in [0.2, 0.4]$ die der maximalen, es bestehen aber keine wesentlichen Unterschiede.

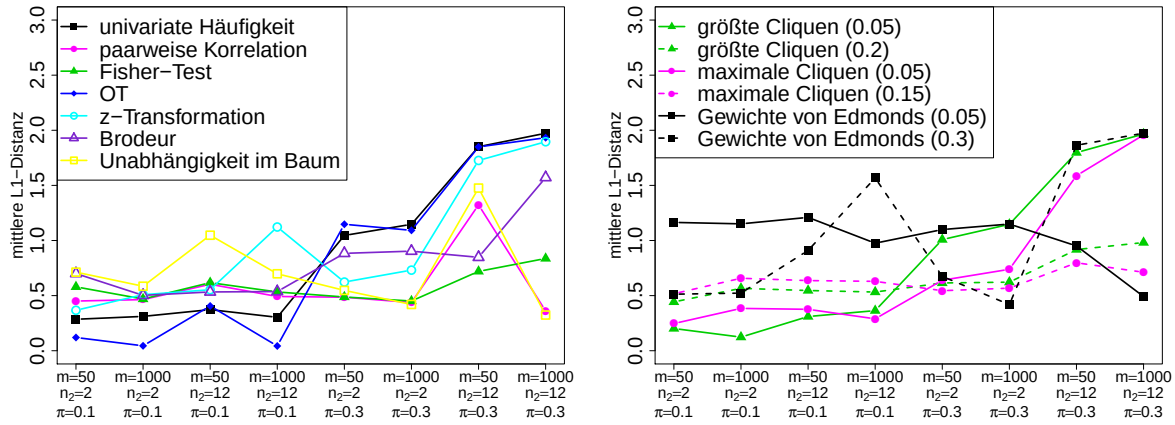



Abbildung 5.6: Vergleich der Variablenselektionsmethoden. Links: eindeutiger bester Schwellenwert, rechts: situationsabhängiger Schwellenwert.

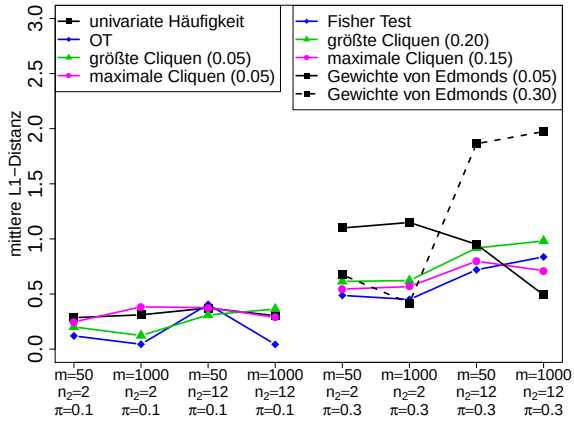



Abbildung 5.7: Vergleich aller Variablenselektionsmethoden. Basierend auf den vorangegangenen Ergebnissen muss zwischen Situationen mit geringer und höherer Wahrscheinlichkeit für das Rauschen unterschieden werden.

### 5.2.3 Auswertung basierend auf Sensitivität und Spezifität

Als Kriterium für ein gutes Variablenselektionsverfahren soll nun nicht mehr ein möglichst kleiner $L_1$-Abstand der Wahrscheinlichkeitsverteilungen des wahren bzw. angepassten Baummodells herangezogen werden, sondern ein direkter Vergleich der ausgewählten bzw. aussortierten Ereignisse. Es kommt darauf an, dass möglichst viele der $n_2$ Störereignisse und möglichst wenige der $n_1$ wahren Ereignisse aussortiert werden. Es müssen somit zwei Kriterien gleichzeitig betrachtet und optimiert werden. Ein optimales Abschneiden bei der Auswahl der wahren Ereignisse (Sensitivität, `sens`) reicht für ein gutes Variablenselektionsverfahren nicht aus, da dies immer erzielt werden kann, indem überhaupt kein Ereignis entfernt wird. Genauso können immer alle Störereignisse korrekterweise entfernt werden (Spezifität, `spec`), wenn auch kein wahres Ereignis ausgewählt wird.

Die Auswertung dieser weiteren Kriterien soll genauso erfolgen, wie bei der vorangegangenen $L_1$-Distanz. Zunächst soll überprüft werden, ob von den 32 Parametereinstellungen einige zu vernachlässigen sind, bevor der beste Schwellenwert und dann die beste Methode bestimmt wird.

**Vorarbeiten**

Wie in Abschnitt 5.2.2 werden auch hier zunächst Dendrogramme betrachtet, um zu beurteilen, ob einige der 32 Parameterkonstellationen redundant sind und somit nicht explizit betrachtet werden müssen.

Beim Vergleich der wahren und ausgewählten Ereignisse kann erfasst werden, wie viele wahre Ereignisse ausgewählt und wie viele Störereignisse aussortiert werden. Bei beiden Kriterien handelt es sich zunächst um absolute Zahlen. Da für jede Datensituation und jede Methode insgesamt 100 Durchläufe betrachtet werden, wird als erstes der Mittelwert dieser absoluten Zahlen berechnet. Dieser gibt an, wie viele Ereignisse im Durchschnitt korrekter- bzw. fälschlicherweise aussortiert werden. Um die Werte vergleichbar zu machen, da es mal 2 und mal 12 Störereignisse und mal 5 und mal 7 wahre Ereignisse gibt, werden die Mittelwerte auf das Intervall $[0, 1]$ normiert, so dass nun prozentuale Angaben vorliegen. Für beide Kriterien gilt, dass ein Wert in der Nähe von 1 für ein gutes Abschneiden der jeweiligen Methode steht. Ein Wert von z.B. 0.9 bei `sens` bedeutet damit, dass im Schnitt 90% aller wahren Ereignisse auch erkannt und ausgewählt werden. Liefert das Kriterium `spec` ein Ergebnis von z.B. 0.8, werden im Schnitt 80% aller Störereignisse entfernt. Somit gibt `sens` den Anteil richtig erkannter wahrer Ereignisse an und `spec` den Anteil korrekt entfernter Störereignisse.

Für die beiden Kriterien `sens` und `spec` sollen nun die durchschnittlichen Anteile korrekt erkannter Ereignisse für alle Methoden, Parameterkombinationen und Schwellenwerte geclustert werden. Mit Hilfe der daraus resultierenden Dendrogramme erhält man einen Überblick darüber, welche Parameterkombinationen ähnliche Ergebnisse aufweisen und kann so eventuell einige Parameter als unwichtig identifizieren und somit die zu betrachtende Anzahl der Parameterkombinationen reduzieren. In Abbildung 5.8 sind sowohl die Ergebnisse für `sens` und `spec` getrennt als auch gemeinsam aufgeführt. Wie oben schon beschrieben, reicht es nicht aus, nur in einem Kriterium gut abzuschneiden. Ein gutes Variablenselektionsverfahren sollte bei beiden Kriterien Werte in der Nähe von 1 erzielen. Daher wird zusätzlich für beide Kriterien das gemeinsame Dendrogramm betrachtet.

Aus dem Dendrogramm für das Kriterium `spec` (Abbildung 5.8 oben rechts) lässt sich ablesen, dass die Vereinigung über die Parameterkombinationen, die sich nur im Parameter $n_1$ unterscheiden (roter Kasten) meistens als erstes erfolgt. Für den Parameter $\alpha_l$ (blauer Kasten) verhält es sich genauso. Hier erfolgt die Vereinigung jeweils im zweiten Schritt. Wie beim Dendrogramm der $L_1$-Distanzen sind also die Parameterkombinationen, die sich nur

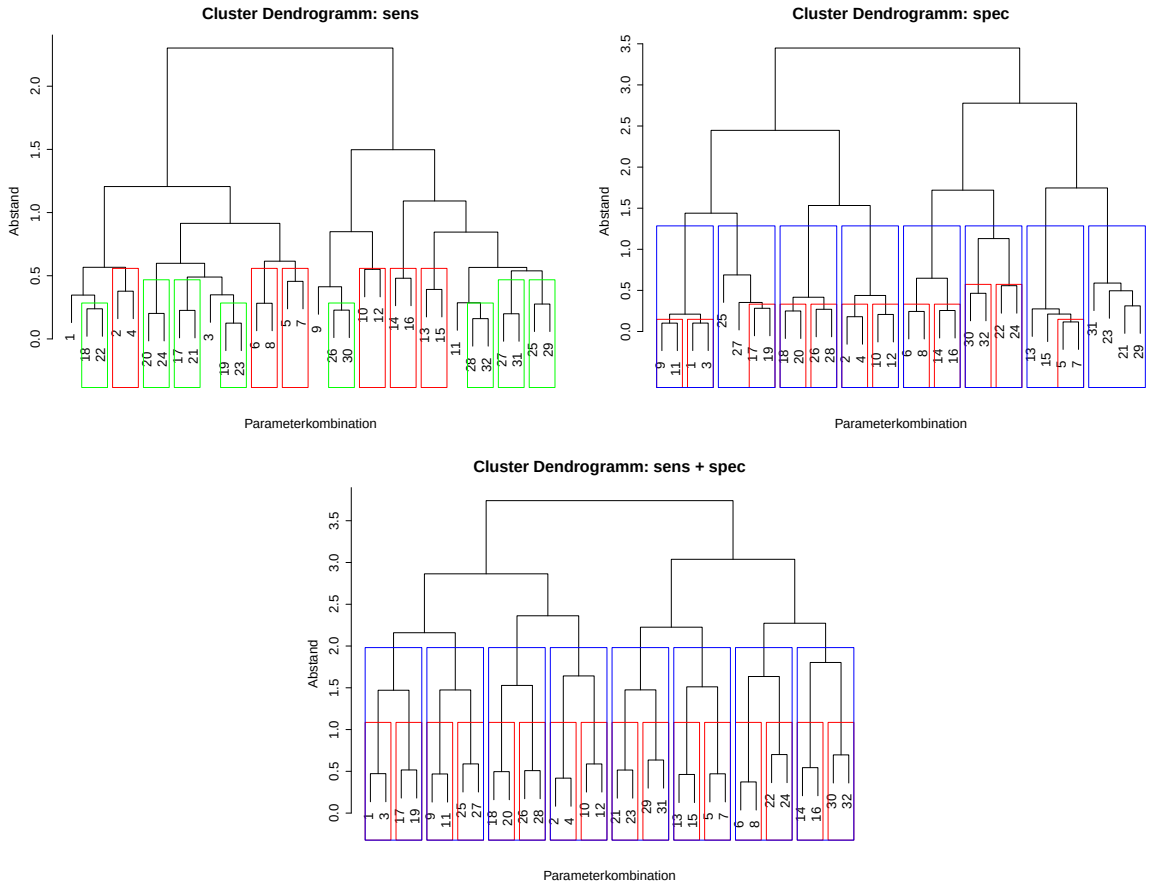



Abbildung 5.8: Cluster Dendrogramme zu den Kriterien `sens` und `spec` unter Verwendung des complete-Linkage Verfahrens um die Anzahl der Parameterkombinationen möglicherweise zu reduzieren. Die Parameterkombinationen, die durch einen roten Kasten zusammengefasst sind, unterscheiden sich nur in der Anzahl $n_1$ der wahren Ereignisse. Die blauen bzw. grünen Kästen stehen für Unterschiede nur bei den Parametern $\alpha_l$ bzw. $\pi_i$. Die 32 Parameterkombinationen sind die aus Tabelle 5.2.

in der Anzahl der wahren Ereignisse bzw. der unteren Grenze für die bedingte Wahrscheinlichkeit bei den Kantengewichten unterscheiden, diejenigen, die sehr ähnliche Ergebnisse erzielen.

Die Ergebnisse für die falsch aussortierten wahren Ereignisse (Abbildung 5.8 oben links) fallen jedoch nicht so eindeutig aus. Die Unterschiede beim Parameter $n_1$ (roter Kasten) sind weiterhin eher gering. Eine Vereinigung erfolgt früh im Dendrogramm. Es ist zu beachten, dass in dieser Grafik nur die Vereinigungen über $n_1$ im ersten Schritt markiert wurden, nicht jedoch die im zweiten. Auffällig ist allerdings, dass für dieses Kriterium in 8 Fällen (grüner Kasten) die erste Vereinigung über die Parameterkombinationen stattfindet, die sich nur durch den Parameter $\pi_i$ unterscheiden. Bei der Betrachtung der $L_1$-Distanzen (siehe Abschnitt 5.2.2) haben die zwei verschiedenen Intervalle von $\pi_i$ zu großen Unterschieden in den Ergebnissen geführt. Die hier beobachteten Resultate stehen aber nicht im Widerspruch zu den bisherigen, da sich das

Kriterium `sens` mit den richtig bzw. falsch aussortierten wahren Ereignissen beschäftigt, der Parameter $\pi_i$ jedoch die Häufigkeit der Störvariablen beeinflusst.

Fasst man beide Kriterien zusammen, ergibt sich das Dendrogramm in Abbildung 5.8 unten. Da eine Methode nur als gut bezeichnet werden kann, wenn sie für beide Kriterien gute Werte liefert, ist es durchaus sinnvoll, das gemeinsame Dendrogramm von `sens` und `spec` zu betrachten. Das Ergebnis ist eindeutig und beschreibt eine perfekte Hierarchie der fünf verschiedenen Parameter. Die Parameterkombinationen, die die ähnlichsten Ergebnisse liefern, unterscheiden sich nur in der Anzahl $n_1$ wahrer Ereignisse (roter Kasten). Als nächstes wird immer über den Parameter $\alpha_l$ zusammengefasst (blauer Kasten). Die nächsten Vereinigungen erfolgen über $n_2$ (dritte Stufe), $m$ (vierte Stufe) und als letztes über den Parameter $\pi_i$.

Zusammenfassend lässt sich sagen, dass es auch für die Kriterien `sens` und `spec` in Frage kommt, die Anzahl der Parameterkombinationen von 32 auf 8 zu reduzieren, indem $n_1 = 5$ und $\alpha_l = 0.2$ festgesetzt wird. Es ist jedoch zu beachten, dass bei alleiniger Betrachtung des Kriteriums `sens` die Ähnlichkeit der Ergebnisse für unterschiedliche Werte von $\alpha_l$ nicht gegeben ist.

**Ergebnisse: Was ist der beste Schwellenwert?**

Wie bei der $L_1$-Distanz werden nun die acht Parameterkombinationen aus Tabelle 5.3 betrachtet, um für jede Methode den besten Schwellenwert zu bestimmen. Die Ergebnisse für die Kriterien `sens` und `spec` sind in den Abbildungen 5.9 und 5.10 dargestellt. Wie man schnell erkennen kann, sind die beiden Kriterien jedoch nicht dazu geeignet, einen besten Schwellenwert für jede Methode zu bestimmen.

Für jede Methode und jede Parametersituation entspricht die Reihenfolge vom besten bis zum schlechtesten Ergebnis immer exakt der aufsteigenden bzw. absteigenden Reihenfolge der Schwellenwerte. Ist diese Reihenfolge für eine Methode beim Kriterium `sens` aufsteigend, so ist sie für das Kriterium `spec` absteigend und umgekehrt.

Für die Methode des exakten Tests von Fisher müsste z.B. der Schwellenwert 0.15 gewählt werden, wenn man den größten Anteil korrekt erkannter wahrer Ereignisse erzielen will. Sollen jedoch möglichst viele Störereignisse erkannt werden, ist der Schwellenwert 0.01 am besten geeignet. Da aber beide Kriterien gleichzeitig berücksichtigt werden sollen, gibt es keinen optimalen Schwellenwert.

Dieses Ergebnis ist nicht überraschend, da man immer das beste Ergebnis für die Kriterien `sens` bzw. `spec` erzielen kann, wenn insgesamt möglichst wenige bzw. viele Ereignisse entfernt werden. Dies entspricht daher immer dem größten oder kleinsten Schwellenwert. Ein optimaler Schwellenwert kann daher nicht mit Hilfe der Kriterien `sens` und `spec` gefunden werden.

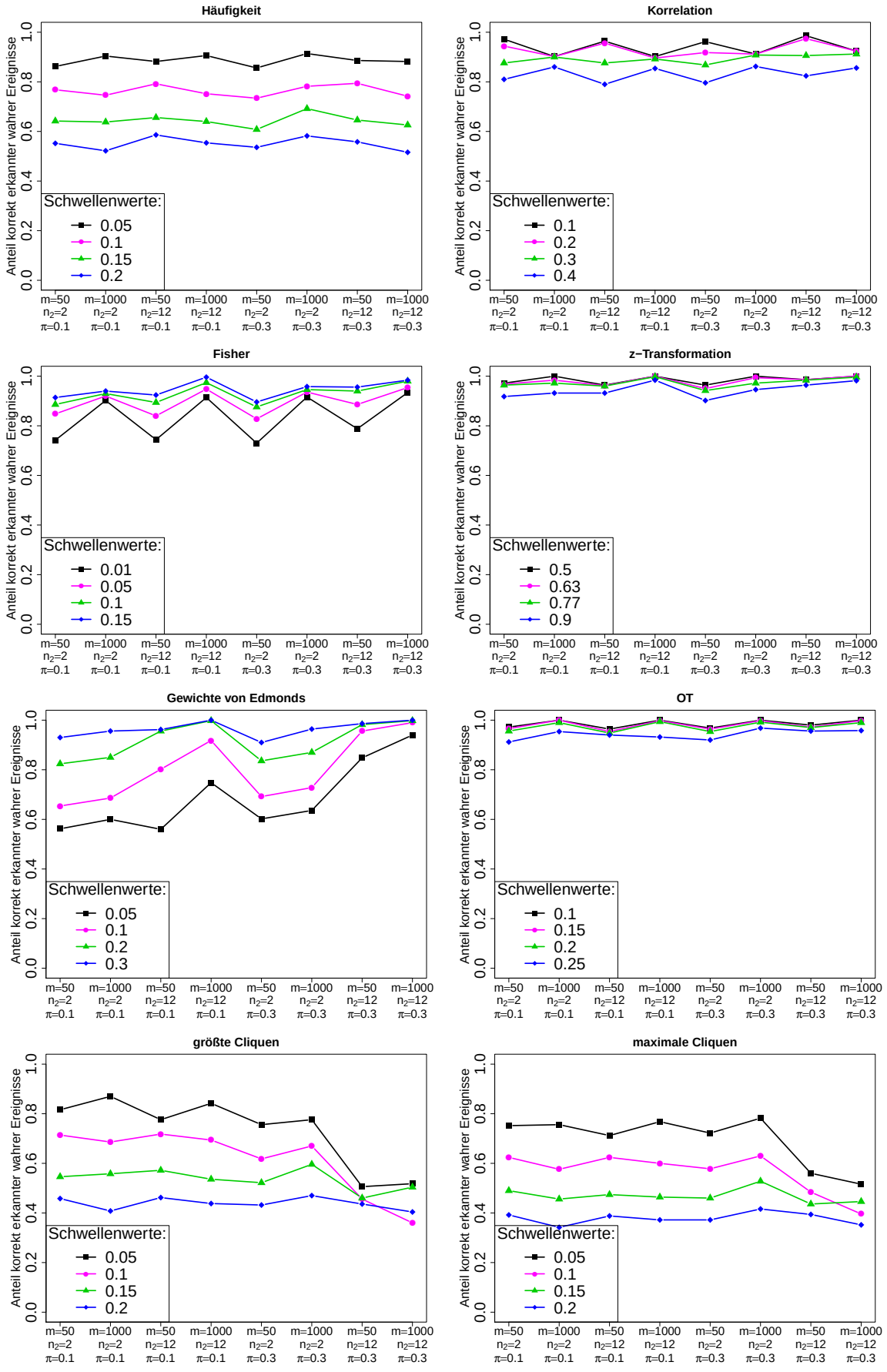



Abbildung 5.9: Ergebnisse der 8 Variablenselektionsverfahren mit je 4 verschiedenen Schwellenwerten für das Kriterium `sens`. Auf der x-Achse sind jeweils die 8 Parameterkombinationen, auf der y-Achse der Anteil korrekt erkannter wahrer Ereignisse dargestellt.

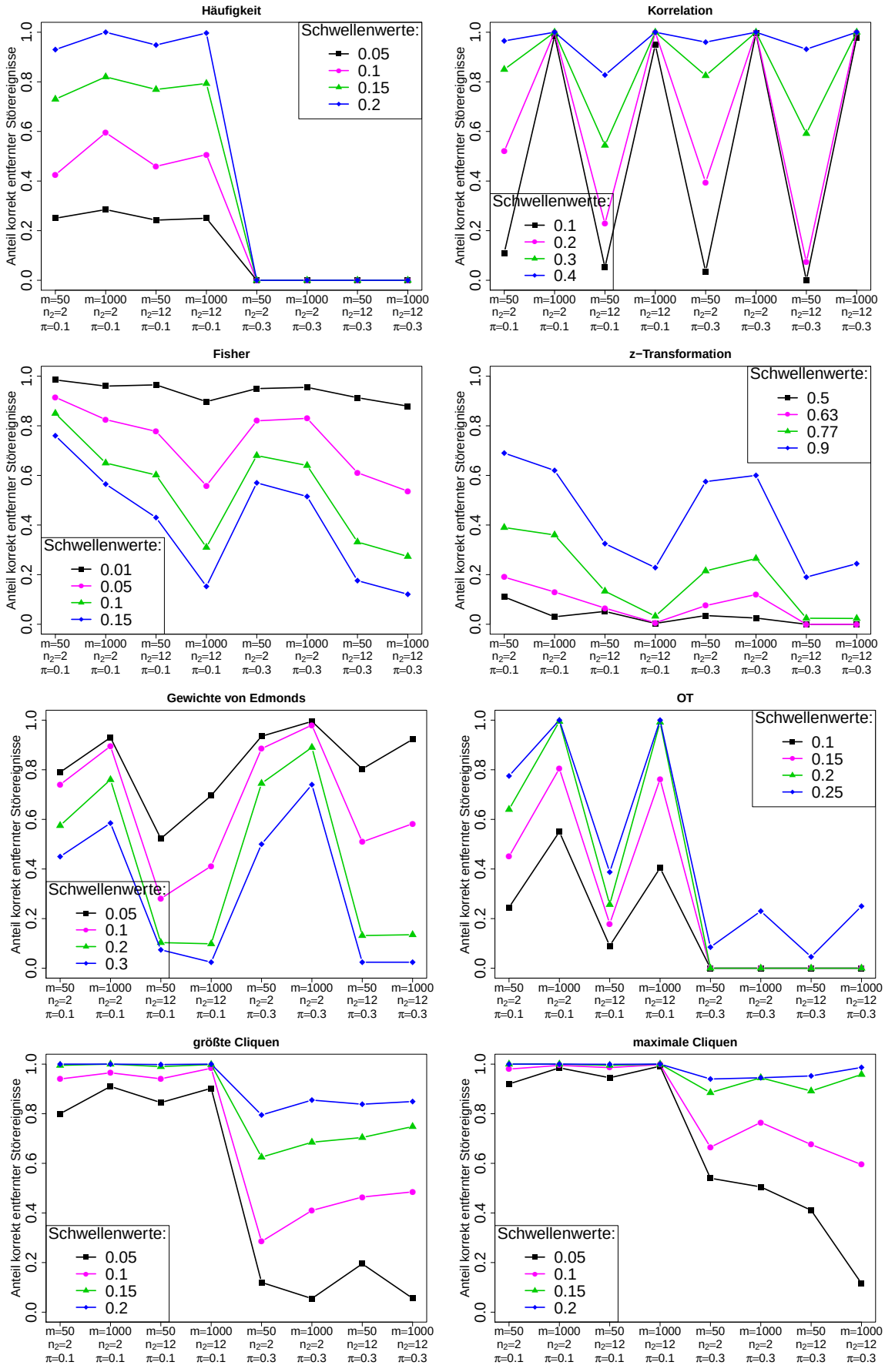



Abbildung 5.10: Ergebnisse der 8 Variablenselektionsverfahren mit je 4 verschiedenen Schwellenwerten für das Kriterium `spec`. Auf der x-Achse sind jeweils die 8 Parameterkombinationen, auf der y-Achse der Anteil korrekt erkannter Störereignisse dargestellt.

Da diese Maße aber dennoch wichtig sind, um die Güte eines Variablenselektionsverfahrens zu beurteilen, werden die optimalen Schwellenwerte aus Abschnitt 5.2.2 verwendet, die anhand der $L_1$-Distanz bestimmt wurden.

#### Ergebnisse: Was ist die beste Methode?

Eine gute Methode ist dadurch gekennzeichnet, dass viele wahre Ereignisse identifiziert und gleichzeitig möglichst viele Störereignisse aussortiert werden. Sowohl für die Sensitivität als auch die Spezifität sollten demnach hohe Werte vorliegen. In Abbildung 5.11 sind die Ergebnisse für die Kriterien `sens` und `spec` für die bekannten acht Parameterkombinationen dargestellt. Wie bei der $L_1$-Distanz werden aus Gründen der Übersichtlichkeit auch hier zunächst die sieben Methoden mit nur einem besten Schwellenwert (links) getrennt von den drei Methoden mit zwei situationsabhängigen Schwellenwerten (rechts) verglichen. Anschließend werden die besten daraus in einer Grafik direkt gegenübergestellt (unten).

Betrachtet man die Ergebnisse zur Sensitivität (Abbildung 5.11, oben), fällt auf, dass von den sieben Methoden mit global bestem Schwellenwert nur die univariate Häufigkeit und die Methode von Brodeur nicht geeignet sind, die wahren Ereignisse korrekt zu identifizieren. Alle anderen erkennen diese meistens in mehr als 80% der Fälle, 4 von 5 wahren Ereignissen werden also immer erkannt. Der Grund für das schlechte Abschneiden der zwei frequentistischen Methoden liegt darin, dass bei kleinen Kantenwahrscheinlichkeiten Ereignisse am Ende eines Pfades recht geringe Auftretenswahrscheinlichkeiten besitzen können und somit nicht ausgewählt werden.

Bei der Betrachtung der übrigen drei Methoden ist zu erkennen, dass sich dieser Sachverhalt auch auf das bisher gute Abschneiden der Cliquenmethoden auswirkt. Für den jeweils größeren Schwellenwert weisen sowohl die größten als auch die maximalen Cliquen keine annehmbaren Ergebnisse auf. Damit ein Ereignis in eine Clique aufgenommen werden kann, muss für alle Ereignispaare aus dieser Clique die Häufigkeit, mit der beide Ereignisse gemeinsam auftreten, über dem festgesetzten Schwellenwert liegen. Liegt aber die Wahrscheinlichkeit eines einzelnen Ereignisses bereits unterhalb des Schwellenwertes ist eine Auswahl dieses Ereignisses von vornherein ausgeschlossen.

Die Cliquenmethoden mit einem Schwellenwert von 0.05 weisen für die vier Situationen mit $\pi_i \in [0, 0.2]$ gute Ergebnisse auf. Die Methode der Gewichte von Edmonds kann bei einem Schwellenwert von 0.3 für alle Parametersituationen fast immer alle wahren Ereignisse identifizieren.

Wie schon gesagt, ist es jedoch nicht ausreichend, möglichst viele wahre Ereignisse zu erkennen. Gleichzeitig müssen auch möglichst viele Störereignisse erfasst werden. Die Ergebnisse dieser Spezifität sind in der mittleren Zeile von Abbildung 5.11 dargestellt. Lediglich der Fisher-Test (links) liefert für alle Parametersituationen zufriedenstellende Ergebnisse. Methoden wie OT oder die z-Transformation, die bei der Sensitivität am besten abgeschnitten haben, gehören hier zu den schlechtesten. Bei den Methoden mit zwei situationsabhängigen

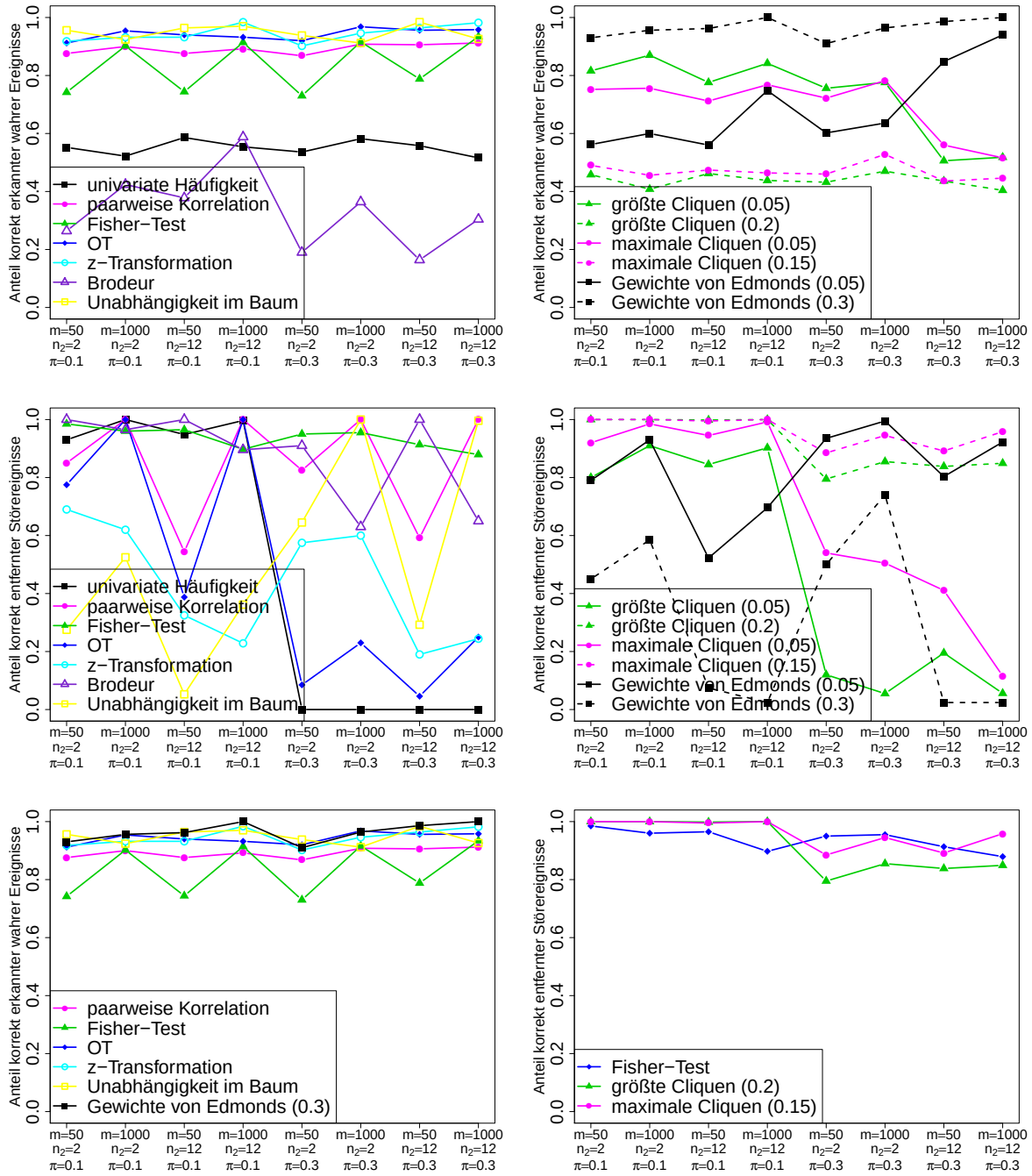



Abbildung 5.11: Vergleich der Variablenselektionsmethoden in Bezug auf die Gütekriterien `sens` (oben) und `spec` (mitte), links: eindeutiger bester Schwellenwert, rechts: situationsabhängiger Schwellenwert; sowie der Vergleich der besten Variablenselektionsmethoden (unten), links für das Gütekriterium `sens`, rechts für `spec`.

Schwellenwerten (rechts) weisen lediglich die Cliquen mit großem Schwellenwert ein gutes Ergebnis auf.

Vergleicht man nun die jeweils besten Methoden (Abbildung 5.11 unten), gibt es lediglich eine einzige Methode, die sowohl beim Erkennen wahrer Ereignisse als auch der Störereignisse gut abschneidet. Es handelt sich dabei um den Fisher-Test.

Im Vergleich zur Analyse der $L_1$-Distanz fällt zusätzlich auf, dass für Sensitivität und Spezifität nicht mehr nach Situationen mit unterschiedlicher Wahrscheinlichkeit $\pi_i$ für die Störereignisse getrennt werden muss. Ist eine Methode mit einem bestimmen Schwellenwert gut, gilt das für alle acht Parametersituationen.

Nun scheint es so, dass die Auswertung der Simulationsstudie bezüglich der $L_1$-Distanz auf der einen Seite und der Sensitivität und Spezifität auf der anderen Seite zu keinem gemeinsamen Ergebnis kommt. Im Hinblick auf den Abstand zwischen wahrem und angepasstem Baummodell schneiden die Cliquenmethoden am besten ab, bei der Unterscheidung zwischen wahren Ereignissen und Störereignissen ist es der Fisher-Test. Es müssen dazu jedoch mehrere Sachverhalte bedacht werden. Zum einen schneidet der Fisher-Test bei der Betrachtung der $L_1$-Distanz nicht schlecht ab. Für $\pi_i \in [0.2, 0.4]$ sind die Ergebnisse sogar sehr gut, für $\pi_i \in [0, 0.2]$ liegen sie im Mittelfeld. Zum anderen ist die Ursache des schlechten Abschneidens der Cliquen beim Erkennen der wahren Ereignisse zu berücksichtigen. Kleine Kantenwahrscheinlichkeiten im wahren Baum führen zu kleinen Auftretenswahrscheinlichkeiten für die Ereignisse am Ende eines Pfades und somit zum Ausschluss dieses Ereignisses. Natürlich kann bei echten Daten nicht davon ausgegangen werden, dass die wichtigen Ereignisse häufig genug auftreten. Gerade seltene Ereignisse sollen auch zuverlässig erkannt werden, wenn sie im Krankheitsprozess in einem späten Stadium eine wichtige Rolle spielen. Für das Auswahlkriterium der Cliquenmethoden ist jedoch die Häufigkeit nicht vollständig zu vernachlässigen. Zusätzlich ist in Abschnitt 5.2.3 darauf hingewiesen worden, dass das Zusammenfassen von Parametersituationen mit unterschiedlichem Wert für $\alpha_l$ für das Kriterium `sens` eventuell fragwürdig ist, da die Ergebnisse hier nicht eindeutig waren. Genau dieser Parameter beeinflusst jedoch die Kantenwahrscheinlichkeiten und damit direkt die Wahrscheinlichkeit für das Auftreten einzelner Ereignisse.

In Abbildung 5.12 sind daher die Ergebnisse für Sensitivität und Spezifität dargestellt, wenn für die acht ausgewählten Parametersituationen nicht $\alpha_l = 0.2$, sondern $\alpha_l = 0.5$ gilt.

Das Ergebnis für die Sensitivität ändert sich in Bezug auf die Cliquenmethoden deutlich. Für beide Schwellenwerte liegt der Anteil korrekt erkannter wahrer Ereignisse für alle Parametersituationen fast durchgehend über 0.8. Die Ergebnisse für den kleineren Schwellenwert 0.05 sind dabei etwas besser. Die Resultate der anderen Methoden ändern sich nicht großartig. Lediglich die univariate Häufigkeit schneidet deutlich besser ab.

Die Ergebnisse für die Spezifität unterscheiden sich generell nicht viel. Die Methode von Brodeur sowie die Cliquen mit großem Schwellenwert erzielen für $\pi_i \in [0.2, 0.4]$ bessere Ergebnisse, während die Cliquen mit kleinem Schwellenwert in diesen Situationen teilweise schlechter abschneiden.

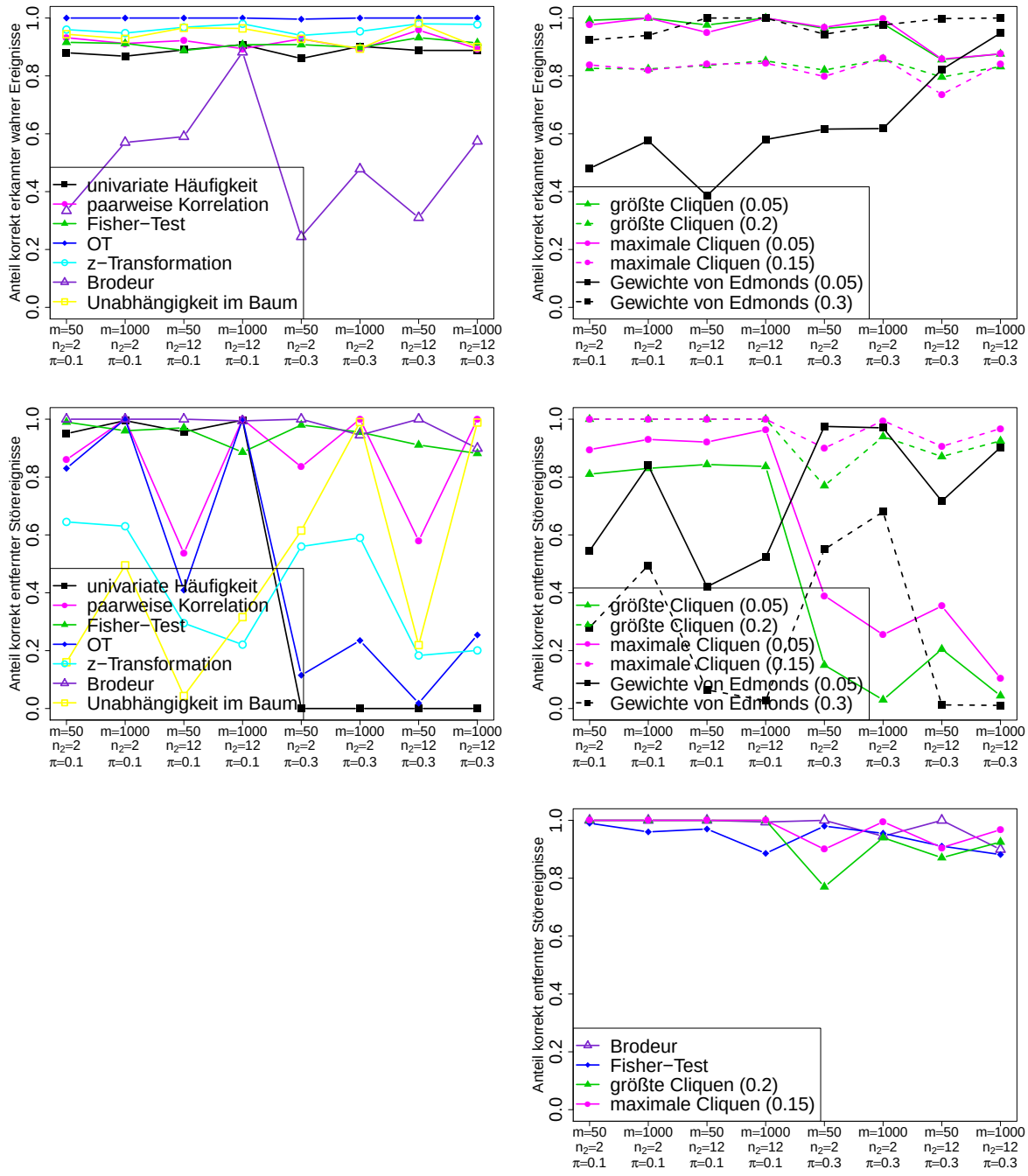



Abbildung 5.12: Die gleichen Grafiken wie in Abbildung 5.11 mit dem Unterschied, dass $\alpha_l = 0.5$ gilt und nicht mehr $\alpha_l = 0.2$: Vergleich der Variablenselektionsmethoden in Bezug auf die Gütekriterien `sens` (oben) und `spec` (mitte), links: eindeutiger bester Schwellenwert, rechts: situationsabhängiger Schwellenwert; sowie der Vergleich der besten Variablenselektionsmethoden für das Gütekriterium `spec` (unten). Auf die Grafik für das Kriterium `sens` wird der Übersichtlichkeit halber verzichtet, da alle bis auf 2 Verfahren gute Ergebnisse liefern (siehe obere Zeile).

Insgesamt sind also fast alle Methoden (bis auf Brodeur und die Gewichte von Edmonds mit Schwellenwert 0.05) dazu geeignet, die wahren Ereignisse zu identifizieren. Zum korrekten Erkennen der Störereignisse sollten allerdings nur die Methode von Brodeur, der Fisher-Test oder die Cliquenmethoden (mit größerem Schwellenwert) verwendet werden. Das bedeutet, dass sowohl die Cliquen als auch der Fisher-Test adäquate Methoden sind, um wahre Ereignisse und Störereignisse zuverlässig zu unterscheiden.

Sind die Kantenwahrscheinlichkeiten im wahren Baum und damit die Häufigkeit der wahren Ereignisse nicht zu gering, stellen die Cliquen (mit Schwellenwert 0.2 bzw. 0.15) einen geeigneten Ansatz zur Variablenselektion dar. Bestehen Zweifel daran, dass alle für die Krankheit wichtigen Ereignisse mit entsprechend großer Häufigkeit auftreten, sind die Cliquen mit kleinem Schwellenwert immer noch ein geeignetes Mittel, falls die Störvariablen nicht mit zu großer Wahrscheinlichkeit auftreten (siehe Abbildung 5.11 oben und mitte rechts für $\pi_i \in [0, 0.2]$).

Eine weitere Möglichkeit, die Ergebnisse der Variablenselektionsmethoden bezüglich Sensitivität und Spezifität darzustellen, sind sogenannte ROC-Kurven. Der Vorteil dieser Darstellung ist, dass direkt ablesbar ist, ob eine Methode als gut oder schlecht zu bewerten ist. Eine Methode ist genau dann zur Variablenselektion geeignet, wenn sie oben links in der Grafik abgebildet ist. Dann werden ausreichend wahre Ereignisse korrekt erkannt und ausreichend viele Störereignisse aussortiert. Der Nachteil ist, dass zur Übersichtlichkeit eine Grafik pro Parametersituation erstellt werden muss. Für eine einzelne Parametersituation lässt sich also auf einen Blick erkennen, welche Methoden gut abschneiden. Es müssen jedoch mehrere Grafiken betrachtet werden, wenn eine Aussage über mehrere Parametersituationen getroffen werden soll. Eine Methode kann nur dann als global gute Methode bezeichnet werden, wenn sie in allen Grafiken in der oberen linken Ecke zu finden ist. Die hier genannten ROC-Kurven für 16 verschiedene Parametersituationen sind in Abbildung 5.13 und 5.14 aufgeführt. Jeweils auf der linken Seite sind die Situationen mit Parameter $\alpha$ = 0.2 dargestellt, für die Grafiken auf der rechten Seite gilt $\alpha$ = 0.5. Die Reihenfolge der Parametersituationen in den Zeilen ist die gleiche, die bisher immer auf der x-Achse verwendet wurde.

### 5.2.4 Rauschen in den Daten

Bei der Durchführung der Simulationsstudie werden im zweiten Schritt Daten aus einem zufällig generierten Baum gezogen (siehe Abschnitt 5.2.1). Bei einem zugrunde liegenden echten Datensatz würde dieser Schritt der Datenerhebung entsprechen. In der Praxis können dabei durchaus Messfehler und kleine Ungenauigkeiten auftreten. Diese sind in der oben beschriebenen Simulationsstudie bislang nicht berücksichtigt. Die echten Daten bzw. die wahren Ereignisse sind alle fehlerfrei und passen exakt zum wahren Baummodell. In der Realität ist dieser Sachverhalt jedoch eher unwahrscheinlich. Um diesem Aspekt gerecht zu werden und die verschiedenen Variablenselektionsmethoden auch auf Daten mit Messfehlern zu bewerten, wird die Simulationsstudie erweitert.

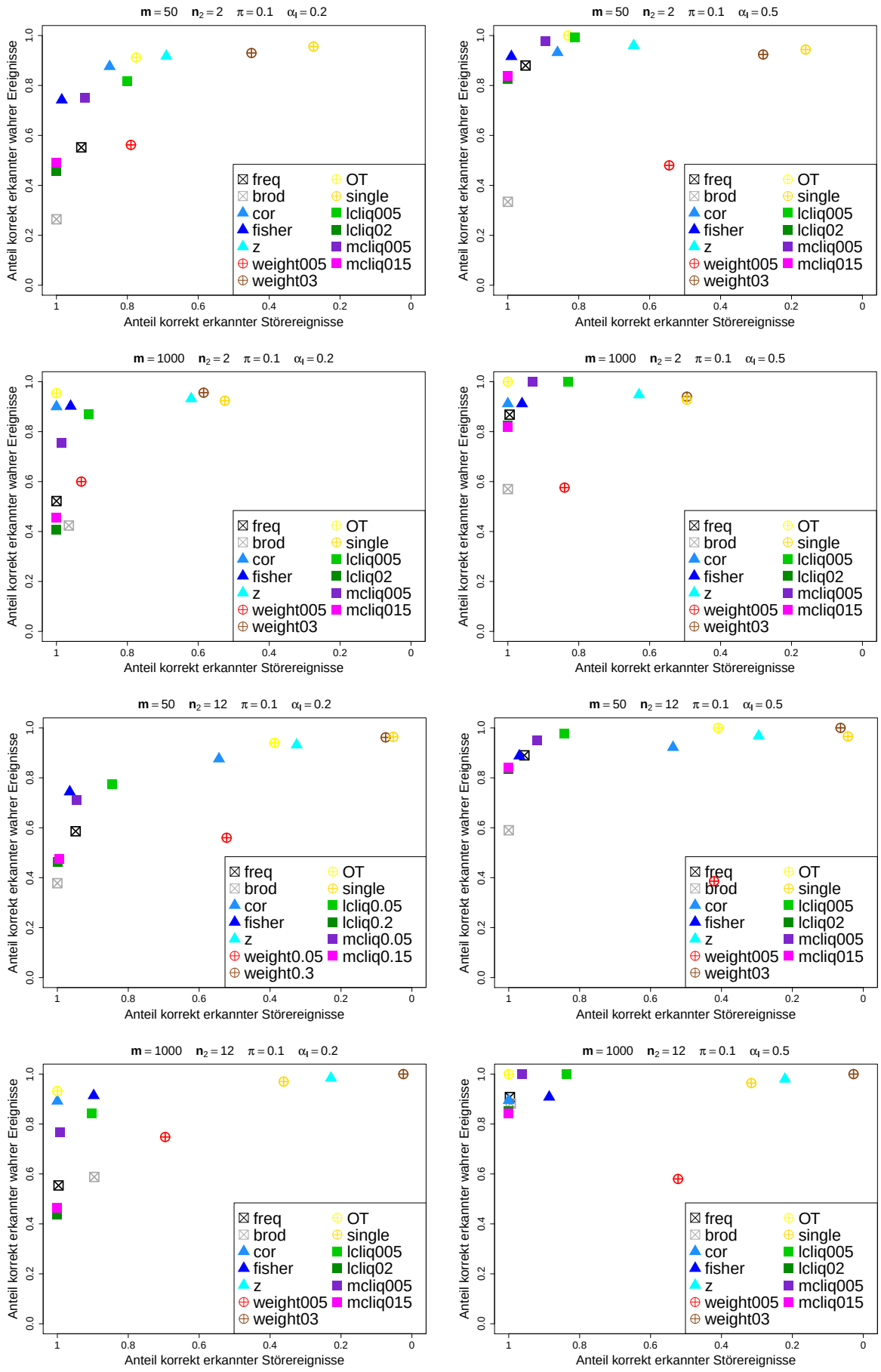



Abbildung 5.13: ROC-Kurven zur simultanen Analyse von Sensitivität und Spezifität für verschiedene Parametersituationen (Teil 1)

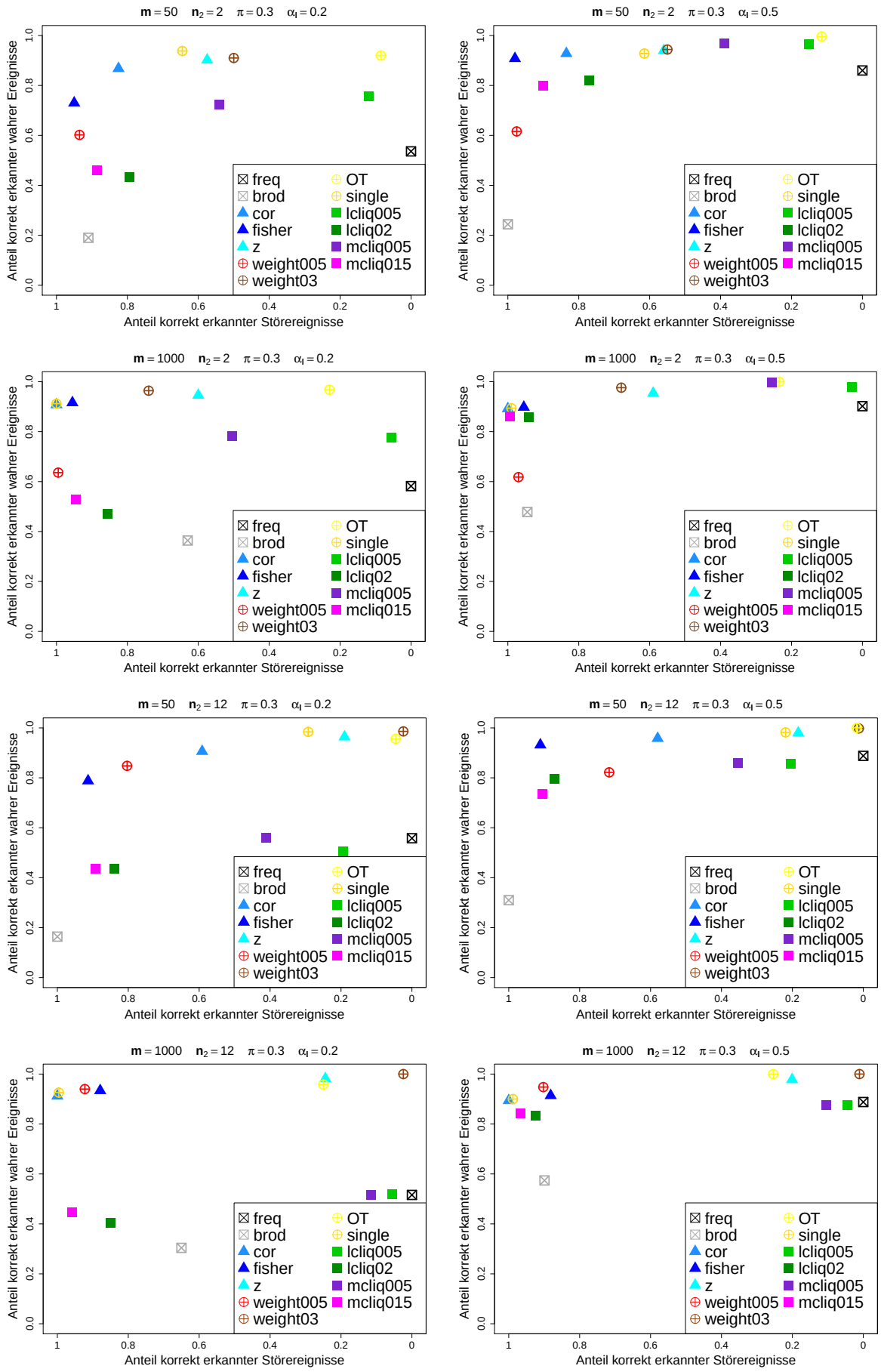



Abbildung 5.14: ROC-Kurven zur simultanen Analyse von Sensitivität und Spezifität für verschiedene Parametersituationen (Teil 2)

In Schritt 2 werden wie bisher Daten aus einem Baummodell gezogen. Bevor diese allerdings weiter verwendet werden, wird Rauschen hinzugefügt. Jeder Eintrag der Datenmatrix wird dabei mit einer festgelegten Wahrscheinlichkeit verändert. Da es sich bei der Datenmatrix um eine Binärmatrix handelt, entspricht diese Veränderung entweder dem Ersetzen einer 0 durch eine 1 oder umgekehrt. Ein Ereignis gilt damit entweder als eingetreten, obwohl es in Wahrheit nicht beobachtet werden kann oder es gilt als nicht eingetreten, obwohl die Veränderung in Wahrheit aufgetreten ist.

Dieses Rauschen in den Daten wird im Folgenden in zwei verschiedenen Varianten untersucht. Jeder Eintrag der Datenmatrix wird einmal mit Wahrscheinlichkeit 0.01 und einmal mit Wahrscheinlichkeit 0.10 verändert, so dass im Mittel 1% bzw. 10% Rauschen vorliegen. Die Ergebnisse der Variablenselektionsverfahren werden sowohl für die $L_1$-Distanz als auch für Sensitivität und Spezifität analysiert.

Der Vergleich der Ergebnisse zwischen echten und verrauschten Daten erfolgt über einen Scatterplot. Für alle zehn Variablenselektionsmethoden (7 mit global bestem Schwellenwert, 3 mit situationsabhängigen Schwellenwerten) und alle drei Abstandsmaße wird in Abbildung 5.15 gezeigt, wie sich die Ergebnisse ändern, wenn die Datenerhebung Messfehlern unterliegt. Hat das Rauschen in den Daten keinen Einfluss, liegen alle Punkte im Scatterplot auf der Diagonalen. Dies ist der Fall für alle Grafiken auf der linken Seite und damit für das 1%-ige Rauschen. Kleine Veränderungen in den Daten wirken sich demnach nicht auf die Ergebnisse aus. Ist das Rauschen etwas größer, d.h. treten Messfehler in 10% aller Fälle auf, unterscheiden sich die Ergebnisse etwas, jedoch nicht gravierend. Die Interpretation der Ergebnisse und damit die Reihenfolge bzw. das Abschneiden der Variablenselektionsmethoden bleibt gleich.

In den meisten Fällen sind die Ergebnisse der Variablenselektionsmethoden schlechter, wenn Rauschen in den Daten vorhanden ist. In einigen Fällen sind die Ergebnisse jedoch besser. Verfälschte Daten wirken sich demnach positiv aus. Warum dies in einigen Fällen vorkommt und welche Situationen davon betroffen sind, wird in Abbildung 5.16 herausgestellt. Diese Abbildung beinhaltet dieselben Grafiken wie Abbildung 5.15 (rechte Seite), jedoch sind die ungewöhnlichen Fälle, in denen verfälschte Daten zu einem besseren Ergebnis führen, farbig markiert. Es ist zu beachten, dass es sich bei der $L_1$-Distanz um Punkte unterhalb der Diagonalen handelt und bei Sensitivität und Spezifität um Punkte oberhalb der Diagonalen. Das liegt daran, dass bei der $L_1$-Distanz möglichst kleine Werte (Abstände zum wahren Modell) und bei Sensitivität und Spezifität möglichst große Werte (Anteile korrekt erkannter wahrer bzw. Störereignisse) erzielt werden sollen.

Die größten Unterschiede bzw. Abstände zur Diagonalen werden für folgende vier Methoden beobachtet: univariate Häufigkeit, Methode von Brodeur, größte und maximale Cliquen. Die Begründung, warum sich Rauschen in den Daten für diese Methoden positiv auswirken kann, hängt mit der Wahrscheinlichkeit für das Auftreten wahrer Ereignisse zusammen. Einige der wahren Ereignisse können eine recht geringe Wahrscheinlichkeit aufweisen, wenn sie erst spät im wahren Baummodell eintreten. Beim Verändern der Datenmatrix mit Wahrscheinlichkeit 0.10 ist es für diese Ereignisse wahrscheinlicher, dass Einträge von 0 zu 1 geändert werden, als an-

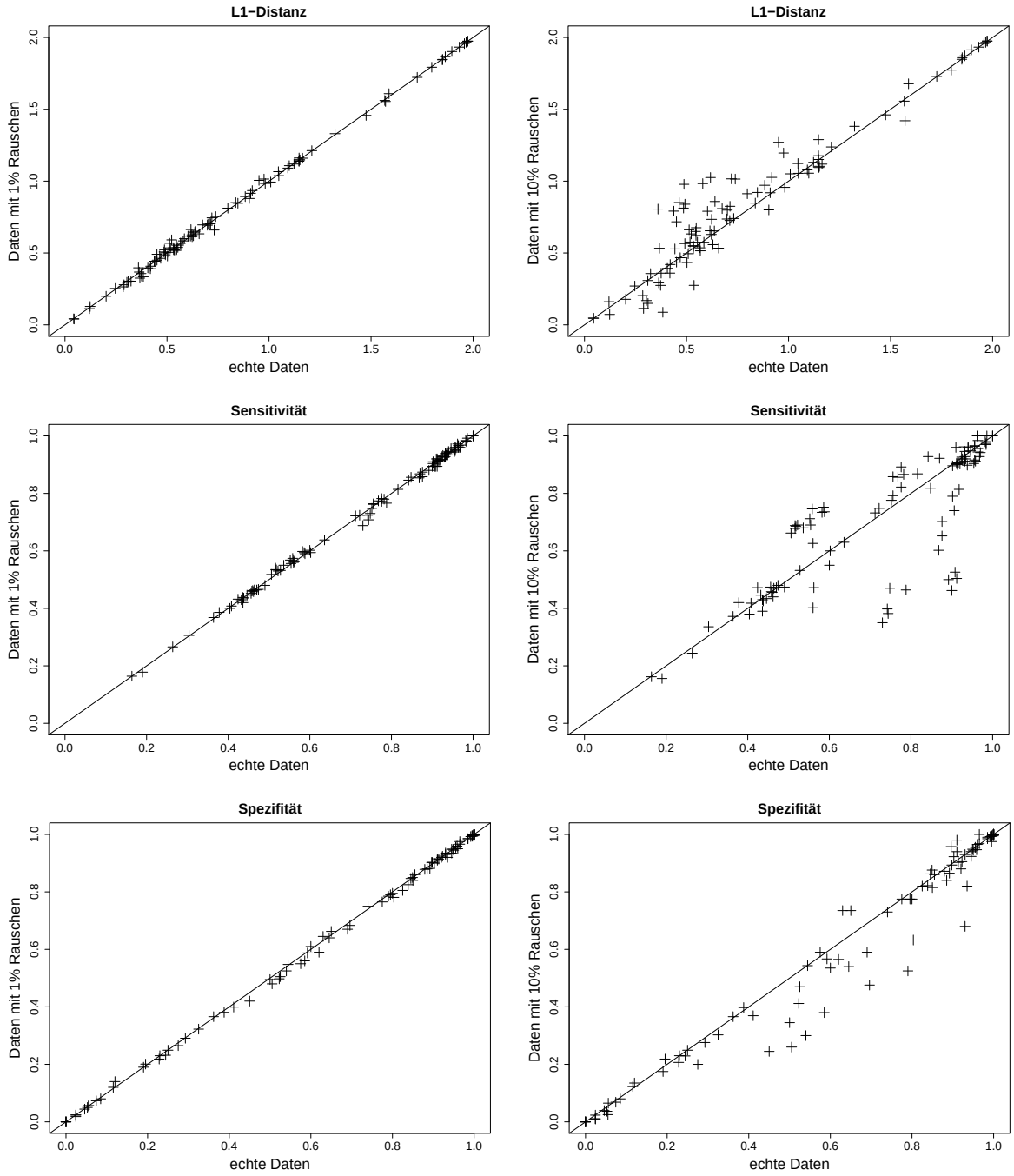



Abbildung 5.15: Scatterplots zur Gegenüberstellung der Ergebnisse bei echten und verrauschten Daten für alle drei Abstandsmaße sowie 1% und 10% Rauschen.

dersherum, da insgesamt mehr Nullen als Einsen vorkommen. Damit erhalten diese Ereignisse aber generell eine größere Auftretenswahrscheinlichkeit.

Für die Methode der univariaten Häufigkeit zeigt sich dadurch ein besseres Ergebnis beim Erkennen der wahren Ereignisse (Sensitivität), da sich deren Wahrscheinlichkeit durch das Rauschen womöglich so sehr erhöht hat, dass der notwendige Schwellenwert überschritten werden kann. Auch für die $L_1$-Distanz hat dies dann positive Auswirkungen, da sich wahrer und

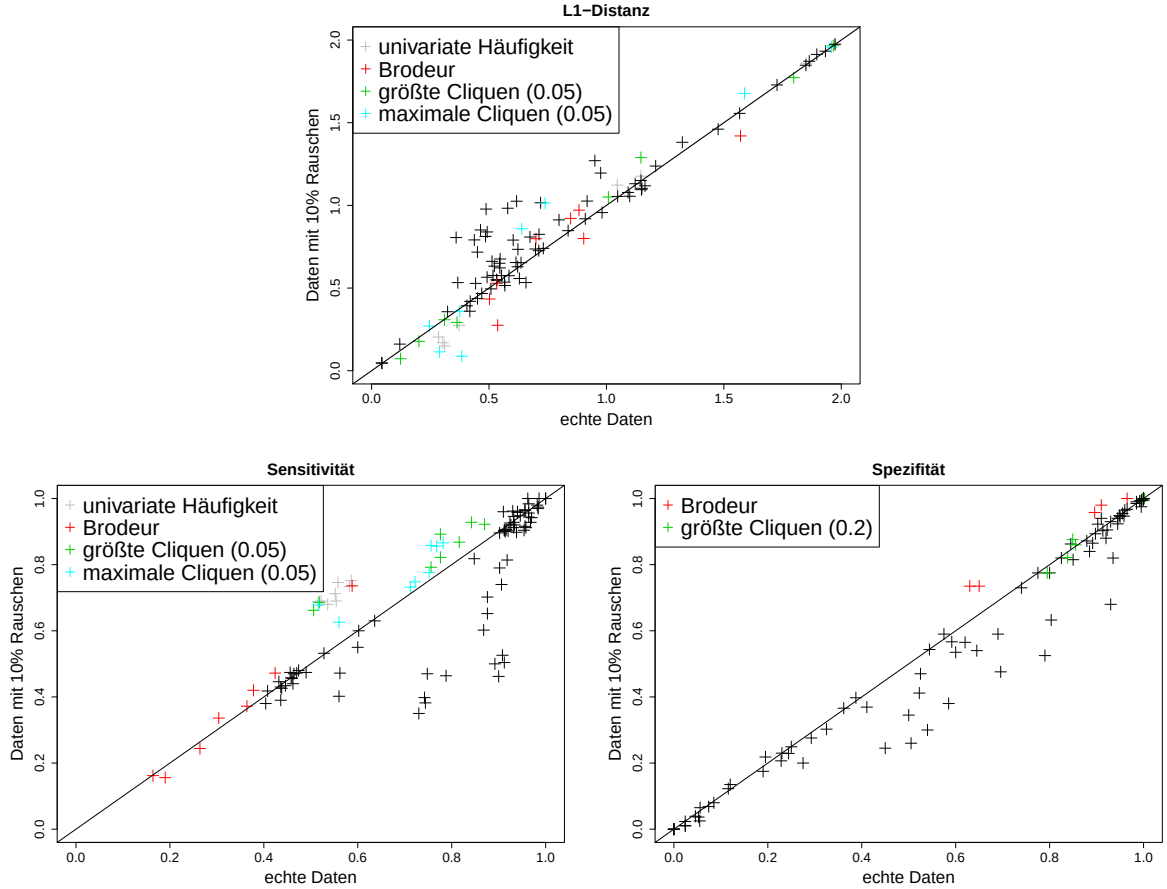



Abbildung 5.16: Scatterplots zur Gegenüberstellung der Ergebnisse bei echten und 10% verrauschten Daten. Methoden, bei denen einige Ergebnisse bei Vorliegen von Rauschen besser sind, sind farbig markiert.

angepasster Baum ähnlicher sind, wenn mehr identische Ereignisse enthalten sind. Bei der Methode von Brodeur führt die höhere Auftretenswahrscheinlichkeit für wahre Ereignisse zu einem höheren Schwellenwert, so dass mehr Störereignisse aussortiert werden können. Die Cliquenmethoden profitieren ebenfalls von diesem Sachverhalt, da nur Ereignisse ausgewählt werden können, die oft genug gleichzeitig mit anderen Ereignissen aufgetreten sind. Ist die Wahrscheinlichkeit für nur das Ereignis an sich schon kleiner als der Schwellenwert, sind sie von vornherein vom Selektionsprozess ausgeschlossen.

Zusammengefasst ergeben sich bei Vorliegen von Rauschen in den Daten zwar einige Abweichungen zu den ursprünglichen Ergebnissen, diese wirken sich aber nicht auf das Abschneiden der Variablenselektionsmethoden aus. Absolut betrachtet mögen die Ergebnisse bei 10% Rauschen zwar etwas schlechter sein, es sind aber immer noch die gleichen Methoden, die gut bzw. schlecht dabei abschneiden.

## 5.3 Anwendung auf Datensätze

Im Folgenden sollen die vorgestellten Variablenselektionsmethoden, die bisher anhand einer Simulationsstudie bewertet wurden, auch auf echte Daten angewendet werden. Dazu werden drei unterschiedliche Datensätze herangezogen und die Ergebnisse mit Resultaten aus der Literatur verglichen.

### 5.3.1 Meningiom

Bei einem Meningiom handelt es sich um einen meistens gutartigen Hirntumor, der operativ entfernt und damit geheilt werden kann. Nur sehr selten treten bösartige Formen auf, die ein aggressives Verhalten zeigen und wiederholt auftreten können (Urbschat et al., 2011).

Der vorliegende Meningiom Datensatz besteht aus 661 Beobachtungen von 9 Ereignissen. Dieser Datensatz wurde bereits von Urbschat et al. (2011) im Zusammenhang mit onkogenetischen Baum-Mischungs-Modellen verwendet und analysiert. So können die Ergebnisse der verwendeten Variablenselektionsverfahren optimal mit bereits gewonnenen Erkenntnissen verglichen werden.

Bei den genannten 9 Ereignissen handelt es sich um chromosomale Gewinne oder Verluste an Chromosomenarmen bzw. ganzen Chromosomen des Hirntumors. Zur Bestimmung des genetischen Zustandes einer Beobachtung, d.h. für die Benennung der vorliegenden genetischen Veränderungen, liegen für jeden Tumor mehrere Klone vor. Urbschat et al. (2011) betrachten zwei verschiedene Vorgehensweisen, um das genetische Muster anzugeben. Zum einen wird das häufigste Muster an genetischen Veränderungen als Repräsentant für eine Beobachtung gewählt, zum anderen das, was am weitesten fortgeschritten ist. In dieser Arbeit wird jedoch nur die erste Variante betrachtet.

Dass der vorliegende Datensatz aus nur 9 Ereignissen besteht, liegt daran, dass für die von Urbschat et al. (2011) beschriebenen Daten bereits eine Häufigkeitsselektion mit dem Parameter $\tau = 0.018$ durchgeführt wurde. Alle anderen möglichen Ereignisse treten somit in weniger als 1.8% aller Fälle auf.

Im Folgenden werden die 10 verschiedenen Variablenselektionsverfahren mit ihren jeweils besten Schwellenwerten auf den Datensatz angewendet. Das Ergebnis, und damit die ausgewählten Ereignisse für jede Methode, ist in Tabelle 5.4 dargestellt.

Die Variablenselektionsmethoden basierend auf dem Fisher-Test, der z-Transformation und der Unabhängigkeit im Baum wählen alle 9 Ereignisse aus, während die beiden Cliquenmethoden mit größerem Schwellenwert gar kein Ereignis auswählen. Sehr viele Ereignisse werden anhand der paarweisen Korrelation, den Gewichten von Edmonds (hoher Schwellenwert) und der bedingten Wahrscheinlichkeiten im Baum ausgewählt. Folgende fünf Methoden selektieren nur ein bis drei Ereignisse: univariate Häufigkeit, Brodeur, Gewichte von Edmonds (niedriger Schwellenwert) und die beiden Cliquenansätze (ebenfalls niedriger Schwellenwert).

Tabelle 5.4: Auflistung der Ereignisse des Meningiom Datensatzes, die von den einzelnen Variablenselektionsverfahren mit ihren jeweils besten Schwellenwerten ausgewählt wurden ($x$ = Ereignis wurde ausgewählt). Die größten und maximalen Cliquen mit Schwellenwert 0.2 bzw. 0.15 wählen keine Ereignisse aus.

| Methode | freq | brod | cor | fisher | z | weight | weight | OT | single | lcliq | mcliq |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\tau$ | 0.2 | 0.1 | 0.3 | 0.01 | 0.9 | 0.05 | 0.3 | 0.25 | - | 0.05 | 0.05 |
| Chr14- | | | x | x | x | x | x | x | x | x | x |
| Chr22- | x | x | x | x | x | | | x | x | x | x |
| Chr1p- | | | x | x | x | x | x | x | x | x | |
| Chr6- | | | x | x | x | | x | x | x | | |
| Chr10- | | | x | x | x | x | x | | x | | |
| Chr18- | | | x | x | x | | x | x | x | | |
| Chr19- | | | x | x | x | | x | | x | | |
| ChrY- | | | | x | x | | | x | x | | |
| ChrX- | | | | x | x | | | | x | | |

In diesem Fall kann von einem niedrigen Anteil an Rauschvariablen ausgegangen werden, da nur 9 Ereignisse mit einer Wahrscheinlichkeit von mehr als 1.8% auftreten. Basierend auf den Ergebnissen der Simulationsstudie, die die Verwendung der Cliquenmethoden (hier mit niedrigem Schwellenwert) nahelegen, werden die Ereignisse $14-, 22-$ und $1p-$ ausgewählt.

Bedingt durch die bereits durchgeführte Häufigkeitsselektion beinhaltet dieser Datensatz nur sehr wenige Ereignisse. Daher werden in einem zweiten Schritt 39 zusätzliche Rauschvariablen hinzugefügt, die die möglichen Gewinne und Verluste an den anderen Chromosomen darstellen sollen.[2] Da diese zusätzlichen Ereignisse alle mit einer Wahrscheinlichkeit von weniger als 1.8% auftreten, werden sie als zufällige Realisierung einer Binomialverteilung mit $\pi = 0.005$ gezogen. Die Ergebnisse der Variablenselektionsmethoden für diesen erweiterten Meningiom Datensatz sind in Tabelle 5.5 dargestellt.

Bemerkenswerterweise wählen nur die Häufigkeitsmethoden `freq` und `brod` sowie die Cliquenmethoden `lcliq` und `mcliq` keine der zusätzlichen Störereignisse aus. Alle anderen Variablenselektionsmethoden wählen einige oder sogar viele falsch positive Ereignisse aus. Zusätzlich wählen die Methoden nach Brodeur und den Gewichten von Edmonds noch weitere der 9 'wahren' Ereignisse aus. Bei allen anderen Methoden stimmt die Auswahl der 'wahren' Ereignisse mit der bei insgesamt nur 9 Ereignissen überein (vergleiche mit Tabelle 5.4).

Wenn man annimmt, dass die 9 ursprünglichen Variablen auch die 'wahren' Ereignisse darstellen, die ausgewählt werden sollen, kann für jede Methode ein optimaler Schwellenwert bestimmt werden, der am besten zwischen den beiden Gruppen unterscheidet. Dieser Schwellenwert wird

[2] Da nur bei einem der 9 beschriebenen Ereignisse ein konkreter Chromosomenarm betrachtet wurde, werden hier nur ganze Chromosomen betrachtet. Insgesamt soll daher eine Zahl von 48 Ereignissen erreicht werden (Chromosomen 1 bis 22, sowie X und Y jeweils mit Gewinn und Verlust).

Tabelle 5.5: Auflistung der Ereignisse des erweiterten Meningiom Datensatzes (39 zusätzliche Ereignisse mit einer zufälligen Häufigkeit von 0.5%), die von den einzelnen Variablenselektionsverfahren mit ihren jeweils besten Schwellenwerten ausgewählt wurden ($x$ = Ereignis wurde ausgewählt). Die größten und maximalen Cliquen mit Schwellenwert 0.2 bzw. 0.15 wählen überhaupt keine Ereignisse aus. Die Zeile '# random' gibt an, wie viele der 39 zusätzlichen Ereignisse ausgewählt wurden. Die letzten beiden Zeilen benennen den Schwellenwert, der gewählt werden muss, wenn alle 9 'wahren' Ereignisse und so wenig Störereignisse wie möglich ausgewählt werden sollen, sowie erneut die Anzahl an ausgewählten Störereignissen ('-1' bedeutet dabei, dass ein 'wahres' Ereignis nicht ausgewählt werden konnte).

| Methode | freq | brod | cor | fisher | z | weight | weight | OT | single | lcliq | mcliq |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\tau$ | 0.2 | 0.04 | 0.3 | 0.01 | 0.9 | 0.05 | 0.3 | 0.25 | - | 0.05 | 0.05 |
| Chr14- | | x | x | x | x | x | x | x | x | x | x |
| Chr22- | x | x | x | x | x | x | x | x | x | x | x |
| Chr1p- | | x | x | x | x | x | x | x | x | x | |
| Chr6- | | | x | x | x | x | x | x | x | | |
| Chr10- | | | x | x | x | x | x | | x | | |
| Chr18- | | | x | x | x | x | x | x | x | | |
| Chr19- | | | x | x | x | x | x | | x | | |
| ChrY- | | x | | x | x | x | x | x | x | | |
| ChrX- | | x | | x | x | x | x | | x | | |
| # random | 0 | 0 | 8 | 6 | 20 | 23 | 28 | 8 | 36 | 0 | 0 |
| 'opt.' $\tau$ | 0.0166 | - | 0.1940 | 0.0015 | $1\text{-}10^{-7}$ | 0.0480 | - | 0.1801 | - | 0.0045 | 0.0030 |
| # random | 0 | - | 11 | 2 | 11 | 21 | - | 12 | - | 0 | -1 |

so gewählt, dass alle 9 'wahren' Ereignisse und so wenige Störereignisse wie möglich selektiert werden. Das Ergebnis ist ebenfalls in Tabelle 5.5 aufgeführt.

Auch hier schaffen es nur die Häufigkeits- und Cliquen-basierten Methoden eindeutig zwischen 'wahren' und zufälligen Ereignissen zu unterscheiden.[3] Alle anderen Variablenselektionsmethoden wählen auch mehr oder weniger viele der künstlich generierten Ereignisse.

Zusätzlich zum Vergleich der ausgewählten Ereignisse der verschiedenen Selektionsmethoden soll nun auch der geschätzte Krankheitsverlauf verglichen werden. Dazu werden zum einen die 9 Ereignisse und das geschätzte Modell aus Urbschat et al. (2011) herangezogen und zum anderen onkogenetische Bäume an die ausgewählten Ereignisse aller Variablenselektionsmethoden angepasst.

In Abbildung 5.17 sind drei verschiedene Baummodelle abgebildet. Zum einen als Art Referenzbaum das Baummodell mit allen Ereignissen, zum anderen die Ergebnisse der zwei vielversprechendsten Selektionsmethoden `freq` und `lcliq` (niedriger Schwellenwert).

[3] Dass bei den maximalen Cliquen ein 'wahres' Ereignis nicht ausgewählt werden kann liegt daran, dass die Ereignisse X- und Y- nie gleichzeitig auftreten. Sie können damit nie in einer gemeinsamen Clique liegen. Bei den größten Cliquen gibt es zwei gleich große, so dass alle Ereignisse aus beiden Cliquen ausgewählt werden. Bei den maximalen Cliquen kann aber insgesamt nur eine Clique ausgewählt werden.

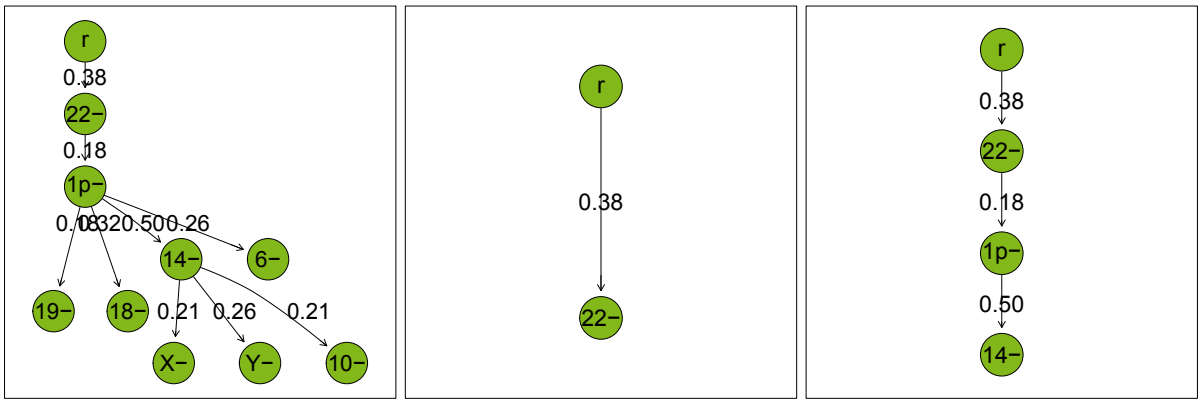



Abbildung 5.17: Onkogenetische Baummodelle für den Meningiom Datensatz. Links der Baum mit allen Ereignissen, in der Mitte und rechts die Bäume basierend auf der Häufigkeits- bzw. Cliquen-Selektion (größte Cliquen, niedriger Schwellenwert).

Von den 9 Ereignissen des Datensatzes werden mit Hilfe der beiden genannten Variablenselektionsverfahren nur ein bzw. drei Ereignisse ausgewählt. Anhand der Baummodelle ist jedoch zu erkennen, dass es sich dabei um die wichtigsten handelt. Das Ereignis 22− ist das erste auftretende Ereignis und auch der Beginn aller folgenden Pfade. Das Auftreten aller anderen Ereignisse hängt somit von diesem Ereignis ab. Die Methode der univariaten Häufigkeit wählt zumindest dieses wichtige Ereignis aus. Die Cliquenmethode `lcliq` mit kleinem Schwellenwert wählt drei Ereignisse aus und deckt damit den häufigsten Pfad ab. Die wichtigsten Ereignisse werden somit von den erfolgversprechendsten Variablenselektionsmethoden ausgewählt.

### 5.3.2 HIV

Auch beim HIV Datensatz handelt es sich um einen umfassend analysierten Datensatz, bei dem die Krankheitspfade bekannt sind (Beerenwinkel et al., 2005). Dieser Datensatz ist im R-Paket Rtreemix (Bogojeska et al., 2008) enthalten und besteht aus 364 Beobachtungen für 6 Ereignisse. Bei diesen Ereignissen handelt es sich jedoch nicht um Veränderungen am menschlichen Genom, sondern um Mutationen im Genom des HI-Virus'. Diese Mutationen wurden im Rahmen einer Behandlung der Patienten durch Monotherapie mit Zidovudin beobachtet.

Erneut werden alle zehn Variablenselektionsverfahren auf diesen Datensatz angewendet. Die Ergebnisse sind in Tabelle 5.6 zu finden.

Von den insgesamt 13 verschiedenen Selektionsverfahren (bei drei Verfahren werden jeweils zwei Schwellenwerte betrachtet) wählen sechs jeweils alle Ereignisse aus während vier Methoden nur jeweils zwei Ereignisse auswählen. Die Methode der univariaten Häufigkeit, sowie beide Cliquenmethoden mit kleinem Schwellenwert selektieren vier oder fünf Ereignisse. Da man bei nur sechs ursprünglichen Ereignissen von einem eher kleinen Anteil des Rauschens

Tabelle 5.6: Auflistung der Ereignisse des HIV Datensatzes, die von den einzelnen Variablenselektionsverfahren mit ihren jeweils besten Schwellenwerten ausgewählt wurden ($x$ = Ereignis wurde ausgewählt).

| Methode | freq | brod | cor | fisher | z | weight | weight | OT | single | lcliq | lcliq | mcliq | mcliq |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\tau$ | 0.2 | 0.1 | 0.3 | 0.01 | 0.9 | 0.05 | 0.3 | 0.25 | - | 0.05 | 0.20 | 0.05 | 0.15 |
| 215 F,Y | x | x | x | x | x | | x | x | x | x | x | x | x |
| 41 L | x | | x | x | x | x | x | x | x | x | x | | x |
| 70 R | x | x | x | x | x | | x | x | x | x | | x | |
| 67 N | x | | x | x | x | | x | x | x | x | | x | |
| 219 E,Q | | | x | x | x | | x | x | x | x | | x | |
| 210 W | | | x | x | x | x | x | x | x | | | | |

ausgehen kann, würde man bei den Cliquenmethoden diesen kleineren Schwellenwert bevorzugen.

In Abbildung 5.18 ist wieder das Baummodell mit allen Ereignissen als Referenzmodell, sowie zum Vergleich die Baummodelle nach Selektion mit univariater Häufigkeit und größtmöglicher Cliquen dargestellt.

Der Baum mit allen Ereignissen zeigt zwei unabhängige Pfade mit jeweils drei Ereignissen. Bei der Betrachtung der anderen beiden Modelle sind diese zwei unabhängigen Pfade ebenfalls zu erkennen. Bei den größten Cliquen fehlt nur das Ereignis $210W$ am Ende des rechten Pfades. Dieses Ereignis fehlt ebenfalls im Baum basierend auf der Selektion nach univariater Häufigkeit. Das zweite fehlende Ereignis im Vergleich zum Referenzbaum ist interessanterweise das mittlere Ereignis im linken Pfad.

Zusammenfassend lässt sich festhalten, dass die Ergebnisse der Variablenselektion für den HIV Datensatz recht ähnlich sind. Aufgrund der wenigen Ereignisse war zu erwarten, dass viele oder alle Ereignisse ausgewählt werden. Dieses haben viele Selektionsmethoden erfüllt inklusive der vielversprechenden Cliquenmethoden.

### 5.3.3 Glioblastom

Bei einem Glioblastom handelt es sich um die häufigste Form eines bösartigen Hirntumors. Da es sich um einen sehr aggressiven Tumor handelt, liegt die mediane Überlebenszeit bei nur 14 Monaten (Tolosi, 2011).

Der vorliegende Datensatz stammt aus der öffentlichen Datenbank 'The Cancer Genome Atlas' und wurde von Laura Tolosi vorverarbeitet (Tolosi, 2011; Tolosi et al., 2013). Er enthält 539 Beobachtungen von 132 Ereignissen. Bei diesen Ereignissen handelt es sich um Gewinn (+), Verlust (-) und Amplifikation (++, deutliche Vervielfältigung) der Chromosomenarme 1 bis 22.

Im Unterschied zu den anderen beiden Datensätzen ist die Anzahl der Ereignisse hier weitaus größer, da noch keine vorgeschaltete Variablenselektion stattgefunden hat. Weiterhin ist bisher kein Baummodell an diese Daten angepasst und analysiert worden, so dass kein explizites Referenzmodell vorliegt. In der Literatur sind jedoch wichtige Ereignisse bekannt, die mit dem

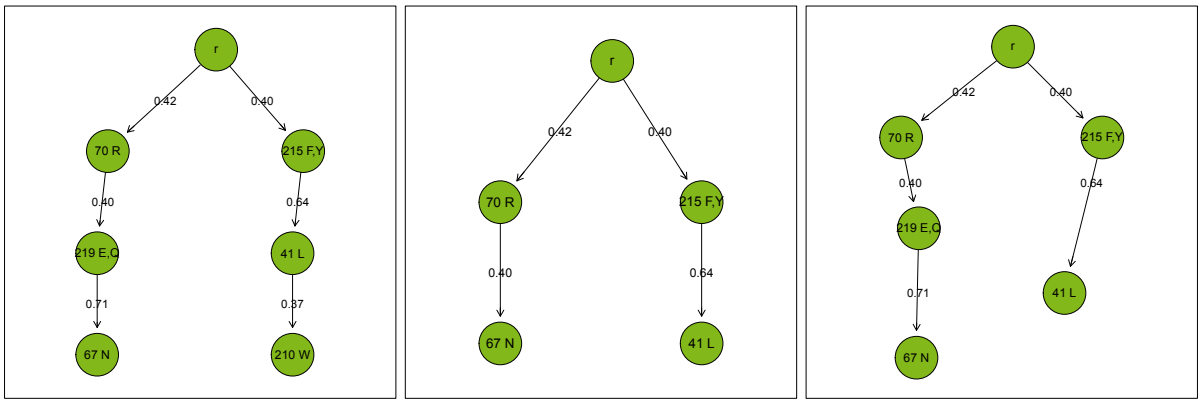



Abbildung 5.18: Onkogenetische Baummodelle für den HIV Datensatz. Links der Baum mit allen Ereignissen, in der Mitte und rechts die Bäume basierend auf der Häufigkeits- bzw. Cliquen-Selektion (größte Cliquen, niedriger Schwellenwert).

Fortschreiten der Krankheit in Verbindung gebracht werden (Ohgaki und Kleihues, 2007). Diese werden daher als Referenz angesehen und mit den Ergebnissen der Variablenselektionsmethoden verglichen.

Beim Anwenden der Variablenselektionsmethoden auf diesen Datensatz fällt auf, dass die Verfahren mit den Schwellenwerten aus der Simulationsstudie oft zu einer sehr großen Anzahl an ausgewählten Ereignissen führen (siehe Tabelle 5.7). Erneut erzielen hier die Häufigkeits- sowie die Cliquenmethoden die sinnvollsten Ergebnisse. Sie wählen zwischen 10 und 29 Ereignissen aus, was eine realisierbare Größenordnung für onkogenetische Baummodelle darstellt. Alle anderen Methoden wählen mindestens 73 Ereignisse aus. Eine solche Anzahl ist jedoch ungeeignet und führt zu sehr instabilen Bäumen.

Aufgrund dieser Beobachtungen soll nun die Anzahl an ausgewählten Ereignissen auf ungefähr 11 beschränkt werden. Diese Zahl basiert darauf, dass die Methode der maximalen Cliquen mit größerem Schwellenwert genau 11 Ereignisse auswählt und sie sich sowohl in der Simulationsstudie als auch bei der Anwendung auf die vorangegangenen Datensätze bewährt hat. In dieser Situation wird der größere Schwellenwert herangezogen, da man bei insgesamt 132 Ereignissen einen höheren Anteil des Rauschens annehmen kann.

Um nun ungefähr 11 Ereignisse auszuwählen, müssen folgende Schwellenwerte für die einzelnen Verfahren verwendet werden:

$$\begin{aligned}
\tau_{\text{freq}} &= 0.41 \\
\tau_{\text{cor}} &= 0.70 \\
\tau_{\text{fisher}} &= 10^{-26} \\
\tau_{\text{weight}} &= 0.0018 \\
\tau_{\text{OT}} &= 0.90 \\
\tau_{\text{lcliq}} &= 0.20 \\
\tau_{\text{mcliq}} &= 0.15
\end{aligned}$$

Tabelle 5.7: Anzahl der Ereignisse des Glioblastom Datensatzes, die von den einzelnen Variablenselektionsverfahren mit ihren jeweils besten Schwellenwerten ausgewählt wurden.

| Meth. | freq | brod | cor | fisher | z | weight | weight | OT | single | lcliq | lcliq | mcliq | mcliq |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\tau$ | 0.2 | 0.1 | 0.3 | 0.01 | 0.9 | 0.05 | 0.3 | 0.25 | - | 0.05 | 0.20 | 0.05 | 0.15 |
| | 23 | 29 | 73 | 99 | 102 | 89 | 102 | 85 | 131 | 22 | 10 | 22 | 11 |

Die z-Transformation wird hier von Beginn an ausgeschlossen, da selbst bei einem Schwellenwert von $\tau_z = 1 - 10^{-16}$ immer noch 60 Ereignisse ausgewählt werden und diese Methode somit keine geeignete Wahl darstellt. Die Methode von Brodeur wählt bei einem implizit berechneten Schwellenwert von $\tau^*_{\text{freq}} = 0.1725$ 29 Ereignisse aus. Die Methode der Unabhängigkeit im Baum, bei der auch kein Schwellenwert vorgegeben werden kann, entfernt nur eins der 132 Ereignisse und ist somit ungeeignet und wird ebenfalls von der weiteren Analyse ausgeschlossen.

Die Ergebnisse der Variablenselektionsverfahren mit den oben angegebenen Schwellenwerten sind in Tabelle 5.8 aufgeführt.

Die Methode der univariaten Häufigkeit wählt sieben der acht in der Literatur (Ohgaki und Kleihues, 2007) bekannten Ereignisse aus und ein paar zusätzliche darüber hinaus. Da die Häufigkeitsauswahl als Variablenselektionsmethode sehr verbreitet ist, ist dieses Ergebnis nicht überraschend. Die Methode von Brodeur selektiert insgesamt 29 Ereignisse, darunter alle acht Ereignisse aus der Literatur. Die Methoden, die auf der paarweisen Korrelation, den bedingten Wahrscheinlichkeiten im onkogenetischen Baum und den Gewichten in Edmonds' Algorithmus basieren, erkennen nur eins bzw. gar kein bekanntes Ereignis, während der exakte Test von Fisher die Hälfte dieser Ereignisse auswählt. Die beiden Cliquenmethoden erfassen fast alle in der Literatur bekannten Ereignisse. Nur das Ereignis $13q-$ (und bei den größten Cliquen zusätzlich $1p-$ und $22q-$) ist nicht in der Auswahl enthalten.

Auch hier sollen wieder verschiedene angepasste Baummodelle betrachtet werden. Dazu wird in Abbildung 5.19 das Modell basierend auf allen in der Literatur bekannten Ereignissen mit den Modellen der Häufigkeitsselektion sowie der maximalen Cliquen verglichen. Das Baummodell basierend auf den Ereignissen aus der Literatur ist exakt im Baummodell der Auswahl nach univariater Häufigkeit enthalten. Dadurch, dass insgesamt mehr Ereignisse ausgewählt werden, ändert sich nicht die Abhängigkeitsstruktur. Lediglich der Pfad $10q- \rightarrow 7p+$ ist insofern anders, als dass das Ereignis $7q+$ in der Mitte eingefügt wurde. Zwei weitere zusätzliche Ereignisse sind als unabhängig direkt am Wurzelknoten angeschlossen und ein drittes ergänzt einen bereits bestehenden Pfad.

Das Baummodell basierend auf der Selektion von maximalen Cliquen enthält ebenfalls die Struktur des Baums der Literatur-Ereignisse (erneut mit der Ergänzung $7q+$ in einem Pfad). Lediglich das bekannte Ereignis $13q-$ fehlt. Diesem ist aber nicht die größte Bedeutung beizumessen, da es sich hierbei um ein unabhängiges Ereignis handelt. Es geht also keine Information über einen wichtigen Pfad verloren. Die weiteren selektierten Ereignisse ergänzen den bereits bestehenden Pfad von $9p-$ und $7q++$. Das bedeutet, dass die Ereignisse $19p+$, $20p+$ und $20q+$ weitere Informationen über das Fortschreiten des Glioblastoms beinhalten

Tabelle 5.8: Auflistung der Ereignisse des Glioblastom Datensatzes, die von den einzelnen Variablenselektionsverfahren mit den angegebenen Schwellenwerten ausgewählt wurden ($x$ = Ereignis wurde ausgewählt). Die Ereignisse sind entsprechend ihrer Auswahlhäufigkeit sortiert. Ereignisse, die in der Literatur erwähnt werden, sind fett markiert.

| Methode | freq | brod | cor | fisher | weight | OT | lcliq | mcliq |
|---|---|---|---|---|---|---|---|---|
| $\tau$ | 0.41 | 0.1725 | 0.70 | $10^{-26}$ | 0.0018 | 0.90 | 0.2 | 0.15 |
| **Chr7p+** | x | x | x | x | | x | x | x |
| Chr7q+ | x | x | x | x | | x | x | x |
| Chr19p+ | x | x | x | x | | | x | x |
| Chr20p+ | | x | x | x | | x | x | x |
| Chr20q+ | | x | x | x | | x | x | x |
| **Chr10p-** | x | x | | x | | | x | x |
| **Chr10q-** | x | x | | x | | | x | x |
| **Chr7p++** | x | x | | x | | | x | x |
| **Chr9p-** | x | x | | | | | x | x |
| Chr19q+ | | x | x | x | | | x | |
| Chr9q++ | | | x | | x | x | | |
| Chr12p++ | | | x | | x | x | | |
| Chr18p++ | | | x | | x | x | | |
| Chr18q++ | | | x | | x | x | | |
| Chr21q++ | | | x | | x | x | | |
| **Chr22q-** | x | x | | | | | | x |
| **Chr1p-** | | x | | | | | | x |
| Chr2q++ | | | x | | x | | | |
| Chr3p++ | | | | | x | x | | |
| Chr8q++ | | | | | x | x | | |
| Chr11p- | | x | | x | | | | |
| Chr11q- | | x | | x | | | | |
| **Chr13q-** | x | x | | | | | | |
| Chr14q- | x | x | | | | | | |
| Chr15q- | x | x | | | | | | |
| Chr1q+ | | x | | | | | | |
| Chr1q- | | x | | | | | | |
| Chr3q- | | x | | | | | | |
| Chr4q- | | x | | | | | | |
| Chr6p- | | x | | | | | | |
| Chr6q- | | x | | | | | | |
| Chr8p- | | x | | | | | | |
| Chr9q- | | x | | | | | | |
| Chr12q+ | | x | | | | | | |
| Chr12q- | | x | | | | | | |
| Chr15q+ | | x | | | | | | |
| Chr21p- | | x | | | | | | |
| Chr7q- | | | | | x | | | |
| Chr13q+ | | | | | x | | | |
| Chr18p+ | | | | | x | | | |

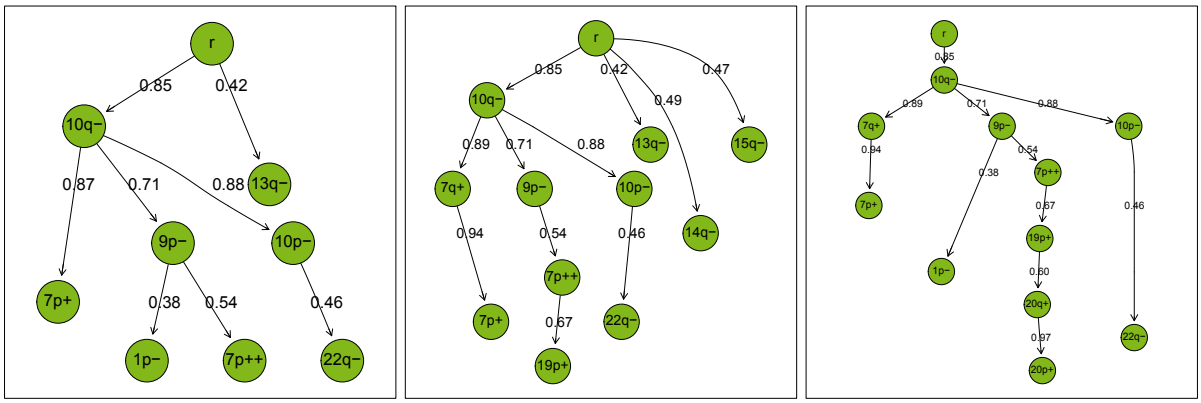
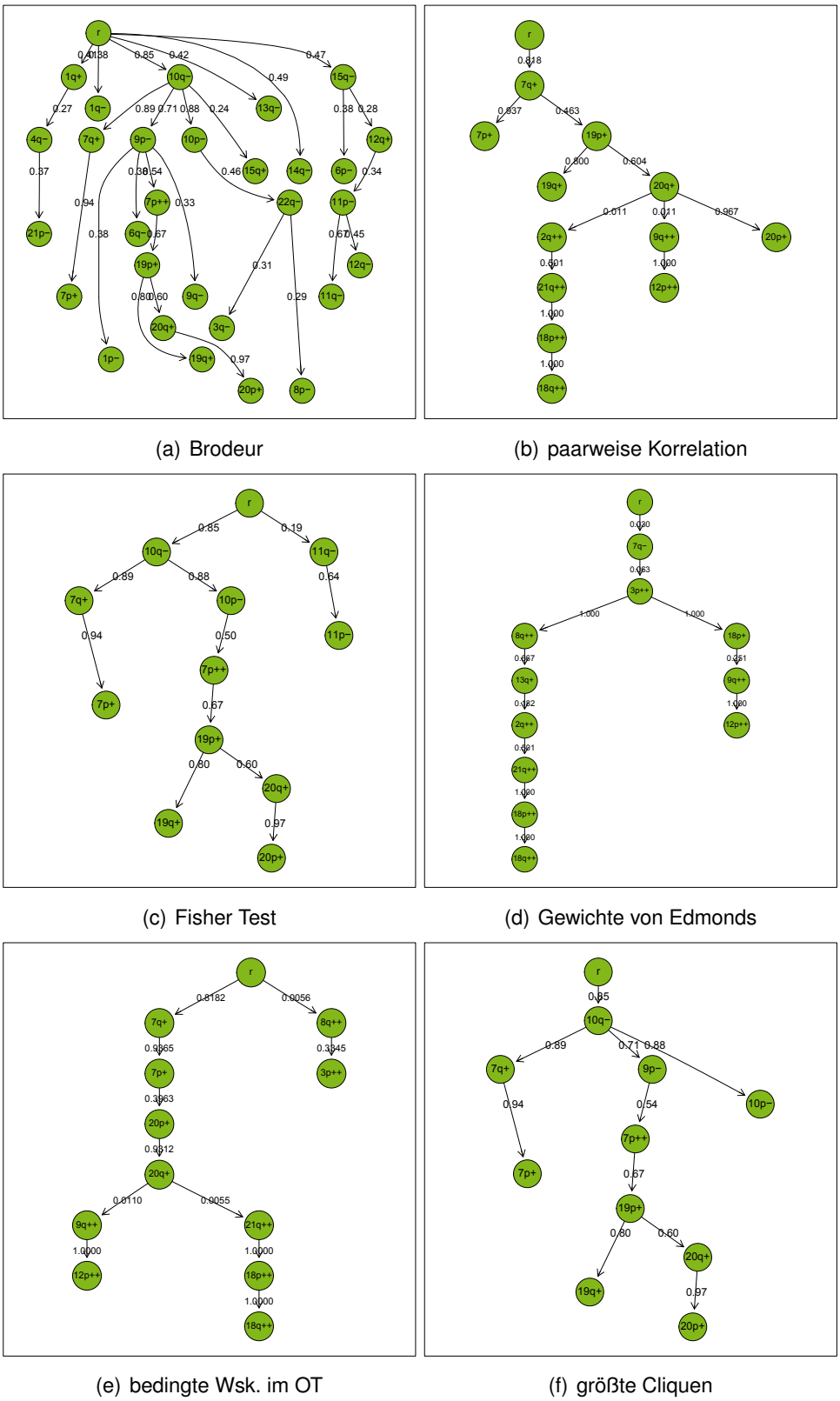



Abbildung 5.19: Onkogenetische Baummodelle für den Glioblastom Datensatz. Links der Baum mit allen Ereignissen, die in der Literatur bekannt sind, in der Mitte und rechts die Bäume basierend auf der Häufigkeits- bzw. Cliquen-Selektion (maximale Cliquen, größerer Schwellenwert).

können. Außerdem werden diese drei Ereignisse von sechs der acht Variablenselektionsverfahren ausgewählt. Nur zwei andere Ereignisse werden noch häufiger selektiert. Die Methode der univariaten Häufigkeit, die als Standard Methode bezeichnet werden kann, erkennt nur eins dieser drei Ereignisse. Daher überzeugt die Cliquenmethode erneut, da sowohl die wichtigen bekannten Ereignisse aus der Literatur als auch vielversprechende neue Ereignisse ausgewählt werden.

Zusätzlich sollen hier auch die Baummodelle der anderen Variablenselektionsverfahren betrachtet werden (siehe Abbildung 5.20). Das Baummodell mit den Ereignissen, die nach der Brodeur-Methode ausgewählt wurden, enthält alle Pfade aus dem Baum der maximalen Cliquen, aber zusätzlich noch viele darüber hinaus. Daher ist es schwierig, die wichtigsten Pfade und Ereignisse zu identifizieren. Der Baum basierend auf den Ergebnissen der paarweisen Korrelation enthält nur zwei Ereignisse aus der Literatur, aber auch den neuen Pfad $19p+ \rightarrow 20p+ \rightarrow 20q+$. Die restlichen sechs Ereignisse sind zwar hoch korreliert (Kantengewicht 1), treten jedoch fast nie auf (Kantengewicht 0.011). Die Fisher-Test-Methode schneidet nur knapp schlechter ab als die maximalen Cliquen und das zugehörige Baummodell enthält somit die wichtigsten Pfade. Die Variablenselektionsmethode, die auf den Gewichten von Edmonds basiert, ist unbrauchbar, da das initiale Ereignis nur in 3% aller Fälle auftritt. Andere Ereignisse des Baums sind zwar hoch korreliert, trotzdem kann dieses Baummodell keine belastbaren Aussagen über die Krankheitsprogression treffen. Dasselbe gilt für den Baum, der anhand der bedingten Wahrscheinlichkeiten geschätzt wurde. Es wurden zwar zwei bekannte Ereignisse aus der Literatur ausgewählt, aber sieben Ereignisse sind in Pfaden mit zu geringem Kantengewicht enthalten. Die Methode der größten Cliquen liefert sehr ähnliche Ergebnisse wie die maximalen Cliquen und deckt somit die wichtigen Ereignisse und Pfade ab.




(a) Brodeur

(b) paarweise Korrelation

(c) Fisher Test

(d) Gewichte von Edmonds

(e) bedingte Wsk. im OT

(f) größte Cliquen

Abbildung 5.20: Onkogenetische Baummodelle der übrigen Variablenselektionsmethoden für den Glioblastom Datensatz.

## 5.4 Abschließende Bemerkungen

Sowohl in der Simulationsstudie als auch bei der Anwendung auf echte Datensätze schneiden die Cliquenmethoden gut ab. Um die Vergleichbarkeit zwischen simulierten und echten Daten zu veranschaulichen, wird die Verteilung der Ereigniswahrscheinlichkeiten betrachtet. Tabelle 5.9 gibt einen Überblick über die entsprechenden Häufigkeiten.

Insgesamt treten in der Simulationsstudie Ereignisse etwas häufiger auf als bei den echten Datensätzen. Anhand der Werte lässt sich jedoch nicht schließen, dass die Ergebnisse der Simulationsstudie sich nicht auf echte Datensätze übertragen lassen.

Weiterhin ist zu bemerken, dass die Variablenselektionsverfahren hier zunächst nur für die grundlegende und weit verbreitete Klasse der onkogenetischen Baummodelle betrachtet wurden. Mit diesen Bäumen kann nicht jede denkbare Ereigniskombination modelliert werden, so dass einige Beobachtungen nicht zum geschätzten Modell passen. Einen Überblick darüber, welcher Anteil der Daten zum berechneten Modell passt, liefert Tabelle 5.10. Trotzdem ist es wichtig, dass neue Variablenselektionsmethoden zunächst für ein einfaches, grundlegendes Modell analysiert werden, um die wesentlichen charakteristischen Eigenschaften zu verstehen. In einem nächsten Schritt sollten die Variablenselektionsverfahren natürlich auf andere Modellklassen erweitert werden.

Alle vorgeschlagenen Variablenselektionsverfahren lassen sich auf andere Modellklassen erweitern (teilweise mit kleiner Modifikation). Die Methoden `freq`, `brod` und `clique` können direkt übernommen werden. Bei den korrelationsbasierten Verfahren `cor`, `fisher` und `z` muss berücksichtigt werden, dass diese Methoden paarweise Interaktionen untersuchen. Für Modellklassen, die z.B. mehrere Elternknoten und somit mehrdimensionale Abhängigkeiten erlauben, können trotzdem die bestehenden Methoden verwendet werden oder sie müssen alternativ für höhere Interaktionen erweitert werden. Die Methoden `weight`, `OT` und `single` sind spezifisch für onkogenetische Bäume. Sehr leicht lässt sich jedoch die Methode der bedingten Wahrscheinlichkeiten im onkogenetischen Baum modifizieren, indem nicht länger die Kantengewichte, sondern einfach die bedingten Wahrscheinlichkeiten verwendet werden. Bei CBNs stehen diese z.B. nicht mehr an den Kanten, sondern an den Knoten. Die Methode der Unabhängigkeit im Baum lässt sich weiterhin unverändert anwenden, da in jedem Modell einzelne Knoten als Nachfolger der Wurzel identifiziert werden können. Für die Gewichte von Edmonds Algorithmus gilt zwar, dass diese speziell zum Schätzen von onkogenetischen Bäumen verwendet werden,

Tabelle 5.9: Überblick über die Auftrittshäufigkeiten aller Ereignisse für die Daten der Simulationsstudie (mit und ohne Rauschvariablen) und die drei Anwendungsbeispiele

| Datensatz | Minimum | 1. Quartil | Medium | Mittelwert | 3. Quartil | Maximum |
|---|---|---|---|---|---|---|
| Simulation | 0.00 | 0.14 | 0.29 | 0.33 | 0.49 | 0.96 |
| Simulation mit Rauschen | 0.00 | 0.12 | 0.23 | 0.26 | 0.36 | 0.96 |
| Meningiom | 0.02 | 0.07 | 0.04 | 0.08 | 0.06 | 0.38 |
| HIV | 0.12 | 0.20 | 0.24 | 0.27 | 0.36 | 0.42 |
| Glioblastom | 0.00 | 0.00 | 0.04 | 0.12 | 0.14 | 0.85 |

Tabelle 5.10: Anteil der Beobachtungen aus den drei Datensätzen, der zum geschätzten Baummodell passt. Für den Glioblastom Datensatz sind die Werte etwas niedriger aufgrund der Baumtiefe von 4 bzw. 6 für `freq` bzw. `cliq`. Für die simulierten Daten lauten Minimum, 1. Quartil, Median, Mittelwert, 3. Quartil und Maximum wie folgt: 0.47, 0.94, 0.99, 0.92, 1.00 und 1.00.

| Datensatz | alle Ereignisse | Häufigkeitsselektion | Cliquenselektion |
|---|---|---|---|
| Meningiom | 0.90 | 1.00 | 0.97 |
| HIV | 0.87 | 0.94 | 0.88 |
| Glioblastom | 0.79 | 0.69 | 0.59 |

da es sich jedoch lediglich um bestimmte Wahrscheinlichkeiten handelt, können sie selbstverständlich für beliebige Datensätze berechnet werden. Zusätzlich ist es natürlich denkbar, weitere Variablenselektionsverfahren zu entwickeln, die auf speziellen Eigenschaften eines konkreten Progressionsmodells beruhen.

# Kapitel 6

# Zusammenfassung und Ausblick

In dieser Arbeit geht es darum, Modelle zur Beschreibung von Krankheitsprogression genauer zu betrachten und mit statistischen Methoden zu vergleichen. Ein tieferes Verständnis der Entstehung und des Verlaufs einer Krankheit ist unerlässlich für eine geeignete medizinische Behandlung. Die genaue Kenntnis darüber, in welchem Stadium des Krankheitsprozesses sich ein Patient befindet, ist bei individuellen Therapieentscheidungen sehr hilfreich. Daher ist es von großer Bedeutung, einzelne Schritte der Krankheitsprogression benennen zu können. Die in dieser Arbeit vorgestellten Progressionsmodelle haben das Ziel, die Abhängigkeitsstruktur der aufgetretenen genetischen Ereignisse zu schätzen und so charakteristische Ereignispfade anzugeben. Die zugrunde liegenden genetischen Ereignisse sind dabei als binäre Zufallsvariablen definiert und beschreiben aufgetretene Mutationen an Chromosomen, Chromosomenarmen oder sogar einzelnen Genen.

In Kapitel 2 werden einige Krankheitsprogressionsmodelle mit ihren Eigenschaften und Schätzalgorithmen vorgestellt. Einige Modelle bauen aufeinander auf, einige wurden jedoch auch ganz unabhängig voneinander entwickelt. Ein intensiver Vergleich verschiedener Modelle, der die Vor- und Nachteile einzelner Modellklassen herausstellt, ist bisher jedoch nicht durchgeführt worden. In Kapitel 3 werden vier verschiedene Modellklassen plus einzelne Unterkategorien mit Hilfe einer Simulationsstudie verglichen. Es stellt sich heraus, dass verbindende Bayes-Netze (CBNs) den onkogenetischen Bäumen (OTs) immer überlegen sind. Dies war zu erwarten, da CBNs alle Eigenschaften eines OTs besitzen, zusätzlich aber noch weitere Abhängigkeitsstrukturen modellieren können. Ein Vergleich zwischen CBNs und onkogenetischen Baum-Mischungs-Modellen (OTMs) fällt jedoch schwerer, da beide Modellklassen bestimmte Eigenschaften aufweisen, die die andere nicht modellieren kann. CBNs können Abhängigkeiten von mehreren vorangegangenen Ereignissen modellieren, während OTMs durch ihre Sternkomponente jeder Ereigniskombination eine positive Wahrscheinlichkeit zuweisen und somit keine Beobachtungen auftreten können, die mit dem Modell nicht zu erklären sind. Distanzbäume weisen eine andere Baumstruktur auf und sind damit schlecht mit den anderen Modellen zu vergleichen bzw. schneiden nur in den Fällen gut ab, in denen die Daten auch aus einem Distanzbaum stammen.

Für weitergehende Untersuchungen ist es aufschlussreich, noch andere Modellklassen in den Vergleich mit einzubeziehen. Dazu müssen einige Schätzalgorithmen jedoch zunächst noch programmiert werden. Auch ein Vergleich unterschiedlicher Schätzalgorithmen für die gleiche Modellklasse ist interessant, da z.B. für einen onkogenetischen Baum im Laufe der

Zeit verschiedene Algorithmen vorgeschlagen wurden. Denkbar ist auch, den Vergleich mit weiteren Abstandsmaßen durchzuführen. Auch wenn die Ergebnisse der induzierten Wahrscheinlichkeitsverteilung und der *graph edit distance* weitestgehend übereinstimmen, können Maße wie *recovery, precision* und *dissimilarity* weitere Eigenschaften einzelner Modellklassen gegebenenfalls genauer benennen bzw. sie noch besser voneinander abgrenzen. Weiterhin könnte auch der Aufbau der Simulationsstudie insofern verändert werden, als dass man als zugrunde liegendes wahres Modell nicht nur einen Vertreter aus jeder Modellklasse wählt, sondern gezielt schwerer werdende Modelle gestaltet und beobachtet, ab wann eine Modellklasse nicht mehr in der Lage ist, die Abhängigkeitsstruktur der Ereignisse geeignet wiederzugeben.

Ist die wahre Modellklasse jedoch nicht bekannt, ist es wünschenswert, anhand eines geeigneten Kriteriums die beste Modellklasse zum Schätzen des Krankheitsverlaufs auszuwählen. Hierzu werden in Kapitel 4 verschiedene Modellwahlkriterien miteinander verglichen. Anhand der durchgeführten Simulationsstudie lässt sich jedoch kein eindeutig bestes Kriterium bestimmen, das in den meisten Fällen eine gute Entscheidung trifft. Für die Modellklassen mit nur einer Baumkomponente treffen alle vier betrachteten Kriterien eine gute Entscheidung, für die OTMs lässt sich jedoch kein eindeutig bestes Kriterium angeben.

Bevor überhaupt ein Progressionsmodell angepasst werden kann, müssen geeignete genetische Ereignisse ausgewählt werden. Selbst wenn z.B. für einen Krebsdatensatz nur Veränderungen an ganzen Chromosomen berücksichtigt werden, liegen schon 48 Ereignisse vor, wenn man die aufgetretenen Mutationen in Gewinn und Verlust an Erbinformation unterteilt. Ein Modell mit so vielen Ereignissen ist jedoch instabil und außerdem schwer zu interpretieren. Man möchte sich daher auf die etwa 5 bis 15 wichtigsten Ereignisse beschränken. Dazu müssen aber die zufällig aufgetretenen Mutationen von den für die Krankheit spezifischen unterschieden werden. Bisher wurde dazu lediglich eine Selektion nach Häufigkeit durchgeführt. Da jedoch Ereignisse, die spät in einem Ereignispfad auftreten, eher eine geringe Wahrscheinlichkeit besitzen, besteht die Gefahr, dass diese niemals mit ausgewählt werden. In Kapitel 5 werden daher diverse Variablenselektionsverfahren vorgestellt, die anhand unterschiedlichster Kriterien geeignete Ereignisse auswählen sollen. Anhand einer Simulationsstudie sowie auch auf konkreten Datensätzen werden diese miteinander verglichen. Es stellt sich heraus, dass die Identifikation von Cliquen sehr vielversprechende Ergebnisse liefert. Sowohl bereits in der Literatur bekannte Ereignisse als auch Erweiterungen einzelner Pfade werden identifiziert.

Der Vergleich verschiedener Variablenselektionsverfahren wurde in dieser Arbeit nur anhand der grundlegenden Modellklasse der onkogenetischen Bäume durchgeführt. Natürlich sollten diese Ergebnisse in einem nächsten Schritt auch für andere Progressionsmodelle überprüft werden.

Ein interessantes Konzept, das für onkogenetische Baum-Mischungs-Modelle entwickelt wurde (Rahnenführer et al., 2005), ist der genetische Progressionsscore (GPS). Zusätzlich zur Identifikation einzelner Krankheitsschritte möchte man eine Aussage darüber treffen, wie weit die Krankheit bei einem Patienten fortgeschritten ist. Bei Krebs werden dazu oft unterschiedliche Stadien angeben, aber ein GPS als stetige Variable könnte diese Information noch genauer

wiedergeben und ist auch eine relevante Variable für die Vorhersage der Überlebenszeit. Auch für andere Progressionsmodelle ist es denkbar, einen solchen genetischen Progressionsscore zu entwickeln. Ein Vergleich des GPS mit den bekannten Stadien eines Tumors sollte dabei auch anhand mehrerer Datensätze überprüft werden.

Insgesamt ist die Beschreibung der Krankheitsprogression und der damit verbundene Erkenntnisgewinn von großer Bedeutung, so dass die Analyse und Weiterentwicklung der hier beschriebenen Modelle ein wichtiges Forschungsthema bleiben sollte.

# Literaturverzeichnis


Agarwala, R., Bafna, V., Farach, M., Paterson, M., und Thorup, M. (1999). On the approximability of numerical taxonomy (fitting distances by tree metrics). *SIAM Journal of Computation*, 28(3):1073–1085.

Agresti, A. (1992). A survey of exact inference for contingency tables. *Statistical Science*, 7(1):131–153.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. N. und Caski, F., Herausgeber, *Proceedings of the Second International Symposium on Information Theory*, Seiten 267–281. Akademiai Kiado.

Attolini, C. S.-O., Cheng, Y.-K., Beroukhim, R., Getz, G., Abdel-Wahab, O., Levine, R. L., Mellinghoff, I. K., und Michor, F. (2010). A mathematical framework to determine the temporal sequence of somatic genetic events in cancer. *PNAS*, 107(41):17604–17609.

Beerenwinkel, N. und Drton, M. (2005). Mutagenetic tree models. In Pachter, L. und Sturmfels, B., Herausgeber, *Algebraic Statistics for Computational Biology*, Kapitel 14, Seiten 278–290. Cambridge University Press.

Beerenwinkel, N. und Drton, M. (2007). A mutagenetic tree hidden Markov model for longitudinal clonal HIV sequence data. *Biostatistics*, 8(1):53–71.

Beerenwinkel, N., Eriksson, N., und Sturmfels, B. (2007). Conjunctive Bayesian networks. *Bernoulli*, 13(4):893–909.

Beerenwinkel, N., Rahnenführer, J., Däumer, M., Hoffmann, D., Kaiser, R., Selbig, J., und Lengauer, T. (2005). Learning multiple evolutionary pathways from cross-sectional data. *Journal of Computational Biology*, 12(6):584–598.

Beerenwinkel, N., Schwarz, R. F., Gerstung, M., und Markowetz, F. (2015). Cancer evolution: Mathematical models and computational inference. *Systematic Biology*, 64:e1–e25.

Beerenwinkel, N. und Sullivant, S. (2009). Markov models for accumulating mutations. *Biometrika*, 96:663–676.

Bogojeska, J. (2016). *Rtreemix: Mutagenetic trees mixture models.* R package version 1.36.0.

Bogojeska, J., Alexa, A., Altmann, A., Lengauer, T., und Rahnenführer, J. (2008). Rtreemix: an R package for estimating evolutionary pathways and genetic progression scores. *Bioinformatics*, 24(20):2391–2392.

Brodeur, G. M., Tsiatis, A. A., Williams, D. L., Luthardt, F. W., und Green, A. A. (1982). Statistical analysis of cytogenetic abnormalities in human cancer cells. *Cancer Genetics and Cytogenetics*, 7:137–152.

Bunke, H. und Shearer, K. (1998). A graph distance metric based on the maximal common subgraph. *Pattern Recognition Letters*, 19:225–259.

Cheng, Y.-K., Beroukhim, R., Levine, R. L., Mellinghoff, I. K., Holland, E. C., und Michor, F. (2012). A mathematical methodology for determining the temporal order of pathway alterations arising during gliomagenesis. *PLoS Computational Biology*, 8(1).

Chernoff, H. (1952). A measure of asymptotic efficiency of tests for a hypothesis based on a sum of observations. *The Annals of Mathematical Statistics*, 23(4):493–507.

Claeskens, G. und Hjort, N. L. (2008). *Model selection and model averaging*. Cambridge University Press.

Conover, W. (1999). *Practical nonparametric statistics*. John Wiley & Sons, 3. Auflage.

Cristea, S., Kuipers, J., und Beerenwinkel, N. (2017). pathTiMEx: Joint inference of mutually exclusive cancer pathways and their progression dynamics. *Journal of Computational Biology*, 24(6):603–615.

Csardi, G. und Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695:1–9.

Desper, R., Jiang, F., Kallioniemi, O.-P., Moch, H., Papadimitriou, C. H., und Schäffer, A. A. (1999). Inferring tree models for oncogenesis from comparative genome hybridization data. *Journal of Computational Biology*, 6(1):37–52.

Desper, R., Jiang, F., Kallioniemi, O.-P., Moch, H., Papadimitriou, C. H., und Schäffer, A. A. (2000). Distance-based reconstruction of tree models for oncogenesis. *Journal of Computational Biology*, 7(6):789–803.

Edmonds, J. (1967). Optimum branchings. *Journal of Research of the National Bureau of Standards B*, 71:233–240.

Farach, M. und Kannan, S. (1996). Efficient algorithms for inverting evolution. In *Proceedings of the ACM Symposium on the Foundations of Computer Science*, Seiten 230–236.

Fearon, E. R. und Vogelstein, B. (1990). A genetic model for colorectal tumorigenesis. *Cell*, 61:759–767.

Geiger, D., Heckerman, D., und Meek, C. (1996). Asymptotic model selection for directed networks with hidden variables. In *Proceedings of the 12th Conference on Uncertainty in Artificial Intelligence*, Seiten 283–290.

Gerstung, M., Baudis, M., Moch, H., und Beerenwinkel, N. (2009). Quantifying cancer progression with conjunctive Bayesian networks. *Bioinformatics*, 25(21):2809–2815.

Hainke, K., Rahnenführer, J., und Fried, R. (2012). Cumulative disease progression models for cross-sectional data: A review and comparison. *Biometrical Journal*, 54(5):617–640.

Hainke, K., Szugat, S., Fried, R., und Rahnenführer, J. (2017). Variable selection for disease progression models: methods for oncogenetic trees and application to cancer and HIV. *BMC Bioinformatics*, 18:358.

Hartung, J., Knapp, G., und Sinha, B. K. (2008). *Statistical Meta-Analysis with Applications*. John Wiley & Sons.

Hastie, T., Tibshirani, R., und Friedman, J. (2009). *The Elements of Statistical Learning*. Springer.

Hjelm, M., Höglund, M., und Lagergren, J. (2006). New probabilistic network models and algorithms for oncogenesis. *Journal of Computational Biology*, 13(4):853–865.

Huang, Z., Desper, R., Schäffer, A. A., Yin, Z., Li, X., und Yao, K. (2004). Construction of tree models for pathogenesis of nasopharyngeal carcinoma. *Genes, Chromosomes & Cancer*, 40:307–315.

Kullback, S. und Leibler, R. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22:79–86.

Li, X. (2009). Mathematical modeling of carcinogenesis based on chromosome aberration data. *Chinese Journal of Cancer Research*, 21(3):240–246.

Loohuis, L. O., Caravagna, G., Graudenzi, A., Ramazzotti, D., Mauri, G., Antoniotti, M., und Mishra, B. (2014). Inferring tree causal models of cancer progression with probability raising. *PLoS ONE*, 9(10).

Ohgaki, H. und Kleihues, P. (2007). Genetic pathways to primary and secondary glioblastoma. *The American Journal of Pathology*, 170:1445–1453.

Pigott, T. D. (2012). *Advances in Meta-Analysis*. Springer, New York.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rachev, S. T. (1991). *Probability metrics and the stability of stochastic models*. John Wiley & Sons.

Radmacher, M. D., Simon, R., Desper, R., Taetle, R., Schäffer, A. A., und Nelson, M. A. (2001). Graph models of oncogenesis with an application to melanoma. *Journal of Theoretical Biology*, 212:535–548.

Rahnenführer, J., Beerenwinkel, N., Schulz, W. A., Hartmann, C., von Deimling, A., Wullich, B., und Lengauer, T. (2005). Estimating cancer survival and clinical outcome based on genetic tumor progression scores. *Bioinformatics*, 21(10):2438–2446.

Ramazzotti, D., Caravagna, G., Loohuis, L. O., Graudenzi, A., Korsunsky, I., Mauri, G., Antoniotti, M., und Mishra, B. (2015). CAPRI: efficient inference of cancer progression models from cross-sectional data. *Bioinformatics*, 31(18):3016–3026.

Robert, C. P. (2001). *The Bayesian choice: From decision-theoretic foundations to computational implementation*. Springer, New York.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464.

Shahrabi Farahani, H. und Lagergren, J. (2013). Leaning oncogenetic networks by reducing to mixed integer linear programming. *PLoS ONE*, 8(6).

Simon, R., Desper, R., Papadimitriou, C. H., Peng, A., Alberts, D. S., Taetle, R., Trent, J. M., und Schäffer, A. A. (2000). Chromosome abnormalities in ovarian adenocarcinoma: III. using breakpoint data to infer and test mathematical models for oncogenesis. *Genes, Chromosomes & Cancer*, 28:106–120.

Spivak, M. (1979). *A Comprehensive Introduction to Differential Geometry 1*. Publish or Perish.

Suppes, P. (1970). *A probabilistic theory of causality*. North-Holland Publishing company.

Tofigh, A. (2009). *Using trees to capture reticulate evolution: Lateral gene transfers and cancer progression*. Dissertation, KTH School of Computer Science and Communications, Stockholm, Sweden.

Tofigh, A., Sjölund, E., Höglund, M., und Lagergren, J. (2011). A global structural EM algorithm for a model of cancer progression. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, NIPS'11, Seiten 163–171, USA. Curran Associates Inc.

Tolosi, L. (2011). *Finding regions of aberrant DNA copy number associated with tumor phenotype*. Dissertation, Saarland University.

Tolosi, L., Theißen, J., Halachev, K., Hero, B., Berthold, F., und Lengauer, T. (2013). A method for finding consensus breakpoints in the cancer genome from copy number data. *Bioinformatics*, 29:1793–1800.

Urbschat, S., Rahnenführer, J., Henn, W., Feiden, W., Wemmert, S., Linsler, S., Zang, K. D., Oertel, J., und Ketter, R. (2011). Clonal cytogenetic progression within intratumorally heterogeneous meningiomas predicts tumor recurrence. *International Journal of Oncology*, 39:1601–1608.

Vogelstein, B., Fearon, E. R., Hamilton, S. R., Kern, S. E., Preisinger, A. C., Leppert, M., Nakamura, Y., White, R., Smits, A. M., und Bos, J. L. (1988). Genetic alterations during colorectal-tumor development. *New England Journal of Medicine*, 319(9):525–532.

von Heydebreck, A. (2003). *oncomodel: Maximum likelihood tree models for oncogenesis*. R package version 0.8.

von Heydebreck, A., Gunawan, B., und Füzesi, L. (2004). Maximum likelihood estimation of oncogenetic tree models. *Biostatistics*, 5(4):545–556.

Wallis, W. D. (2000). *A beginner's guide to graph theory*. Birkhäuser.

Yin, J., Beerenwinkel, N., Rahnenführer, J., und Lengauer, T. (2006). Model selection for mixtures of mutagenetic trees. *Statistical Applications in Genetics and Molecular Biology*, 5(1):17.

# Anhang

## A.1 Vergleich der Modelle

### A.1.1 Ausgewählte zugrunde liegende wahre Modelle für die Simulationsstudie

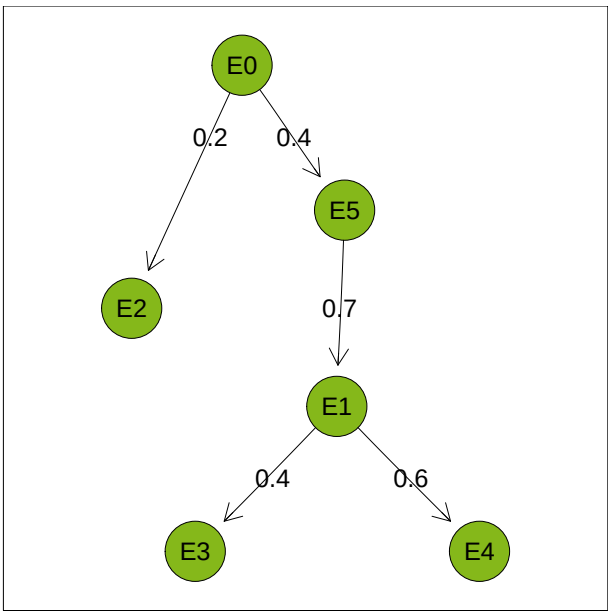



Abbildung A.1: Ausgewähltes wahres Modell aus der Klasse der onkogenetischen Bäume

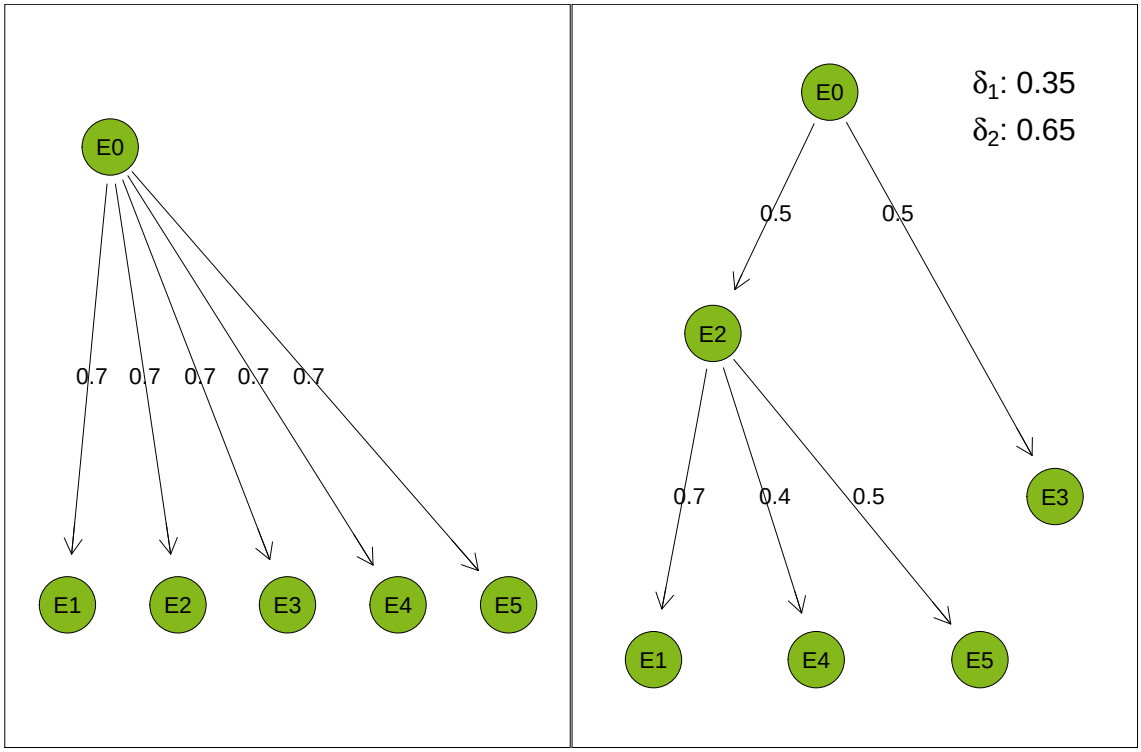



Abbildung A.2: Ausgewähltes wahres Modell aus der Klasse der Baum-Mischungs-Modelle mit 2 Komponenten und gleichen Kantengewichten beim Stern

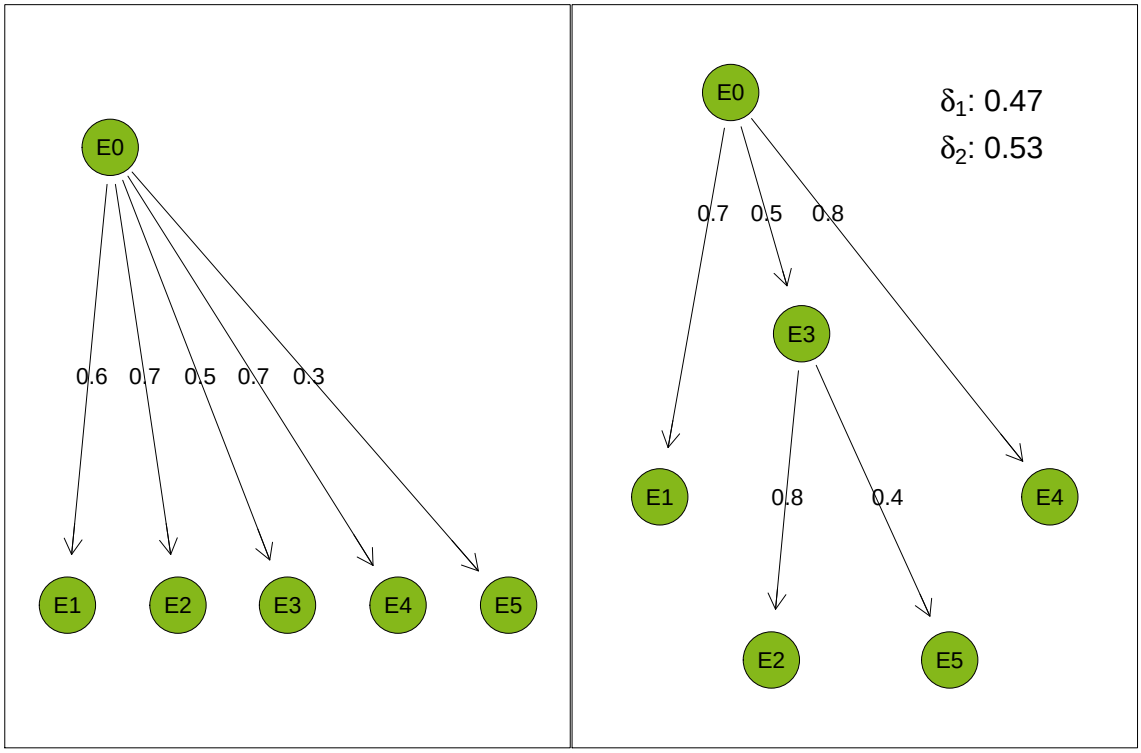



Abbildung A.3: Ausgewähltes wahres Modell aus der Klasse der Baum-Mischungs-Modelle mit 2 Komponenten und ungleichen Kantengewichten beim Stern

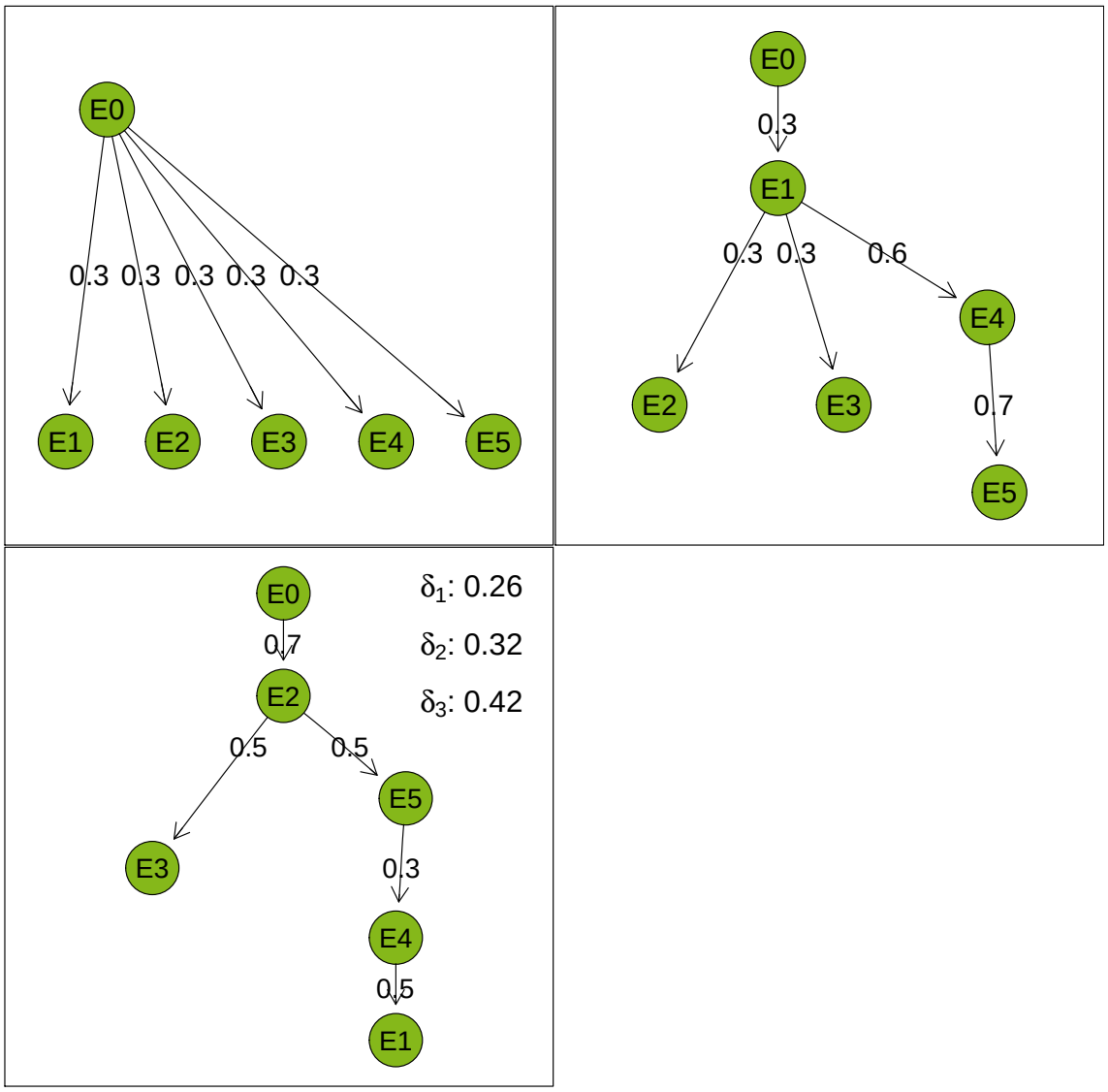



Abbildung A.4: Ausgewähltes wahres Modell aus der Klasse der Baum-Mischungs-Modelle mit 3 Komponenten und gleichen Kantengewichten beim Stern

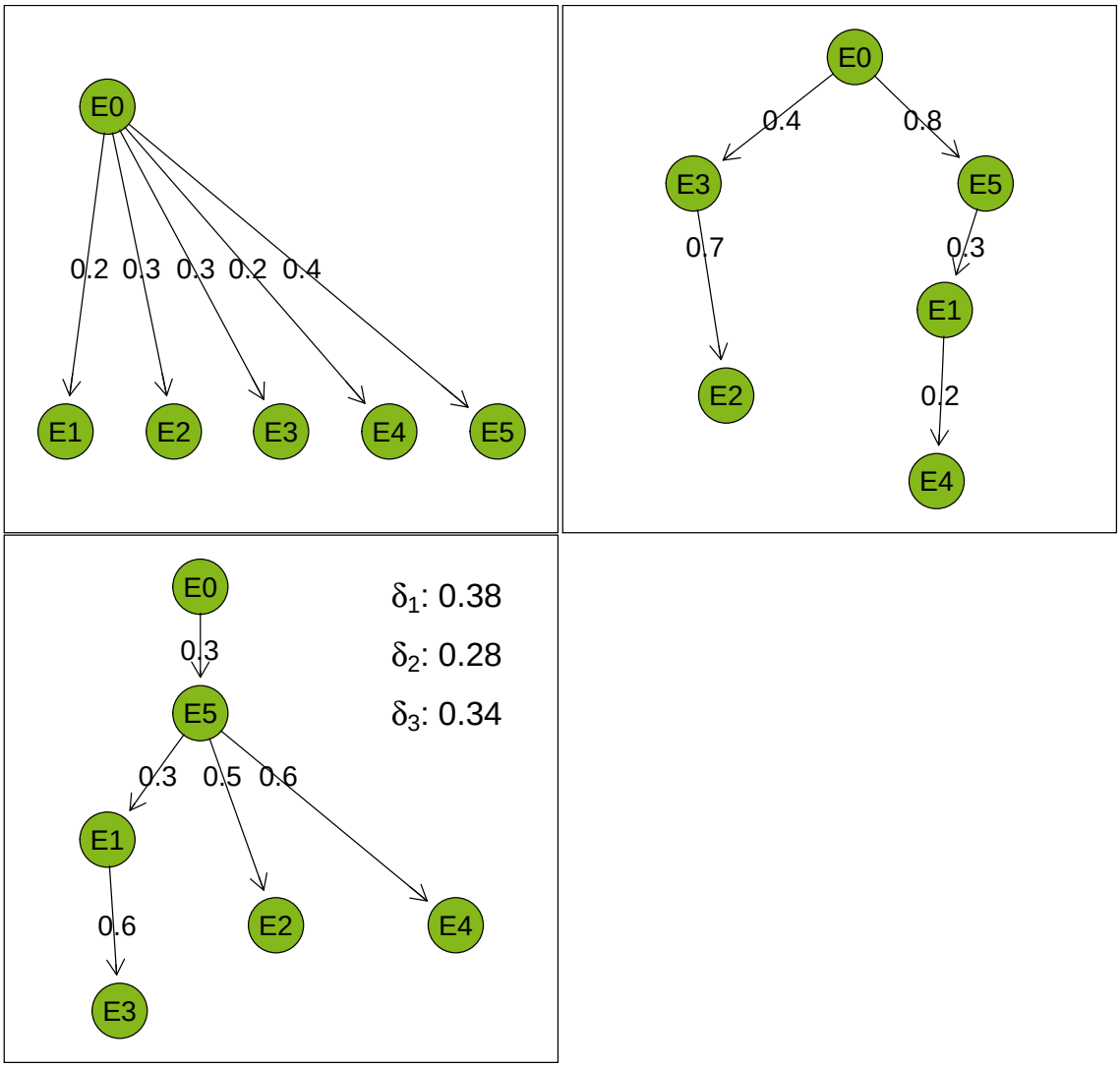



Abbildung A.5: Ausgewähltes wahres Modell aus der Klasse der Baum-Mischungs-Modelle mit 3 Komponenten und ungleichen Kantengewichten beim Stern

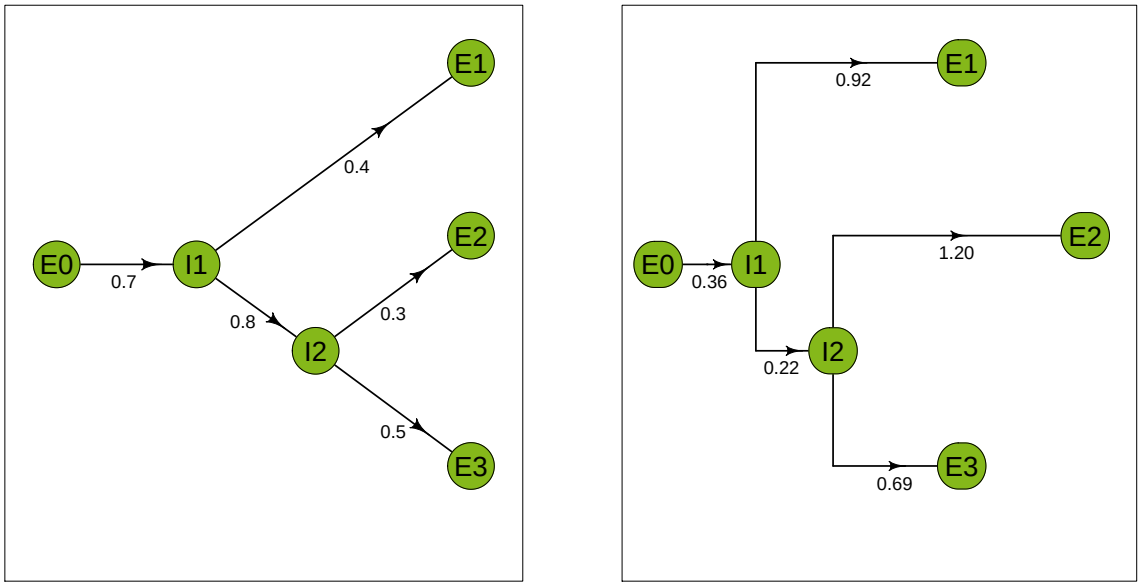



Abbildung A.6: Ausgewähltes wahres Modell aus der Klasse der Distanzbäume (links: mit Wahrscheinlichkeiten, rechts: mit Distanzen)

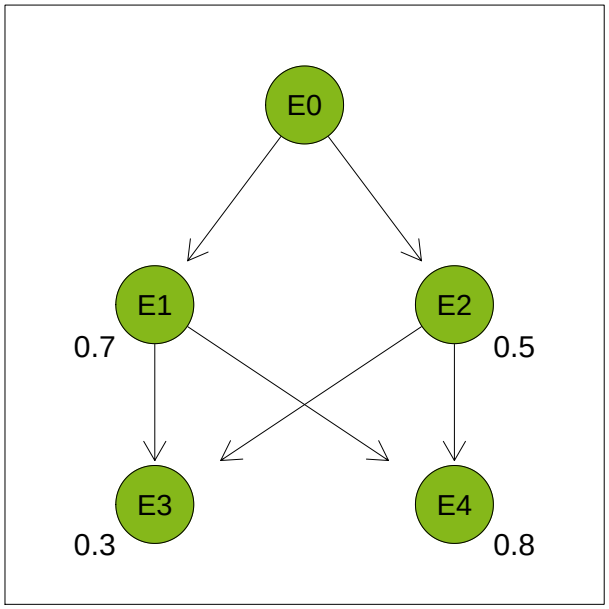



Abbildung A.7: Ausgewähltes wahres Modell aus der Klasse der verbindenden Bayes Netze

### A.1.2 Ergänzendes Beispiel zum Ranking der Modellklassen

Tabelle A.1: Ergebnisse des Wilcoxon-Mann-Whitney Tests (T: $p$-Wert $< 0.05$, F: $p$-Wert $\geq 0.05$) und die resultierende Reihenfolge vom besten (1) zum schlechtesten (8) Modell beim Anpassen eines OTM2ne, unter Verwendung der $L_1$-Distanz bei $N = 200$ Beobachtungen.

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | # T's | ranking |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 OT | - | F | F | F | F | F | T | F | 1 | 7 |
| 2 OTM2e | T | - | F | F | F | F | T | F | 2 | 5 |
| 3 OTM2ne | T | T | - | T | T | T | T | T | 7 | 1 |
| 4 OTM3e | T | T | F | - | F | T | T | T | 5 | 2.5 |
| 5 OTM3ne | T | T | F | F | - | T | T | T | 5 | 2.5 |
| 6 DIST | T | F | F | F | F | - | T | T | 3 | 4 |
| 7 CBN | F | F | F | F | F | F | - | F | 0 | 8 |
| 8 edf | T | F | F | F | F | F | T | - | 2 | 6 |
| # F's | 1 | 4 | 7 | 6 | 6 | 4 | 0 | 3 | | |