Entwicklung statistischer Tests mit computergestützter Algebra

Dissertation

zur Erlangung des Grades
eines Doktors der Naturwissenschaften
der Technischen Universität Dortmund

technische universität
dortmund

Der Fakultät Statistik
der Technischen Universität Dortmund

vorgelegt von
Anne Krampe

Dortmund 2008

1. Gutachter Prof. Dr. Ursula Gather

2. Gutachter PD. Dr. Sonja Kuhnt

Tag der mündlichen Prüfung: 20.10.2008

Inhaltsverzeichnis

Kapitel 1

Einleitung

Die Computergestützte Algebra hat sich in den 1960er Jahren als Teilgebiet der Mathematik etabliert und hat seitdem immer mehr an Bedeutung gewonnen. Bruno Buchberger leistete hierzu 1965 mit seiner Dissertationsschrift einen weitreichenden Beitrag. Begründet auf eine Idee seines Doktorvaters Wolfgang Gröbner entwickelte er den nach ihm benannten Algorithmus zur Bestimmung von Gröbner-Basen. Eine englische Übersetzung der Dissertationsschrift ist 2006 zu Ehren Buchbergers erschienen, Buchberger (2006). Neben dem Fortschritt in der Computertechnologie ist vor allem die gute Anwendbarkeit vieler Methoden der computergestützten Algebra im ingenieur- und naturwissenschaftlichen Bereich ausschlaggebend für die heutige Bedeutung dieses Teilgebiets der Mathematik, siehe beispielsweise Buchberger und Winkler (1998), Heldt et al. (2006) sowie Carbonell und Siekmann (2004).

Diaconis und Sturmfels begründeten ein relativ junges Forschungsgebiet, die Verwendung der computergestützten Algebra in der Statistik, mit einer Arbeit, die 1993 als Manuskript eingereicht und schließlich 1998 als Artikel in den Annals of Statistics veröffentlicht wurde. Für die Analyse kategorialer Daten verwendeten sie Markov Chain Monte Carlo-Methoden als Brücke zwischen computergestützter Algebra und Statistik. Dazu haben sie statistische Fragestellungen als polynomiale Gleichungssysteme dargestellt, welche anschließend mit Methoden der computergestützten Algebra gelöst werden konnten. Die zweite grundlegende Arbeit für dieses Forschungsgebiet lieferten Pistone und Wynn (1996), in der sie den Nutzen von Gröbner-Basen in der statistischen Versuchsplanung aufzeigten. Pistone et al. (2000) prägten schließlich den Begriff der „algebraischen Statistik". Verschiedene Autoren haben die Ansätze von Diaconis und Sturmfels (1998) und Pistone et al. (2000) aufgegriffen und damit vielfältige Anwendungsmöglichkeiten der algebraischen Statistik geschaffen, vergleiche z.B. Viana und Richards (2002). Giglio und Wynn (2004) verwendeten Verfahren der computergestützten Algebra in der Zuverlässigkeitstheorie. Der Nutzen der algebraischen Statistik für biologische und genetische Untersuchungen wurde unter anderem von Pachter und Sturmfels (2005), Sturmfels und Sullivant (2005) sowie Allman und Rhodes (2007) herausgestellt. Ferner erwiesen sich Methoden der computergestützten Algebra als hilfreich für Maximum-Likelihood-Schätzungen, siehe beispielsweise Dinwoodie (2002), Sturmfels (2002), Abschnitt 8, und Drton (2006). Drton et al. (2007) beschrieben eine algebraische Faktoranalyse. Angestoßen durch den Artikel von Diaconis und Sturmfels (1998) erwiesen sich insbesondere graphische Modelle als wichtiges Anwendungsgebiet der algebraischen Statistik. Ein herausragendes Ergebnis ist hier die algebraische Formulierung des bekannten Hammersley-Clifford Theorems, siehe Geiger et al. (2006).

Die Beschreibung struktureller Zusammenhänge zwischen Komponenten multivariater Zufallsvektoren ist eine wichtige Aufgabe in der Statistik. Dazu wird ein Modell gesucht, so dass die gemeinsame Verteilungsfunktion des Zufallsvektors geeignet beschrieben werden kann. Die Beurteilung der Anpassungsgüte des betrachteten Modells erfolgt anhand verschiedener Kriterien; in der vorliegenden Arbeit werden dazu Anpassungstests verwendet, wobei die Testentscheidung traditionell auf asymptotischen oder exakten Ergebnissen basiert. Oftmals ist dabei eine Annäherung der Verteilung einer zur Modellüberprüfung geeigneten Teststatistik nicht gerechtfertigt, beispielsweise wenn der Stichprobenumfang nicht „ausreichend" groß ist. Andererseits ist die exakte Berechnung häufig sehr aufwändig. Das Ziel dieser Arbeit ist die Entwicklung neuer algebraischer Tests unter Verwendung des Diaconis-Sturmfels-Algorithmus (1998). Diese Tests erweisen sich als wertvolle Ergänzung zu anderen herkömmlichen Methoden. Soweit nicht anders erwähnt, werden hierzu die Programme `CoCoA 4.3` sowie `R 2.5.0` verwendet.

In Kapitel 2 werden log-lineare und graphische Modelle zur Beschreibung möglicher Abhängigkeitsstrukturen in einem Datensatz herangezogen. Liegen ausschließlich kategoriale Merkmale vor, so können mit log-linearen Modellen verschiedene Abhängigkeitsrelationen dargestellt werden. Graphische Modelle repräsentieren Unabhängigkeitsbeziehungen zwischen den Komponenten eines multivariaten Zufallsvektors durch einen Graphen. Darroch et al. (1980) bringen log-lineare und graphische Modelle miteinander in Verbindung. Die Arbeit von Wermuth (1976) bildet die Grundlage für die Verwendung graphischer Modelle bei multivariat normalverteilten Zufallsvektoren. Lauritzen und Wermuth (1989) beschreiben graphische Modelle für gemischt stetige-diskrete Daten. Wesentliche Vorteile graphischer Modelle sind ihre gute Anwendbarkeit in verschiedenen Anwendungsgebieten der Statistik und die intuitive Darstellung komplexer Abhängigkeitsstrukturen durch Kombination kleinerer Teilgraphen.

In Kapitel 3 werden die verwendeten Begriffe der computergestützten Algebra, Ideale und Varietäten, eingeführt und ihre Struktur charakterisiert. Insbesondere kann jedes Polynomideal durch eine Gröbner-Basis erzeugt werden. Darauf aufbauend wird der Algorithmus von Diaconis und Sturmfels (1998) eingeführt. Dieser ist der Eliminationstheorie der computergestützten Algebra entlehnt und ermöglicht die Simulation aus der bedingten Verteilung einer diskreten Exponentialfamilie mit beobachteter suffizienter Statistik.

Nachfolgend werden neue Verfahren aus dem Bereich der algebraischen Statistik vorgeschlagen. In Kapitel 4 werden algebraische Testverfahren entwickelt. Liegen der statistischen Untersuchung so genannte gepaarte Beobachtungen zugrunde, so sind häufig strukturelle Übereinstimmungen der Realisationen von besonderem Interesse. Der Bowker-Test prüft, ob der vorliegende Datensatz der Hypothese perfekt symmetrischer Zellwahrscheinlichkeiten widerspricht. Weitere Tests auf perfekte Symmetrie wie der stetigkeitskorrigierte Anpassungstest, vgl. Edwards (1948), oder der Test von May und Johnson (2001) werden ebenfalls betrachtet. Häufig ist jedoch diese Symmetrieforderung zu restriktiv und tatsächlich vorhandene Symmetriestrukturen werden nicht entdeckt. Daher werden zudem alternative, „gewichtete“ Symmetriemodelle wie das bedingte, das diagonale sowie das ordinale Quasi-Symmetriemodell, vgl. McCullagh (1978), Goodman (1979) und Agresti (1983), betrachtet. Die Testentscheidungen beruhen üblicherweise auf traditionellen asymptotischen oder exakten Methoden. Unter Verwendung des Algorithmus von Diaconis und Sturmfels werden für die vorgestellten Tests neue algebraische Verfahren eingeführt. Die Eigenschaften des neuen und der herkömmlichen Verfahren werden in einer Simulationsstudie vergleichend untersucht.

Bei der Analyse kategorialer Daten ist häufig nicht nur von Interesse, ob strukturelle Abhängigkeiten zwischen den erhobenen Merkmalen bestehen, sondern es interessiert, wie stark diese Zusammenhänge sind. Ein wichtiges Beispiel hierzu findet man in der Epidemiologie, in der die Identifizierung von Risikofaktoren für Krankheiten von besonderer Bedeutung ist. Für den Fall einer 2×2-Kontingenztafel wird üblicherweise das Odds Ratio (OR) und das zugehörige Konfidenzintervall als Maß für den Zusammenhang zwischen einer Krankheit und der Exposition mit einem potenziellen Risikofaktor verwendet. Die Konfidenzintervalle werden häufig entweder approximativ oder exakt berechnet, siehe Kreienbrock und Schach (1995). In der vorliegenden Arbeit wird ein alternatives algebraisches Konfidenzintervall für das Odds Ratio entwickelt. Das Konzept des neuen Konfidenzintervalls für das OR für eine 2×2-Tafel kann auf den allgemeineren Fall eines $2 \times 2 \times K$-Datensatzes übertragen werden. In einer Simulationsstudie werden anschließend die Eigenschaften des vorgeschlagenen algebraischen sowie des asymptotischen und exakten Konfidenzintervalls untersucht.

Am Beispiel der Modellselektion für graphische Modelle sowie für die in Kapitel 4 beschriebenen Symmetriemodelle wird in Kapitel 5 gezeigt, wie der nötige Simulationsaufwand für algebraische Tests reduziert werden kann. Für eine Analyse der Abhängigkeitsstrukturen in einem Datensatz wird oftmals aus einer Menge möglicher Modelle dasjenige Modell ausgewählt, das gemäß einem gewählten Kriterium am geeignetsten erscheint. In dieser Arbeit werden Modellselektionsstrategien verwendet, die auf Anpassungstests beruhen, wobei die Testentscheidungen traditionell auf der Asymptotik der kritischen Werte oder exakten Berechnungen beruhen. Liegt der Untersuchung ein kategorialer Datensatz zugrunde, so ermöglicht der Diaconis-Sturmfels-Algorithmus alternative algebraische Anpassungstests. Der Simulationsaufwand für eine algebraische Modellselektion nach Diaconis und Sturmfels ist jedoch erheblich und in der Praxis oft nicht realisierbar, da für jedes Modell eine separate Simulation erforderlich ist. Weisen die interessierenden Modelle eine hierarchische Struktur auf, so kann der nötige Simulationsaufwand z. T. erheblich reduziert werden. Insbesondere wird bewiesen, dass das Ideal eines Modells im Ideal eines Untermodells enthalten ist; diese Hierarchie wirkt sich entsprechend auf die Gröbner-Basen aus. Für die graphischen Modelle wird gezeigt, dass mit der Gröbner-Basis des Unabhängigkeitsmodells jede weitere Gröbner-Basis der übrigen Modelle dargestellt werden kann. Entsprechend ist eine Simulation gemäß dem Unabhängigkeitsmodell ausreichend, um algebraische Tests für alle darauf aufbauenden Modelle durchzuführen. Für eine neue algebraische Modellselektion ersetzen die so generierten algebraischen p-Werte die entsprechenden asymptotischen bzw. exakten p-Werte in bekannten Modellselektionsverfahren. Die in Kapitel 4 vorgestellten Symmetriemodelle weisen ebenfalls eine hierarchische Struktur auf; so reicht hier eine Simulation gemäß dem perfekten Symmetriemodell, um einen algebraischen Test für das bedingte, diagonale und ordinale Quasi-Symmetriemodell durchzuführen. In einer Simulationsstudie werden asymptotische Modellwahlverfahren, die „herkömmliche" sowie die neue algebraische Modellselektionsprozedur miteinander verglichen und ihre Eigenschaften analysiert.

Weitere Herausforderungen und offene Forschungsfragen der algebraischen Statistik werden in Kapitel 6 diskutiert. Der Fokus liegt hier auf der Verwendung des Diaconis-Sturmfels-Algorithmus bei gemischt stetigen-diskreten Datensätzen. Hierzu werden Lösungsansätze und offene Probleme beschrieben. Abschließend werden in Kapitel 7 die Ergebnisse der vorliegenden Arbeit zusammenfassend dargestellt und diskutiert.

Kapitel 2

Strukturmodellierung

In diesem Kapitel werden Methoden für die Beschreibung struktureller Zusammenhänge zwischen Komponenten eines multivariaten Zufallsvektors vorgestellt. Für eine geeignete Darstellung der gemeinsamen Verteilungsfunktion des betrachteten Zufallsvektors existieren in der Statistik verschiedene Ansätze. Eine so genannte Copula-Funktion ermöglicht beispielsweise die Zerlegung der gemeinsamen multivariaten Verteilungsfunktion in die einzelnen univariaten Randverteilungen sowie die zugehörige Abhängigkeitsstruktur, siehe z. B. Nelsen (2006). In dieser Arbeit werden log-lineare sowie graphische Modelle betrachtet. Liegen der Untersuchung ausschließlich kategoriale Merkmale zugrunde, so ermöglichen log-lineare Modelle die Beschreibung verschiedener Abhängigkeitsstrukturen, wie z. B. Symmetrie- oder Unabhängigkeitsbeziehungen. Die Grundlage graphischer Modelle bildet die Graphentheorie, deren Konzepte in verschiedenen Wissenschaftsgebieten wie z. B. in der Physik und der Genetik angewendet werden, siehe z. B. Volkmann (1996), Bodendiek und Lang (1995). Für graphische Modelle ist kennzeichnend, dass Unabhängigkeitsbeziehungen zwischen einzelnen Komponenten eines multivariaten Zufallsvektors durch einen Graphen repräsentiert werden und dadurch ein Teil der gemeinsamen Verteilungsfunktion spezifiziert wird. Darroch et al. (1980) stellen eine erste Beziehung zwischen graphischen Modellen und log-linearen Modellen für die Analyse kategorialer Daten her. Die Klasse der log-linearen Modelle umfasst unter anderem ungerichtete, einfache graphische Modelle mit multinomialverteiltem Zufallsvektor. Bereits 1976 beschreibt Wermuth eine Analogie zwischen log-linearen Modellen und Kovarianz-Selektionsmodellen für multivariat normalverteilte Zufallsvariablen. Die damit verbundene Anwendungsmöglichkeit graphischer Modelle für stetige Merkmale wird von Lauritzen und Wermuth (1989) für gemischt stetige-diskrete Variablen erweitert. Die Vorteile dieses Ansatzes sind vielfältig, so umfasst die Klasse der graphischen Modelle zum einen verschiedene bekannte multivariate Analysemodelle als Spezialfälle; zum anderen werden graphische Modelle z. B. auch erfolgreich in der Zeitreihenanalyse eingesetzt, vgl. beispielsweise Dahlhaus (2000). Ein weiterer wichtiger Vorteil dieser Modellklasse ist die intuitive und einfache Darstellung komplexer Abhängigkeitsstrukturen durch Kombination kleinerer Teilgraphen. Nachfolgend werden zunächst log-lineare Modelle für die Analyse kategorialer Datensätze beschrieben, anschließend werden graphische Modelle eingeführt.

2.1 Log-lineare Modelle

In verschiedenen Anwendungsgebieten der Statistik wie z. B. der Medizin oder den Sozialwissenschaften werden häufig kategoriale Daten erhoben, d. h. das Skalenniveau ist nominal oder ordinal mit endlich vielen möglichen Realisationen. Die Modellklasse der log-linearen Modelle ist dadurch gekennzeichnet, dass der Logarithmus der erwarteten Zellhäufigkeiten, oder äquivalent der Zellwahrscheinlichkeiten, als lineare Funktion unbekannter Parameter repräsentiert wird. Auf diese Weise können verschiedene Abhängigkeitsstrukturen in einem kategorialen Datensatz beschrieben werden. Oftmals werden hierarchische log-lineare Modelle betrachtet. Es sei nachfolgend stets angenommen, dass die Haupteffekte aller untersuchten Variablen im Modell vorhanden sind. Die getroffenen Aussagen stützen sich im Wesentlichen auf Agresti (2002), Bishop et al. (1995), Christensen (1997) sowie Lauritzen (1998).

Im Weiteren sei $\Delta = \{1, 2, \ldots, q\}$ eine endliche Indexmenge und $X_\Delta = (X_\delta)_{\delta\in\Delta}$ ein q-dimensionaler, kategorialer Zufallsvektor. Die Realisation eines beliebigen Merkmals X_δ, $\delta \in \Delta$, sei mit i_δ bezeichnet und nehme einen Wert in der endlichen Menge $\mathcal{I}_\delta = \{1, \ldots, I_\delta\}$ an. Die gemeinsame Häufigkeitsverteilung kategorialer Zufallsvariablen wird üblicherweise in einer Kontingenztafel tabellarisch dargestellt. Die Zellen $i = (i_\delta)_{\delta\in\Delta}$ einer solchen Tafel stammen aus dem Kartesischen Produkt der Mengen der möglichen Realisationen der einzelnen Variablen: $\mathcal{I} = \times_{\delta\in\Delta}\mathcal{I}_\delta$. Die Anzahl der Variablen ist die Dimension der Tafel, d. h. es liegt eine q-dimensionale bzw. $I_1 \times I_2 \times \ldots \times I_q$-Kontingenztafel vor. Ein Zelleintrag der Tafel sei n_i, die beobachtete Anzahl der untersuchten Objekte mit Merkmalskombination $i \in \mathcal{I}$, d. h. $n_i = n_{(i_1,\ldots,i_q)'}$. Für den Stichprobenumfang n gilt: $n = \sum_{i \in \mathcal{I}} n_i$. Alternativ kann auch die Wahrscheinlichkeit $\pi_i = \pi_{(i_1,\ldots,i_q)'}$ mit der für X_Δ die Ausprägung i beobachtet wird ein Eintrag der Kontingenztafel sein, dabei gilt $\sum_{i \in \mathcal{I}} \pi_i = 1$. Häufig werden so genannte Randtafeln betrachtet. Die beobachteten Objekte werden hier gemäß einer Teilmenge der untersuchten Merkmale $a \subseteq \Delta$ in einer $|a|-$dimensionalen Tafel angeordnet. Damit stammen die Zellen einer Randtafel $i_a = (i_\delta)_{\delta \in a}$ aus der Menge $\mathcal{I}_a = \times_{\delta \in a} \mathcal{I}_\delta$. Durch Summation über die Ausprägungen der übrigen Merkmale $\Delta \setminus \{a\}$ können die Zelleinträge n_{i_a} der Randtafel berechnet werden.

Im Folgenden sei angenommen, dass die kategoriale Stichprobe aus einer Multinomialverteilung stammt. Die Ergebnisse dieser Arbeit lassen sich allerdings auch auf die Poisson- oder die Produktmultinomial-Erhebungstechnik übertragen. Gemäß Edwards (2000), Kapitel 2.2, wird ein log-lineares Modell wie folgt definiert:

Definition 2.1.1 ((Hierarchisches) log-lineares Modell)
Es sei $X_\Delta = (X_1, \ldots, X_q)'$ ein Zufallsvektor und $i = (i_1, \ldots, i_q)'$, $i \in \mathcal{I} = \times_{\delta \in \Delta} \mathcal{I}_\delta$ sei eine Realisation von X_Δ. Die erwarteten absoluten Zellhäufigkeiten werden als m_i, $i \in \mathcal{I}$, beschrieben. Weiter seien $\lambda^a_{i_a} \in \mathbb{R}$, $a \subseteq \Delta$, unbekannte Parameter, Effekte genannt.
Das saturierte oder auch vollständige log-lineare Modell kann dargestellt werden als $\log(m_i) = \sum_{\{a \subseteq \Delta\}} \lambda^a_{i_a}$. Ein hierarchisches log-lineares Modell wird durch eine Erzeugendenmenge $d = \{d_1, \ldots, d_r\}$ mit $d_j \subseteq \Delta,\ j = 1, \ldots, r$, und $\log(m_i) = \sum_{\{a \subseteq \Delta | \exists j:\, a \subseteq d_j\}} \lambda^a_{i_a}$ definiert.

Das saturierte log-lineare Modell enthält alle weiteren möglichen hierarchischen log-linearen Modelle als Spezialfälle, indem nicht wirksame Effekte sowie alle darauf aufbauenden Effekte wie höhergradige Interaktionen gleich Null gesetzt werden. Um die Eindeutigkeit und damit die Schätzbarkeit der Parameter zu gewährleisten, müssen zusätzliche Restriktionen an diese gestellt werden, z. B. durch Festlegung einer Referenzzelle i^* mit $\lambda^a_{i_a} = 0$, falls $i_\delta = i^*_\delta$, $\delta \in a \subseteq \Delta$. Weitere mögliche Restriktionen finden sich in McCullagh und Nelder (1999), Kapitel 3.5.

2.2 Graphische Modelle

Graphische Modelle beschreiben die Beziehungen zwischen den einzelnen Komponenten eines multivariaten Zufallsvektors durch einen Graphen, siehe z. B. Edwards (2000), Lauritzen (1998) und Whittaker (1990).

Ein Graph $\mathcal{G} = (V, E)$ ist ein Paar, das aus einer endlichen *Knotenmenge* V (set of vertices) und einer *Kantenmenge* E (set of edges) besteht. Üblicherweise wird die Knotenmenge V mit einer Teilmenge der natürlichen Zahlen $\{1, 2, \dots, k\}$ identifiziert. Die einzelnen Knoten repräsentieren die untersuchten Variablen, daher wird V auch abkürzend synonym für die Menge der zugrunde liegenden Zufallsvariablen verwendet. Häufig wird die Menge der diskreten Variablen durch Δ und die der stetigen durch Γ bezeichnet. Liegen der Untersuchung gemischt stetige-diskrete Merkmale zugrunde, so ist also $V = \Delta \cup \Gamma$; der zugehörige Graph wird dann als *markierter Graph* (marked graph) bezeichnet. Die diskreten Variablen werden durch einen Punkt (dot für discrete) bzw. die stetigen durch einen Kreis (circle für continuous) im Graphen symbolisiert. Die Beziehungen zwischen den Variablen werden durch die Kantenmenge E, d. h. durch Verbindungen bzw. durch das Fehlen von Verbindungen zwischen den jeweiligen Variablen im Graphen spezifiziert. Allgemein ist E eine Menge von geordneten Paaren unterschiedlicher Knoten und damit eine Teilmenge von $V \times V$. Es existiert eine *gerichtete Kante* (directed edge) im Graphen $\mathcal{G}$ von v nach w, $v, w \in V$, falls das geordnete Paar (v, w) in der Kantenmenge E enthalten ist. Ist zusätzlich $(w, v) \in E$, so ist die *Kante ungerichtet* (undirected edge). Ein Graph heißt ungerichtet, falls alle Kanten ungerichtet sind (undirected graph). Wenn bekannt ist, dass ein ungerichteter Graph vorliegt, wird üblicherweise eine ungerichtete Kante $(v, w), (w, v) \in E$ abkürzend durch $(v, w) \in E$ dargestellt. Soweit nicht anders erwähnt, werden in dieser Arbeit ausschließlich ungerichtete, einfache (simple) Graphen betrachtet, d. h. es werden weder multiple Kanten (multiple edges) noch Schleifen (loops) zugelassen. Zwei Knoten v und w heißen *benachbart* (neighbors), wenn sie durch eine ungerichtete Kante miteinander verbunden sind, d. h. $(v, w) \in E$. Die Menge aller Nachbarn (boundary) eines Knoten v wird mit $bd(v)$ bezeichnet und es gilt: $bd(v) = \{w : (v, w) \in E, v, w \in V, v \neq w\}$. Es bezeichne $A \subseteq V$ eine Teilmenge von Knoten. Dann ist die Menge aller Nachbarn von A gegeben durch

$bd(A) = \{w : w \in bd(v), v \in A \subseteq V, w \in V \setminus \{A\}\}$. Der Abschluss (closure) von A wird mit $cl(A)$ abgekürzt und besteht aus der Menge aller Nachbarn von A und A selbst, es ist also $cl(A) = bd(A) \cup A$. Der durch A *induzierte Untergraph* $\mathcal{G}_A$ enthält ausschließlich Knoten $v \in A$ und Kanten von $\mathcal{G}$, die nur Knoten aus A verbinden. Ein Graph oder ein Untergraph heißt *vollständig* (complete), wenn alle Knoten benachbart sind. Eine Teilmenge $A \subseteq V$ heißt vollständig, wenn sie einen vollständigen Graphen bzw. Untergraphen induziert. Eine vollständige Teilmenge heißt *Clique* (clique), wenn sie einen maximalen vollständigen Untergraphen induziert, d. h. durch Hinzufügen jedes weiteren Knotens wäre der so induzierte Untergraph nicht mehr vollständig. Ein *Pfad* (path) der Länge l von Knoten v zu Knoten w ist eine Folge $v = v_0, \ldots, v_l = w$ von verschiedenen Knoten, für die gilt: $(v_{e-1}, v_e) \in E$ für alle $e = 1, \ldots, l$. Existiert ein Pfad zwischen v und w, so sind diese Knoten verbunden (connected). Eine Teilmenge von Knoten $S \subseteq V$ *separiert* (separate) die Knoten v und w, $v, w \in V$, wenn alle Pfade von v nach w durch Knoten s aus S führen. Weiter seien A, B, S paarweise disjunkte Teilmengen von V. Dann wird A von B durch S separiert, wenn jeder Pfad für alle $v \in A$ nach $w \in B$ durch mindestens einen Knoten $s \in S$ führt. Eine nützliche Eigenschaft von Graphen ist ihre *Zerlegbarkeit* in Teilgraphen. Dazu seien A, B, C paarweise disjunkte Teilmengen der Knotenmenge $V = A \cup B \cup C$ eines ungerichteten, markierten Graphen $\mathcal{G}$. A, B und C zerlegen (strong decomposition) den Graphen $\mathcal{G}$ in die Komponenten $\mathcal{G}_{A\cup C}$ und $\mathcal{G}_{B\cup C}$, wenn die folgenden Bedingungen erfüllt sind: *i)* C separiert A von B; *ii)* C ist eine vollständige Teilmenge von V und *iii)* $C \subseteq \Delta$ oder $B \subseteq \Gamma$. Für schwache Zerlegbarkeit muss die Bedingung *iii)* nicht notwendigerweise gelten. Diese zerlegbaren Modelle können sukzessiv in kleinere, den einzelnen Cliquen entsprechende Modelle zerlegt werden. Zudem erleichtern verschiedene theoretische Ergebnisse die Handhabung solcher graphischen Modelle, vgl. Frydenberg und Lauritzen (1989), und es existiert die explizite Darstellung der Maximum-Likelihood-Schätzer.

Im Folgenden bezeichne $X = (X_\alpha)_{\alpha\in V} = (X_1, \ldots, X_k)'$ einen Zufallsvektor und $V = \{1, \ldots, k\}$ die entsprechende Indexmenge für die Knoten, die die untersuchten Variablen repräsentieren. Wie üblich beschreibe $\{\mathscr{H}, \mathfrak{A}, \mathcal{P}\}$ das zugrunde liegende statistische Modell und $\{\mathscr{H}, \mathfrak{A}, P\}$ entsprechend den Wahrscheinlichkeitsraum, wobei die nicht-leere Menge $\mathscr{H}$ den Stichprobenraum symbolisiert; $\mathfrak{A}$ ist eine σ-Algebra über $\mathscr{H}$ und als solche der Definitionsbereich für jedes $P \in \mathcal{P}$, der Familie von Wahrscheinlichkeitsmaßen P auf $\mathfrak{A}$. Die Kanten aus E bzw. vielmehr das Fehlen von Kanten in dem Graphen spezifizieren die Abhängigkeitsstrukturen zwischen den betrachteten Merkmalen X_α, $\alpha = 1, \ldots, k$, und charakterisieren somit die gemeinsame Verteilung P von X. Je nachdem, ob die fehlenden Kanten bedingte oder marginale Unabhängigkeiten beschreiben, heißen die resultierenden Graphen (bedingter) Unabhängigkeitsgraph oder Konzentrationsgraph. Die vorhandenen Kanten werden durch eine durchgezogene bzw. gestrichelte Linie symbolisiert. In der vorliegenden Arbeit werden ausschließlich Unabhängigkeitsgraphen betrachtet.

Definition 2.2.1 (Unabhängigkeitsgraph, graphisches Modell)
Es sei X ein k-dimensionaler Zufallsvektor. Ein bedingter Unabhängigkeitsgraph (oder kurz: Unabhängigkeitsgraph) von X ist der ungerichtete Graph $\mathcal{G} = (V, E)$ mit Knotenmenge $V = \{1, \ldots, k\}$ und Kantenmenge E, für die gilt: Eine Kante (v, w), $v, w \in V$, ist genau dann nicht in E enthalten, wenn X_v bedingt unabhängig ist von X_w, gegeben $X_{V\setminus\{v,w\}}$; in Zeichen: $X_v \perp\!\!\!\perp X_w \mid X_{V\setminus\{v,w\}}$. Ein zugehöriges graphisches Unabhängigkeitsmodell $M(\mathcal{G})$ (oder kurz: graphisches Modell) ist eine Familie von Wahrscheinlichkeitsmaßen $\mathcal{P} = \{P_\theta,\, \theta \in \Theta\}$ aller $P \in \mathcal{P}$ für die gilt:

$$X_v \perp\!\!\!\perp X_w \mid X_{V\setminus\{v,w\}}, \quad \forall\, (v, w), \notin E.$$

Definition 2.2.1 basiert auf der paarweisen Markov-Eigenschaft. Wichtig werden im Folgenden aber auch lokale und globale Markov-Eigenschaften sein.

Definition 2.2.2 (Markov-Eigenschaft)
Es sei $X = (X_\alpha)_{\alpha\in V}$ ein Zufallsvektor und $\mathcal{G} = (V, E)$ ein ungerichteter Graph. Ein Wahrscheinlichkeitsmaß P über dem Graphen $\mathcal{G} = (V, E)$ erfüllt

- *die paarweise Markov-Eigenschaft, wenn für jedes Paar (v, w) nicht benachbarter Knoten gilt: $v \perp\!\!\!\perp w \mid V \setminus \{v, w\}$;*

- *die lokale Markov-Eigenschaft, wenn für jeden Knoten $v \in V$ gilt: $v \perp\!\!\!\perp V \setminus cl(v) \mid bd(v)$;*

- *die globale Markov-Eigenschaft, wenn für jedes Tripel* (A, B, S) *paarweise disjunkter Teilmengen von* V *gilt:*
  S *separiert* A *von* B *in* $\mathcal{G}$*, dann folgt* $A \perp\!\!\!\perp B \,|\, S$.

Eine Familie von Wahrscheinlichkeitsmaßen $\mathcal{P}$ *weist die paarweise, lokale oder globale Markov-Eigenschaft auf, wenn jedes* $P \in \mathcal{P}$ *die jeweiligen Bedingungen erfüllt.*

In dieser Arbeit werden Wahrscheinlichkeitsmaße betrachtet, für die die Äquivalenz dieser drei Eigenschaften gilt. Ausgehend vom Wahrscheinlichkeitsmaß P kann der zugehörige Unabhängigkeitsgraph konstruiert werden. Umgekehrt können vom Graphen $\mathcal{G}$ alle bedingten Unabhängigkeiten abgelesen und somit auch die Verteilung des betrachteten multivariaten Zufallsvektors X charakterisiert werden. Ähnlich dem Multiplikationssatz (vgl. Mood, Graybill und Boes (1974), Theorem 13) kann ein zum Graphen $\mathcal{G}$ gehörendes Wahrscheinlichkeitsmaß P wie folgt faktorisiert werden:

Korollar 2.2.3 (Faktorisierungssatz)
Es seien $\{\mathscr{H}, \mathfrak{A}, P\}$ *ein Wahrscheinlichkeitsraum,* $\mathcal{G} = (V, E)$ *ein zugehöriger Graph und* $C \subseteq V$ *eine vollständige Teilmenge von* V*. O. B. d. A. sei angenommen, dass* C *eine Clique von* $\mathcal{G}$ *ist. Die Menge aller Cliquen von* $\mathcal{G}$ *sei mit* $\mathscr{C}$ *bezeichnet. Dann faktorisiert das Wahrscheinlichkeitsmaß* P *gemäß* $\mathcal{G}$*, wenn für alle* $C \in \mathscr{C}$ *eine nichtnegative Funktion* ψ_C *sowie ein Produktmaß* $\mu = \times_{v \in V} \mu_v$ *auf* $\mathfrak{A}$ *existiert, so dass die Wahrscheinlichkeitsfunktion von* P *bezüglich* μ *gegeben ist durch*

$$f(x) = \prod_{C \in \mathscr{C}} \psi_C(x_C).$$

Die Beziehung zwischen der Faktorisierung von P und den Markov-Eigenschaften wird für einen ungerichteten Graphen $\mathcal{G}$ durch die folgende Hierarchie beschrieben: Kann das Wahrscheinlichkeitsmaß P gemäß $\mathcal{G}$ faktorisiert werden, so folgt daraus, dass P die globale, die lokale und schließlich die paarweise Markov-Eigenschaft erfüllt, Lauritzen (1998), S. 35.

Hammersley und Clifford haben gezeigt, dass unter bestimmten Voraussetzungen für das Wahrscheinlichkeitsmaß P die Äquivalenz der Faktorisierungseigenschaft (Korollar 2.2.3) und der paarweisen Markov-Eigenschaft (Definition 2.2.2) gilt:

Theorem 2.2.4 (Hammersley und Clifford)
Es sei $\mathcal{G} = (V, E)$ ein ungerichteter Graph und $\{\mathscr{H}, \mathfrak{A}, P\}$ der zugehörige Wahrscheinlichkeitsraum. Das Wahrscheinlichkeitsmaß P besitze eine positive Wahrscheinlichkeitsfunktion f bezüglich eines Produktmaßes μ. Dann erfüllt P die paarweise Markov-Eigenschaft bezüglich $\mathcal{G}$ genau dann, wenn P gemäß $\mathcal{G}$ faktorisiert werden kann.

Beweis: Besag (1974), S. 196 ff. für den diskreten Fall;
Lauritzen (1998), S. 36 f für den stetigen Fall.

Die Multinomialverteilung als gemeinsame Verteilung der Zellhäufigkeiten erfüllt zum Beispiel Theorem 2.2.4. Wie bereits erwähnt ist ein graphisches Modell mit multinomialverteiltem Zufallsvektior immer auch ein hierarchisches log-lineares Modell. Die Umkehrung gilt, wenn die Erzeuger $d_1, \ldots, d_r$ des hierarchischen log-linearen Modells den Cliquen des Unabhängigkeitsgraphen entsprechen, vgl. Whittaker (1990), Proposition 7.3.1.

Bisher wurden Modellierungsmöglichkeiten für Datensätze mit ausschließlich diskreten Variablen vorgestellt. Liegen der Untersuchung nur stetige Merkmale zugrunde, so wird den Variablen oftmals eine multivariate Normalverteilung unterstellt. Die Arbeit von Wermuth (1976) begründet für diesen Fall die Verwendung graphischer Modelle für eine Analyse der Abhängigkeitsstrukturen. Die so genannten graphischen Gauß-Modelle sind ausführlich z. B. in den Büchern von Lauritzen (1998) und Whittaker (1990) beschrieben und werden in dieser Arbeit nicht weiter erläutert. Vielmehr interessieren nun graphische Modelle für gemischt stetige-diskrete Variablen. Dazu schlagen Lauritzen und Wermuth (1989) die Familie der bedingten Gaußverteilung (conditional Gaussian, CG-Verteilung) als Verteilungsannahme für den zugrunde liegenden Zufallsvektor vor. Diese Verteilungsfamilie wird im Folgenden beschrieben und anschließend in einen graphentheoretischen Kontext gesetzt. Für eine ausführliche Darstellung sei auf Lauritzen und Wermuth (1989) sowie Lauritzen (1998) verwiesen.

Es sei angenommen, dass insgesamt q diskrete und r stetige Merkmale betrachtet werden. Die Indexmenge der diskreten Variablen wird mit $\Delta = \{1, \ldots, q\}$, die der stetigen mit $\Gamma = \{1, \ldots, r\}$ bezeichnet. Wie im rein diskreten Fall werden diese Mengen synonym für die jeweilige Menge der Variablen und im graphentheoretischen Kontext für die Knotenmenge $V = \Delta \cup \Gamma$ verwendet. Ein Randvektor ist gegeben durch $(X_a', X_b')'$ mit $a \subseteq \Delta$ und $b \subseteq \Gamma$. Als gemeinsame Verteilung des Zufallsvektors $X = (X_\alpha)_{\alpha \in V} = (X_\Delta', X_\Gamma')'$ wird die bedingte Gaußverteilung eingeführt. Alternative Modellverteilungen wie z. B. die Koehler-Symanowski Verteilung, vgl. Caputo (1998), werden in dieser Arbeit nicht berücksichtigt.

Definition 2.2.5 (bedingte Gaußverteilung, CG-Verteilung)

Es sei $X = (X_\Delta', X_\Gamma')'$ ein Zufallsvektor bestehend aus q diskreten und r stetigen Variablen. Eine Beobachtung von X sei mit $(i', y')'$ bezeichnet, wobei $y \in I\!R^{|\Gamma|}$, $i \in \mathcal{I} = \times_{\delta \in \Delta} \mathcal{I}_\delta$, und $\mathcal{I}_\delta$ ist die Menge aller möglichen Ausprägungen für X_δ, $\delta \in \Delta$. Die Wahrscheinlichkeit, dass für X_Δ die Realisation i beobachtet wird, sei $P(X_\Delta = i) =: \pi_i$, $i \in \mathcal{I}$, und es gilt $\sum_{i \in \mathcal{I}} \pi_i = 1$. Ferner sei $X_\Gamma | X_\Delta$ multivariat normalverteilt mit Erwartungswert $E(X_\Gamma | X_\Delta = i) = \mu_i$ und Varianz $Var(X_\Gamma | X_\Delta = i) = \Sigma_i$.
Dann besitzt X_V eine strikt positive Dichte $f_{(X_\Delta, X_\Gamma)}$ bezüglich des Produktes des Zählmaßes auf $\mathcal{I}$ mit dem Lebesgue-Maß auf $I\!R^{|\Gamma|}$, genannt bedingte Gaußverteilung, die sich wie folgt darstellen lässt:

$$f_{(X_\Delta, X_\Gamma)}(i, y) = \pi_i \cdot |2\pi\Sigma_i|^{-\frac{1}{2}} \cdot \exp\{-\frac{1}{2}(y - \mu_i)'\Sigma_i^{-1}(y - \mu_i)\}.$$

Die Parameter $\{\pi_i, \mu_i, \Sigma_i\}_{i \in \mathcal{I}}$ sind die so genannten Momentenparameter der (heterogenen) CG-Verteilung. Gilt zusätzlich für alle $i \in \mathcal{I}$, dass Σ_i konstant ist, d. h. $\Sigma_i = \Sigma$, so heißt die bedingte Gaußverteilung homogen.

Neben der oben beschriebenen Darstellung der CG-Verteilung existieren verschiedene weitere Parametrisierungen. Eine sehr wichtige ist die kanonische, mit der die Dichte dargestellt werden kann als

$$f_{(X_\Delta, X_\Gamma)}(i, y) = \exp\left\{\alpha_i + \beta_i' y - \frac{1}{2} y'\Omega_i y\right\};$$

die Parameter $\{\alpha_i, \beta_i, \Omega_i\}_{i \in \mathcal{I}}$ heißen entsprechend kanonische Parameter der CG-Verteilung. Die Transformation von der Momenten-Parametrisierung in die kanonische erfolgt durch $\Omega_i = \Sigma_i^{-1}$, $\beta_i = \Sigma_i \mu_i$, , $\alpha_i = \log(\pi_i) - \frac{1}{2}\log(|\Sigma_i|) - \frac{1}{2}\mu_i'\Sigma_i^{-1}\mu_i - \frac{r}{2}\log(2\pi)$. Die Reparametrisierung von der kanonischen in die Momenten-Darstellung ist analog und wird daher nicht vorgeführt.

Für eine Analyse der Abhängigkeiten zwischen Variablen aus $X = (X'_\Delta, X'_\Gamma)'$ wird im Weiteren die kanonische Darstellung der CG-Verteilung verwendet. Die entsprechenden Parameter $\{\alpha_i, \beta_i, \Omega_i\}_{i\in\mathcal{I}}$ können nicht den verschiedenen Unabhängigkeitsstrukturen zugeordnet werden. Um dies zu gewährleisten und damit die Testbarkeit aller möglichen graphischen Modelle herzustellen, werden die kanonischen Parameter der bedingten Gaußverteilung „ausgedehnt". Mögliche solcher Erweiterungen sind beispielsweise in Darroch und Speed (1983) beschrieben. In der vorliegenden Situation erfolgt die Parameterausdehnung in Analogie zu der Darstellung log-linearer Modelle, vgl. Lauritzen und Wermuth (1989), d. h. $\alpha_i = \sum_{\{a:a\subseteq\Delta\}} \lambda^a_{i_a}, \quad \beta_i = \sum_{\{a:a\subseteq\Delta\}} \eta^{\gamma;a}_{i_a}, \quad \Omega_i = \sum_{\{a:a\subseteq\Delta\}} \psi^{\gamma,\mu;a}_{i_a}$.
Um die Eindeutigkeit der Parameter zu gewährleisten, müssen zusätzliche Bedingungen erfüllt sein. Analog zu den Restriktionen für die Parameter im log-linearen Modell wird auch hier eine Referenzzelle i^* festgelegt und es gilt:

$$\begin{aligned}
\lambda^a_{i_a} &= 0, \text{ falls } i_\delta = i^*_\delta,\ \delta \in a \subseteq \Delta,\\
\eta^{\gamma;a}_{i_a} &= 0, \text{ falls } i_\delta = i^*_\delta,\ \delta \in a \subseteq \Delta,\ \gamma \in \Gamma,\\
\psi^{\gamma,\mu;a}_{i_a} &= 0, \text{ falls } i_\delta = i^*_\delta,\ \delta \in a \subseteq \Delta,\ \gamma,\ \mu \in \Gamma.
\end{aligned}$$

Damit kann nun die Dichte der CG-Verteilung wie folgt beschrieben werden:

$$f_{(X_\Delta, X_\Gamma)}(i,y) = \exp\left\{ \sum_{\{a:a\subseteq\Delta\}} \lambda^a_{i_a} + \sum_{\{a:a\subseteq\Delta\}} \sum_{\gamma\in\Gamma} \eta^{\gamma;a}_{i_a} y_\gamma - \frac{1}{2} \sum_{\{a:a\subseteq\Delta\}} \sum_{\gamma,\,\mu\in\Gamma} \psi^{\gamma,\mu;a}_{i_a} y_\gamma y_\mu \right\}.$$

$\lambda^a_{i_a}$ beschreibt die Wechselwirkungen der diskreten Variablen aus $a \subseteq \Delta$. Ist $|a| = 1$, so wird $\lambda^a_{i_a}$ als Haupteffekt des Merkmals a bezeichnet; für $a = \emptyset$ ist $\lambda^\emptyset$ eine Normalisierungskonstante. Die Parameter $\eta^{\gamma;a}_{i_a}$ und $\psi^{\gamma,\mu;a}_{i_a}$ repräsentieren die Interaktionen zwischen den stetigen und den diskreten Variablen, bzw. für $a = \emptyset$ die jeweiligen Haupteffekte der stetigen Variablen. Speziell beschreibt $\eta^{\gamma;a}_{i_a}$ die so genannte „lineare Wechselwirkung", d. h. die Wechselwirkung zwischen einer stetigen Variable γ und einer Menge a von diskreten Merkmalen. Ist $a = \emptyset$, so ist $\eta^\emptyset$ der lineare Haupteffekt von γ. $\psi^{\gamma,\mu;a}_{i_a}$ stellt entsprechend die „quadratische Wechselwirkung", also die Interaktion zwischen zwei stetigen Variablen γ, μ und $a \subseteq \Delta$ dar. Ein quadratischer Haupteffekt liegt vor, wenn $\gamma = \mu$ ist und $a = \emptyset$. Lauritzen und Wermuth (1989) haben für diese Darstellung der CG-Verteilung gezeigt, dass zwei Merkmale genau dann voneinander bedingt unabhängig sind, gegeben alle anderen Variablen, wenn alle Interaktionseffekte, die diese beiden Variablen beinhalten, gleich Null sind.

Satz 2.2.6

Es sei $X = (X'_\Delta, X'_\Gamma)'$ ein $q+r$-dimensionaler Vektor gemischt stetiger-diskreter Zufallsvariablen mit einer CG-Verteilung als Modellannahme gegeben. Diese CG-Verteilung erfüllt die Markov-Eigenschaften gemäß Definition 2.2.2 über dem Graphen $\mathcal{G}$, wenn die zugehörigen erweiterten Interaktionsparameter der CG-Verteilung folgende Bedingungen erfüllen:

$$\begin{aligned} \lambda_{i_a}^{a} &= 0, &\quad& \textit{außer wenn } a \subseteq \Delta \textit{ vollständig ist,} \\ \eta_{i_a}^{\gamma;a} &= 0, && \textit{außer wenn } a \cup \{\gamma\} \textit{ vollständig ist, } a \subseteq \Delta,\ \gamma \in \Gamma, \\ \psi_{i_a}^{\gamma,\mu;a} &= 0, && \textit{außer wenn } a \cup \{\gamma, \mu\} \textit{ vollständig ist, } a \subseteq \Delta,\ \gamma, \mu \in \Gamma. \end{aligned}$$

Beweis: Lauritzen (1998), S. 174 f.

Die Repräsentation dieses Modells durch einen (ungerichteten) Unabhängigkeitsgraphen erfolgt gemäß Definition 2.2.1. Dabei bleibt unberücksichtigt, ob die zugrunde liegende CG-Verteilung homogen ist, d. h. ein Graph stellt immer sowohl das homogene als auch das entsprechende heterogene Modell dar (vgl. Definition 2.2.5).

Gilt für die Menge der stetigen Variablen $\Gamma = \emptyset$, so entspricht die Klasse der graphischen Modelle der Klasse der graphischen log-linearen Modelle; werden ausschließlich stetige Merkmale untersucht, d. h. $\Delta = \emptyset$, so reduziert sich die Klasse der graphischen Modelle zur Klasse der graphischen Gauß-Modelle.

Kapitel 3

Algebraische Statistik

In dem folgenden Kapitel werden die verwendeten Begriffe sowie Methoden der algebraischen Statistik beschrieben. Hierbei handelt es sich um ein junges Forschungsgebiet der Statistik, das Verfahren aus der computergestützten Algebra zur Lösung statistischer Fragestellungen anwendet. Diaconis und Sturmfels (1998) entwickeln neue algebraische Methoden zur Analyse kategorialer Daten. Die zweite grundlegende Arbeit stammt von Pistone und Wynn (1996), in der ein algebraischer Ansatz für die statistische Versuchsplanung beschrieben wird.

Im Fokus der vorliegenden Arbeit steht der von Diaconis und Sturmfels vorgeschlagene Algorithmus, siehe Diaconis und Sturmfels (1998). Dieser verbindet wichtige Begriffe der computergestützten Algebra wie Ideale, Varietäten und Gröbner-Basen mit Markov Chain Monte Carlo- (MCMC-) Methoden, um eine Stichprobe aus der bedingten Verteilung einer diskreten Exponentialfamilie mit beobachteter suffizienter Statistik zu simulieren. Auf dieser Grundlage können beispielsweise neue algebraische Testverfahren entwickelt werden, die als bedingte Tests in k-parametrigen Exponentialfamilien bestimmte Optimalitätseigenschaften aufweisen (vgl. Witting (1985), Kapitel 3.3 und Ferguson (1967), Kapitel 5). Damit liefert der Algorithmus von Diaconis und Sturmfels eine wichtige und nützliche Ergänzung zu traditionellen asymptotischen und exakten Verfahren.

Es folgt zunächst eine kurze Einführung in die computergestützte Algebra. Dazu werden Varietäten und Ideale definiert und ihre Struktur anhand von Gröbner-Basen spezifiziert. Mit dieser Grundlage kann anschließend der Diaconis-Sturmfels-Algorithmus eingeführt werden. Als verwendete Literatur sei auf die Bücher von Cox et al. (1997), Kreuzer und Robbiano (2000) sowie Pistone et al. (2000) verwiesen.

3.1 Varietäten und Ideale

In diesem Abschnitt werden Ideale und Varietäten als Grundbegriffe der computergestützten Algebra beschrieben. In der vorliegenden Arbeit erweist es sich als zweckmäßig, diese direkt über den Spezialfall eines Polynomringes zu definieren. Anschließend wird die Beziehung zwischen Idealen und Varietäten spezifiziert. Für den algebraischen Hintergrund sei auf Heinhold und Riedmüller (1975), Fischer und Sacher (1978) sowie Artin (1993) verwiesen.

Definition 3.1.1 (Monom und Polynom)

i) Es seien $\alpha_1, \ldots, \alpha_n \in \mathbb{Z}_{\geq 0}$. Ein Produkt der Form $x^\alpha := x_1^{\alpha_1} \cdot x_2^{\alpha_2} \cdot \ldots \cdot x_n^{\alpha_n}$ wird als Monom bezeichnet. Der Grad des Monoms beträgt $|\alpha| := \alpha_1 + \ldots + \alpha_n$.

ii) Es seien K ein beliebiger Körper und N^* eine endliche Teilmenge von $\mathbb{Z}^n_{\geq 0}$. Dann heißt ein Ausdruck der Form

$$p = \sum_{\alpha \in N^*} a_\alpha \, x^\alpha = \sum_{(\alpha_1,\ldots,\alpha_n)' \in N^*} a_{(\alpha_1,\ldots,\alpha_n)} \, x_1^{\alpha_1} \cdot \ldots \cdot x_n^{\alpha_n}$$

mit $a_\alpha \in K$ und $\alpha = (\alpha_1, \ldots, \alpha_n)' \in \mathbb{Z}^n_{\geq 0}$ Polynom in den Unbestimmten $x_1, \ldots, x_n$ über dem Körper K. Weiter gilt:

- *Ist $a_\alpha \neq 0$, so heißt $a_\alpha x^\alpha$ Term des Polynoms p*
- *Der totale Grad von p, $deg(p)$, ist der maximale Grad $|\alpha|$ aller im Polynom vorkommenden Monome.*

Es bezeichne $K[x]$ die Menge aller Polynome in einer Unbestimmten x über dem Körper K. Mit der üblichen Definition für die Summation und Multiplikation von Polynomen ist $(K[x], +, \cdot)$ ein kommutativer Polynomring mit Einselement (vgl. z.B. Heinhold und Riedmüller (1975), S. 56f.). Üblicherweise wird für den Ring $(K[x], +, \cdot)$ abkürzend $K[x]$ geschrieben, wenn bekannt ist, welche Ringstruktur auf $K[x]$ zugrunde liegt. Da in dieser Arbeit ausschließlich kommutative Polynomringe mit Einselement betrachtet werden, wird der Begriff des Polynomrings synonym verwendet. Grundlegend für die weiteren Ausführungen sind Polynomringe in endlich vielen Unbestimmten. Diese können rekursiv aus Polynomringen in einer Unbestimmten hergeleitet werden, siehe z.B. Fischer und Sacher (1978), Kapitel 2. Dies ermöglicht nun die Definition von Varietäten und Idealen:

Definition 3.1.2 (Varietät)
Es seien K ein Körper und $\{p_1, \ldots, p_s\}$ eine Menge von Polynomen in $K[x_1, \ldots, x_n]$. Dann heißt

$$V(p_1, \ldots, p_s) = \{(a_1, \ldots, a_n) \in K^n : p_i(a_1, \ldots, a_n) = 0 \ \textit{ für alle } 1 \leq i \leq s\}$$

eine (affine) Varietät. Sie wird bestimmt durch $\{p_1, \ldots, p_s\}$.

Als Lösungsmenge des polynomialen Gleichungssystems $p_1(x_1, \ldots, x_n) = \ldots = p_s(x_1, \ldots, x_n) = 0$ hängt eine Varietät von der Menge ab, aus der die Unbestimmten $x_1, \ldots, x_n$ entstammen, so hat z.B. $x^2 - 1 = 0$ in $I\!R$ keine Lösung. Damit ist $V(x^2 - 1)$ die leere Menge in $I\!R$; in $\mathbb{C}$ ist die Lösung $\{i, -i\}$.

Definition 3.1.3 (Polynomideal)
Es sei $K[x_1, \ldots, x_n]$ ein kommutativer Polynomring mit Einselement. Eine Teilmenge $\mathscr{I}$ von $K[x_1, \ldots, x_n]$ heißt Polynomideal (oder kurz: Ideal) von $K[x_1, \ldots, x_n]$, wenn gilt: i) $0 \in \mathscr{I}$;

ii) Sind $p, g \in \mathscr{I}$, dann folgt $p + g \in \mathscr{I}$,;

iii) Ist $p \in \mathscr{I}$ und $h \in K[x_1, \ldots, x_n]$, dann ist $(h \cdot p) \in \mathscr{I}$.

Die Kombination der Bedingungen *ii)* und *iii)* aus Definition 3.1.3 führt zur folgenden Darstellung eines Ideals:

Definition 3.1.4
Es sei $\{p_1, \ldots, p_s\}$ eine Menge von Polynomen in $K[x_1, \ldots, x_n]$. Setze

$$< p_1, \ldots, p_s > = \{\sum_{i=1}^{s} h_i p_i : h_1, \ldots, h_s \in K[x_1, \ldots, x_n]\}.$$

Dann beschreibt $< p_1, \ldots, p_s >$ *ein durch* $\{p_1, \ldots, p_s\}$ *erzeugtes Ideal von* $K[x_1, \ldots, x_n]$ *(ideal generated by polynomials).* $\{p_1, \ldots, p_s\}$ *wird als Basis des Polynomideals bezeichnet.*

Die Menge $\{p_1, \ldots, p_s\}$ erzeugt ein Ideal, vgl. Cox et al. (1997), S. 29. Im Allgemeinen hat ein Ideal keine eindeutige Basis, siehe Cox et al. (1997), S. 31 für ein Beispiel.

In dem folgenden Satz wird die Beziehung zwischen Idealen und Varietäten beschrieben. So hängen Varietäten von Polynomidealen bzw. deren Basen ab, d.h. verschiedene Basen eines Ideals erzeugen dieselbe Varietät.

Satz 3.1.5

Es seien $\{p_1, \ldots, p_s\}$ *und* $\{g_1, \ldots, g_u\}$ *Basen desselben Ideals in* $K[x_1, \ldots, x_n]$*, so dass* $< p_1, \ldots, p_s >=< g_1, \ldots, g_u >$*. Dann folgt:* $V(p_1, \ldots, p_s) = V(g_1, \ldots, g_u)$*.*

Beweis: Diese Eigenschaft kann unter Verwendung von Definition 3.1.4 verifiziert werden.

3.2 Struktur von Polynomidealen – Gröbner-Basis

Im Weiteren wird der Aufbau und die Struktur eines Ideals charakterisiert. Die Polynomdivision erweist sich hierfür als wertvolles Hilfsmittel. Ihre Anwendung bei Polynomen in mehreren Unbestimmten erfordert eine Vergleichbarkeit der Monome, d. h. es wird eine Ordnung in den Unbestimmten benötigt. Häufig existiert keine natürliche Reihenfolge, so dass eine geeignete Ordnung vorzugeben ist. Zur Beschreibung der Idealstruktur eignet sich insbesondere eine so genannte Gröbner-Basis.

Monome $x^\alpha = x_1^{\alpha_1} \cdot x_2^{\alpha_2} \cdot \ldots \cdot x_n^{\alpha_n}$ lassen sich aus den Exponentenvektoren $\alpha = (\alpha_1, \ldots, \alpha_n)' \in \mathbb{Z}^n_{\geq 0}$ rekonstruieren. Daher kann eine Ordnung $\succ$ auf $\mathbb{Z}^n_{\geq 0}$ auf Monome in $K[x_1, \ldots, x_n]$ übertragen werden.

Definition 3.2.1 (Monomordnung)

Eine Monomordnung auf $K[x_1, \ldots, x_n]$ *ist eine Ordnung* $\succ$ *auf* $\mathbb{Z}^n_{\geq 0}$ *oder äquivalent eine Ordnung* $\succ$ *auf der Menge der Monome* $x^\alpha, \alpha \in \mathbb{Z}^n_{\geq 0}$*, für die gilt:*

i) $\succ$ *ist eine totale Ordnung auf* $\mathbb{Z}^n_{\geq 0}$*, d.h.* $\alpha \succ \beta$*,* $\alpha = \beta$ *oder* $\alpha \prec \beta$ *für alle* $\alpha, \beta \in \mathbb{Z}^n_{\geq 0}$*;*

ii) Ist $\alpha \succ \beta$ *und* $\alpha, \beta, \gamma \in \mathbb{Z}^n_{\geq 0}$*, dann ist* $\alpha + \gamma \succ \beta + \gamma$*;*

iii) $\succ$ *ist wohlgeordnet auf* $\mathbb{Z}^n_{\geq 0}$*, d.h. es existiert ein kleinstes Element.*

Es gibt verschiedene Monomordnungen; im Folgenden werden zwei sehr gebräuchliche kurz vorgestellt. Die *lexikographische Ordnung* erstellt die Reihenfolge der Terme eines Polynoms gemäß einer zugrunde liegenden Ordnung wie dem Alphabet:
Es seien $\alpha = (\alpha_1, \cdots, \alpha_n)'$ und $\beta = (\beta_1, \cdots \beta_n)' \in \mathbb{Z}^n_{\geq 0}$. Dann gilt:

$$\alpha \succ_{lex} \beta \Leftrightarrow \text{ es existiert ein } i \in \{1, 2, \ldots, n\} \text{ mit: } \alpha_1 - \beta_1 = 0, \alpha_2 - \beta_2 = 0, \cdots, \alpha_{i-1} - \beta_{i-1} = 0 \text{ und } \alpha_i - \beta_i > 0,$$

d.h. der erste Eintrag der Vektordifferenz $\alpha - \beta \in \mathbb{Z}^n$ ungleich Null ist positiv. Damit ist $x^\alpha \succ_{lex} x^\beta$, wenn $\alpha \succ_{lex} \beta$.

Die *gradlexikographische Ordnung* berücksichtigt den Grad des Monoms sowie zusätzlich die lexikographische Ordnung:
Es seien $\alpha = (\alpha_1, \cdots, \alpha_n)'$ und $\beta = (\beta_1, \cdots \beta_n)' \in \mathbb{Z}^n_{\geq 0}$. Dann gilt:

$$\alpha \succ_{grlex} \beta \Leftrightarrow |\alpha| = \sum_{i=1}^n \alpha_i > |\beta| = \sum_{i=1}^n \beta_i \text{ oder } |\alpha| = |\beta| \text{ und } \alpha \succ_{lex} \beta.$$

Betrachtet wird beispielsweise das Polynom $p = 5x^2y + y^2 - 7xz^3 \in \mathbb{Z}[x, y, z]$. Die lexikographische Anordnung der Monome dieses Polynoms ist $p = 5x^2y - 7xz^3 + y^2$. Für die gradlexikographische Monomordnung gilt $p = -7xz^3 + 5x^2y + y^2$.

Gemäß der gewählten Monomordnung kann nun eine Reihenfolge der Monome innerhalb eines Polynoms erstellt werden. Zudem ergeben sich nützliche Merkmale von Polynomen bezüglich der zugrunde liegenden Monomordnung.

Definition 3.2.2

Es sei $p = \sum_\alpha a_\alpha x^\alpha$ *ein von Null verschiedenes Polynom in* $K[x_1, \ldots, x_n]$ *und* $\succ$ *sei eine beliebige Monomordnung. Dann bedeutet*

- *Multigrad von p: multideg(p)* $= max(\alpha \in \mathbb{Z}^n_{\geq 0} : a_\alpha \neq 0)$.
  (Die Bildung des Maximums erfolgt bezüglich der Monomordnung.)

· *Leitkoeffizient von p:* $LC(p) = a_{multideg(p)} \in K$.

· *Leitmonom von p:* $LM(p) = x^{multideg(p)}$, *wobei der Koeffizient 1 beträgt.*

· *Leitterm von p:* $LT(p) = LC(p) \cdot LM(p)$.

Damit kann die Polynomdivision wie folgt beschrieben werden:

Definition 3.2.3 (Polynomdivision mit Rest in mehreren Unbestimmten)
Es seien $\succ$ *eine Monomordnung auf* $\mathbb{Z}^n_{\geq 0}$ *und* $(p_1, \ldots, p_s)$ *ein entsprechend geordnetes s-Tupel von Polynomen in* $K[x_1, \ldots, x_n]$. *Dann ist jedes* $p \in K[x_1, \ldots, x_n]$ *darstellbar als* $p = a_1p_1 + a_2p_2 + \ldots + a_sp_s + r$ *mit* $a_i, r \in K[x_1, \ldots, x_n], i = 1, \ldots, s$, *und es gilt:*

i) Der Rest r *ist gleich Null oder* r *ist eine Linearkombination von Monomen und keines von ihnen ist teilbar durch* $LT(p_1), \ldots, LT(p_s)$.

ii) Ist $a_ip_i \neq 0$, *dann folgt:* $multideg(p) \geq multideg(a_ip_i)$.

Eine wichtige Eigenschaft der Polynomdivision ist, dass das Ergebnis von der Reihenfolge der Divisoren abhängt, siehe Cox et al. (1997), S. 60 und S. 64 für ein Beispiel.

Für die Strukturbeschreibung eines Ideals ist die Definition des so genannten Leittermideals notwendig.

Definition 3.2.4
Es sei $\mathscr{I} \subset K[x_1, \ldots, x_n]$ *ein von Null verschiedenes Ideal und* $\succ$ *die gewählte Monomordnung. Dann ist*

i) $LT(\mathscr{I}) = \{cx^\alpha : \text{es existiert ein } p \in \mathscr{I} \text{ mit } LT(p) = cx^\alpha\}$
die Menge der Leitterme von Elementen aus $\mathscr{I}$ *bezüglich* $\succ$.

ii) $< LT(\mathscr{I}) >$ *das von den Elementen von* $LT(\mathscr{I})$ *erzeugte Ideal, das Leittermideal bezüglich* $\succ$ *(ideal of leading terms).*

Mit Hilfe der Polynomdivision und Definition 3.2.4 kann nun der Aufbau eines Ideals $\mathscr{I} \subset K[x_1, \ldots, x_n]$ wie folgt beschrieben werden:

Theorem 3.2.5 (Hilbertscher Basissatz)

Jedes Ideal $\mathscr{I} \subset K[x_1, \ldots, x_n]$ hat eine endliche Basis, d.h. es existiert eine Menge $\{g_1, \ldots, g_u\}$, so dass gilt: $\mathscr{I} =< g_1, \ldots, g_u >$.

Beweis: Cox et al. (1997), S. 74.

Mit Definition 3.2.3 und Theorem 3.2.5 wird die Bedeutung der Polynomdivision für die Strukturbeschreibung eines Ideals deutlich: Ein Polynom p ist Element eines Ideals $\mathscr{I} =< g_1, \ldots, g_u >$, wenn es sich als Linearkombination (ohne Rest) durch die Basispolynome darstellen lässt. Wie bereits angemerkt, kann ein Ideal verschiedene Basen besitzen. Zudem ist das Ergebnis der Polynomdivision im Allgemeinen von der Reihenfolge der Divisoren abhängig. Eine Gröbner-Basis ist eine spezielle Basis eines Ideals mit besonders nützlichen Eigenschaften.

Definition 3.2.6 (Gröbner-Basis)

Es seien $\succ$ eine festgelegte Monomordnung auf $K[x_1, \ldots, x_n]$, $\mathscr{I} \subset K[x_1, \ldots, x_n]$ ein Ideal und $\mathscr{G} = \{g_1, \ldots, g_u\}$ eine endliche Teilmenge des Ideals $\mathscr{I}$. Dann heißt $\mathscr{G}$ Gröbner-Basis, wenn gilt: $< LT(g_1), \ldots, LT(g_u) >=< LT(\mathscr{I}) >$.

Für eine beliebige Basis $\{p_1, \ldots, p_s\}$ gilt für alle $1 \leq i \leq s$: $LT(p_i) \in LT(\mathscr{I}) \subseteq < LT(\mathscr{I}) >$. Daraus folgt, dass $< LT(p_1), \ldots, LT(p_s) > \subseteq < LT(\mathscr{I}) >$, wobei $LT(\mathscr{I})$ strikt größer sein kann. Mit der Definition einer Gröbner-Basis wird außerdem deutlich, dass $\mathscr{G}$ eine Gröbner-Basis bezüglich der festgelegten Monomordnung $\succ$ ist.

Satz 3.2.7

Es sei eine Monomordnung $\succ$ auf $K[x_1, \ldots, x_n]$ festgelegt. Dann besitzt jedes von $\{0\}$ verschiedene Ideal $\mathscr{I} \subset K[x_1, \ldots, x_n]$ eine Gröbner-Basis. Zudem ist jede Gröbner-Basis eines Ideals $\mathscr{I}$ eine Basis des Ideals.

Beweis: Cox et al. (1997), S.75.

Die Bedeutung der Gröbner-Basis wird anhand ihrer Eigenschaften deutlich:

Satz 3.2.8
Es sei $\mathscr{G} = \{g_1, \ldots, g_u\}$ *eine Gröbner-Basis für ein Ideal* $\mathscr{I} \subset K[x_1, \ldots, x_n]$, *und ein Polynom* p *sei aus* $K[x_1, \ldots, x_n]$. *Dann gilt:*

i) *Wird* p *durch* $g_1, \ldots, g_u$ *dividiert, so existiert ein eindeutiger Rest* r *mit den folgenden Eigenschaften:*

 · *Kein Term von* r *ist durch ein* $LT(g_1), \ldots, LT(g_u)$ *dividierbar.*

 · *Es gibt ein* $g \in \mathscr{I}$, *so dass* $p = g + r$ *gilt.*

ii) *Es gilt* $p \in \mathscr{I}$ *genau dann, wenn der Rest* r *bei der Division von* p *durch* $\mathscr{G}$ *Null ist.*

Beweis: Cox et al. (1997), S. 79f.

Begründet auf eine Idee seines Doktorvaters Wolfgang Gröbner entwickelte Buchberger in seiner Dissertation 1965 den nach ihm benannten Algorithmus zur Bestimmung von Gröbner-Basen, vgl. Čižmár (2002). Aufgrund der zentralen Bedeutung dieser Dissertation ist 2006 eine englische Übersetzung erschienen (Buchberger (2006)). Für eine Beschreibung des Buchberger-Algorithmus sei auf den Anhang A.1 verwiesen. Insbesondere liefert dieser für jedes Polynomideal $\mathscr{I} \neq \{0\} \subset K[x_1, \ldots, x_n]$ in endlich vielen Schritten eine Gröbner-Basis. Der damit verbundene Rechenaufwand und somit auch die benötigte Rechenzeit kann allerdings noch optimiert werden. Für eine Beschreibung einiger Modifikationen des Buchberger-Algorithmus sei z.B. auf Cox et al. (1997), Kapitel 2.9, verwiesen. Das Vorgehen von Buchberger liefert eine Gröbner-Basis, die häufig größer ist als nötig.

Definition 3.2.9 (reduzierte Gröbner-Basis)
Es sei $\mathscr{G}$ eine Gröbner-Basis für ein Polynomideal $\mathscr{I}$. Gilt zusätzlich

i) $LC(p) = 1$ *für alle Polynome* $p \in \mathscr{G}$,

ii) *für alle Polynome* $p \in \mathscr{G}$ *ist kein Monom von* p *Element von* $< LT(\mathscr{G} \setminus \{p\}) >$,

dann heißt $\mathscr{G}$ *die reduzierte Gröbner-Basis für* $\mathscr{I}$.

Die reduzierte Gröbner-Basis weist zwei wichtige Eigenschaften auf: Zum einen ist sie minimal, d.h. sie enthält keine überflüssigen Basispolynome. Zum anderen ist sie eindeutig bestimmt, siehe z.B. Cox et al. (1997), S. 89f.

3.3 Der Diaconis-Sturmfels-Algorithmus

Diaconis und Sturmfels (1998) leisten einen herausragenden Beitrag zur Verwendung der computergestützten Algebra in der Statistik. Dabei nutzen sie Markov Chain Monte Carlo-Verfahren für die Simulation einer Stichprobe aus der bedingten Verteilung einer diskreten Exponentialfamilie mit gegebener suffizienter Statistik. Die Schnittstelle beider Disziplinen bildet der Metropolis-Hastings-Algorithmus als verwendete MCMC-Methode. Im Folgenden werden zunächst Markov-Ketten sowie der Metropolis-Hastings-Algorithmus vorgestellt. Für eine ausführliche Darstellung dieser Thematik sei auf Fahrmeir et al. (1981), Chib und Greenberg (1995), Sørensen und Gianola (2002), Robert und Casella (2004) sowie Rubinstein und Kroese (2008) verwiesen. Anschließend wird die Beziehung zwischen dem Metropolis-Hastings-Algorithmus und der computergestützten Algebra spezifiziert: Diaconis und Sturmfels (1998) schlagen eine Gröbner-Basis für die Konstruktion der Markov-Kette vor.

Ein stochastischer Prozess $\mathscr{X} = \{X_s, s \in \mathbb{N}_0\}$ mit diskretem Zustandsraum E heißt genau dann (diskrete) Markov-Kette, wenn für alle $s \in \mathbb{N}_0$ und für alle $j,\ i,\ i_{s-1},\ \ldots,\ i_0 \in E$ gilt: $P(X_{s+1} = j \,|\, X_s = i, X_{s-1} = i_{s-1}, \ldots, X_0 = i_0) = P(X_{s+1} = j \,|\, X_s = i)$.

Die stochastische Abhängigkeit zweier Zustände i und j einer Markov-Kette wird mit Übergangswahrscheinlichkeiten $p^{(s+1,s)}(i,j) := P(X_{s+1} = j | X_s = i)$ beschrieben. In der vorliegenden Arbeit werden ausschließlich so genannte homogene Markov-Ketten betrachtet, d.h. es gilt für alle $s,\ s' \in I\!N_0$, dass $p^{(s+1,s)}(i,j) = p^{(s'+1,s')}(i,j) =: p(i,j)$. Diese werden in der Übergangsmatrix $\mathtt{P} := \{p(i,j)\}_{i,j\in E}$ angeordnet und für alle $i,\ j \in E$ gilt: $p(i,j) \geq 0$ und $\sum_{j\in E} p(i,j) = 1$. Die Zustandswahrscheinlichkeit einer Markov-Kette $\mathscr{X} = \{X_s, s \in I\!N_0\}$ wird durch $\pi^{(s)}(i) := P(X_s = i)$, $i \in E$, beschrieben. Entsprechend stellt $\pi^{(s)} := (\pi^{(s)}(i))_{i\in E}$ die Zustandsverteilung der Markov-Kette dar. Für $s = 0$ heißt $\pi^{(0)}$ Anfangsverteilung. Mit dem Satz von der totalen Wahrscheinlichkeit gilt für alle $s > 0$, dass $\pi^{(s)} = \pi^{(s-1)} \cdot \mathtt{P}$. Damit ist die Zustandsverteilung einer Markov-Kette gegeben durch $\pi^{(s)} = \pi^{(0)} \cdot \mathtt{P}^s$. Konvergiert $\pi^{(s)}$ für $s \to \infty$ unabhängig von $\pi^{(0)}$ gegen eine Verteilung, so wird die Markov-Kette stationär genannt. Der Zeilenvektor $\pi = (\pi(i))_{i\in E}$ heißt stationäre Verteilung dieser Markov-Kette genau dann, wenn gilt: $\pi = \pi \cdot \mathtt{P}$ und $\pi \cdot \mathbb{1} = 1$, wobei $\mathbb{1}$ der entsprechende Einservektor ist.

Jede irreduzible, aperiodische, positiv rekurrente Markov-Kette besitzt eine stationäre Verteilung, siehe etwa Fahrmeir et al. (1981). Diese Eigenschaften werden nachfolgend kurz erläutert. Eine Markov-Kette heißt irreduzibel, wenn jeder Zustand der Markov-Kette aus jedem anderen Zustand in endlich vielen Schritten erreicht werden kann, d.h. für alle i und j gibt es mindestens ein $s \geq 0$, so dass gilt $P(X_{s+r} = j|X_r = i) > 0$, $i, j \in E$, $r \in I\!N_0$. Ein Zustand $i \in E$ einer Markov-Kette heißt aperiodisch, wenn es ein $r \in I\!N_0$ gibt, so dass gilt: $P(X_r = i|X_0 = i)$ und $P(X_{r+1} = i|X_0 = i) > 0$. Eine Markov-Kette heißt aperiodisch, wenn alle Zustände $i \in E$ aperiodisch sind. Ein Zustand i heißt positiv rekurrent, falls die Rückkehrwahrscheinlichkeit gleich Eins und die erwartete Rückkehrzeit endlich ist, vergleiche Sørensen und Gianola (2002), Kapitel 10.

Markov Chain Monte Carlo-Methoden ermöglichen die Simulation von Stichproben aus einer interessierenden Verteilung, der so genannten Zielverteilung. Dazu wird eine Markov-Kette generiert, deren stationäre Verteilung die Zielverteilung ist. Ein besonders wichtiges MCMC-Verfahren ist der Metropolis-Hastings-Algorithmus. Dieser basiert auf Arbeiten einer Forschergruppe um Nicholas C. Metropolis im Jahr 1953 und wurde 1970 von W. Keith Hastings verallgemeinert.

Der Metropolis-Hastings-Algorithmus kann für Markov-Ketten mit stetigem und diskretem Zustandsraum E angewendet werden. Befindet sich die Markov-Kette derzeit im Zustand $i \in E$, wird zunächst mit Wahrscheinlichkeit $q(i,j)$ der Zustand $j \in E$

vorgeschlagen. $q(i,j)$ bezeichnet die so genannte Vorschlagsdichte (proposal distribution) und im Fall eines diskreten Zustandsraums E ist $\sum\limits_{j \in E} q(i,j) = 1$. Anschließend wird der potenzielle neue Zustand j der Markov-Kette angenommen oder abgelehnt. Die so genannte Reversibilitätsbedingung sichert ab, dass die Markov-Kette genau so häufig von i nach j wie von j nach i wechselt. Diese Bedingung wird durch Einführung der Akzeptanzwahrscheinlichkeit $\alpha(i,j)$ kontrolliert. Damit wechselt die Markov-Kette von i nach j, i, $j \in E$ gemäß

$$p(i,j) := \begin{cases} q(i,j) \cdot \alpha(i,j), & \text{falls } i \neq j, \\ 0, & \text{sonst.} \end{cases}$$

Unter Beachtung der Reversibilitätsbedingung wird die Akzeptanzwahrscheinlichkeit α bestimmt durch

$$\alpha(i,j) = \begin{cases} min(\frac{\pi^*(j)q(j,i)}{\pi^*(i)q(i,j)}, 1), & \text{falls } \pi^*(i)q(i,j) > 0, \\ 1, & \text{sonst,} \end{cases}$$

wobei π^* die Wahrscheinlichkeitsfunktion der Markov-Kette bezeichne. Für die Anwendung des Metropolis-Hastings-Algorithmus muss daher π^* bekannt sein. Ist q symmetrisch, so entspricht der Metropolis-Hastings-Algorithmus dem ursprünglichen Metropolis-Algorithmus. Üblicherweise wird dieses Verfahren dennoch als Metropolis-Hastings-Algorithmus bezeichnet. Da diese MCMC-Methode für die vorliegende Arbeit von großer Bedeutung ist, wird der Pseudocode angegeben.

Korollar 3.3.1 (Metropolis-Hastings-Algorithmus)

Initialisierung:

- *Wahl des Startwertes i_0 und der Vorschlagsdichte $q(\cdot,\cdot)$*
- *Festlegung der Kettenlänge l*

Wiederhole für $s = 1, \cdots, l$:

- *Ziehung einer Zufallszahl j aus $q(i,\cdot)$ und u aus Gleichverteilung auf $[0,1]$*
- *Gilt: $u \leq \alpha(i,j)$, so wird j als neuer Zustand der Markov-Kette angenommen: setze $X_s = j$ andernfalls $X_s = i$*

Stoppe: *wenn $s = l$*

Ausgabe: *Markov-Kette $\{X_1, X_2, \cdots, X_l\}$*

Aufgrund der Konstruktion des Metropolis-Hastings-Algorithmus ist die so generierte Kette reversibel. Die Markov-Kette ist aperiodisch und irreduzibel, wenn die Vorschlagsdichte $q(\cdot,\cdot)$ positiv ist und den gleichen Träger hat wie $\pi^*(\cdot)$ oder wenn $q(\cdot,\cdot)$ einen beschränkten Träger hat, vgl. Chib und Greenberg (1995).

Die Wahl der Vorschlagsdichte bestimmt den Metropolis-Hastings-Algorithmus entscheidend. Oftmals ist es schwierig, ein geeignetes $q(\cdot,\cdot)$ zu finden; Chib und Greenberg (1995) beschreiben hierzu einige Möglichkeiten. Diaconis und Sturmfels (1998) schlagen für die Simulation aus der bedingten Verteilung einer diskreten Exponentialfamilie mit beobachteter suffizienter Statistik eine Markov-Basis zur Bestimmung von $q(\cdot,\cdot)$ vor.

Der Stichprobenraum $\mathscr{H}$ sei eine endliche Menge. Die Grundlage weiterer Betrachtungen sind diskrete Exponentialfamilien, d.h. die gemeinsame Wahrscheinlichkeitsfunktion von i. i. d. verteilten Zufallsvariablen $X_1,\ldots,X_N$ ist gegeben durch

$$P_\theta(X_1 = x_1,\ldots,X_N = x_N) = a(\theta)^N\, e^{\sum\limits_{i=1}^{d} B_i(\theta) \sum\limits_{j=1}^{N} T_i(x_j)} \prod_{j=1}^{N} c(x_j)$$

mit Parameter $\theta \in \Theta$, $B := (B_1,\ldots,B_d) : \Theta \to I\!R^d$ und $x_j \in \mathscr{H}$, $j = 1,\ldots,N$. $a(\theta)$ ist eine Normierungskonstante und c eine Funktion $c : \mathscr{H} \to [0,\infty)$. Die suffiziente Statistik T für $\theta \in \Theta$ wird beschrieben durch $T := (T_1,\ldots,T_d)'$ mit $T : \mathscr{H} \to I\!N^d$ und es ist $T(X) = \sum\limits_{j=1}^{N} T(X_j)$, siehe beispielsweise Witting (1985), Kapitel 1.7. Damit kann die Menge aller möglichen Datensätze mit beobachteter suffizienter Statistik $t = (t_1,\ldots,t_d)'$ dargestellt werden als

$$\mathscr{Z}_t = \{z : \mathscr{H} \to I\!N : \sum_{x\in\mathscr{H}} z(x)T^*(x) = t\},$$

wobei $T^* : \mathscr{H} \to I\!N^d$ aufgrund der getroffenen Modellannahmen festgelegt ist. Die Funktion $z : \mathscr{H} \to I\!N$ stellt die Zelleinträge der Kontingenztafel dar. Die Menge $\mathscr{Z}_t$ ist endlich und nicht leer. Die bedingte Verteilung einer diskreten Exponentialfamilie mit beobachteter suffizienter Statistik ist die Verteilung auf $\mathscr{Z}_t$; diese ist hypergeometrisch mit Dichte $H(z) = \frac{n!}{|\{(x_1,\ldots,x_N)\in\mathscr{H}^N : \sum_{j=1}^{N} T(x_j)=t\}|} \prod_{x\in\mathscr{H}} \frac{1}{z(x)!}$, vergleiche Diaconis und Sturmfels (1998), S. 365 ff. und Rapallo (2003).

Definition 3.3.2 (Markov-Basis)

Eine Menge von Funktionen $m_1, m_2, \ldots, m_L : \mathscr{H} \to \mathbb{Z}$, wird als Markov-Basis oder „Bewegungen“ bezeichnet, wenn gilt:

i) Für jedes i mit $1 \leq i \leq L$ ist $\sum_{x \in \mathscr{H}} m_i(x)T(x) = 0$;

ii) Für jedes t und $z, z' \in \mathscr{Z}_t$ existiert eine Folge von „Bewegungen“ $(m_{i_1}, \ldots, m_{i_A})$ sowie eine Folge $(\epsilon_1, \ldots, \epsilon_A)$ mit $\epsilon_j = \pm 1$, $j = 1, \ldots, A$, $A \in I\!N$, so dass

$$z' = z + \sum_{j=1}^{A} \epsilon_j m_{i_j} \quad \textit{und} \quad z + \sum_{j=1}^{a} \epsilon_j m_{i_j} \geq 0 \quad \textit{für } 1 \leq a \leq A.$$

Die Forderungen an eine Markov-Basis stellen sicher, dass der Wert t der suffizienten Statistik im neuen Zustand z' der Markov-Kette unverändert bleibt und dass die so generierte Markov-Kette irreduzibel ist, vgl. Rapallo (2003). Im Allgemeinen ist es schwierig, eine geeignete Markov-Basis zu finden. Diaconis und Sturmfels (1998) führen eine polynomiale Darstellung der interessierenden statistischen Fragestellung ein und ermöglichen damit die Anwendung von Methoden der computergestützten Algebra zur Bestimmung der gesuchten Markov-Basis. Dazu wird im Folgenden die verwendete Notation eingeführt und wichtige Begriffe erläutert.

Die Grundlage weiterer Betrachtungen sei ein Polynomring $K[\mathscr{H}]$ mit endlichem Stichprobenraum $\mathscr{H}$. Eine Funktion $f : \mathscr{H} \to I\!N$ sei im Folgenden durch ein Monom $\prod_{x \in \mathscr{H}} x^{f(x)}$ repräsentiert. Weiter sei eine Funktion T^* beschrieben durch den Ringhomomorphismus

$$\begin{array}{rcl} \varphi_{T^*} : K[\mathscr{H}] & \to & K[t_1, \ldots, t_d] \\ x & \to & t_1^{T_1^*(x)} t_2^{T_2^*(x)} \cdots t_d^{T_d^*(x)} =: t^{T^*(x)}. \end{array}$$

Der Kern des Ringhomomorphismus φ_{T^*} ist für diese Arbeit von zentraler Bedeutung; er entspricht einem so genannten „Torischen Ideal“ (toric ideal) $\mathscr{I}_{T^*} = \{p \in K[\mathscr{H}] : \varphi_{T^*}(p) = 0\}$, siehe Diaconis und Sturmfels, Kapitel 3.1. Für eine formale Definition sowie effizienter Berechnungsalgorithmen Torischer Ideale sei auf Bigatti et al. (1999) sowie Bigatti und Robbiano (2001) verwiesen.

Allgemein kann jede Funktion $m : \mathscr{H} \to \mathbb{Z}$ durch die Differenz $m(x) = m^+(x) - m^-(x)$, $m^+, m^- : \mathscr{H} \to I\!N$, mit $m^+(x) := \max(m(x), 0)$ und $m^-(x) := \max(-m(x), 0)$ dargestellt werden. Daraus ergibt sich dann die polynomiale Beschreibung einer Markov-Basis: $x^{m^-} - x^{m^+}$.

Theorem 3.3.3

Eine Menge von Funktionen $m_1, \ldots, m_L : \mathscr{H} \to \mathbb{Z}$ ist genau dann eine Markov-Basis gemäß Definition 3.3.2, wenn die Menge $\mathscr{H}^{m_i^+(x)} - \mathscr{H}^{m_i^-(x)} := \prod_{x \in \mathscr{H}} x^{m_i^+(x)} - \prod_{x \in \mathscr{H}} x^{m_i^-(x)}$, $1 \leq i \leq L$, das Polynomideal $\mathscr{I}_{T^*} = \{p \in K[\mathscr{H}] : \varphi_{T^*}(p) = 0\}$ erzeugt.

Beweis: Diaconis und Sturmfels (1998), S. 375 f.

Da nach Satz 3.2.7 jedes Ideal eine Gröbner-Basis besitzt, ist die Suche nach einer geeigneten Markov-Basis äquivalent zur Bestimmung einer Gröbner-Basis des Ideals $\mathscr{I}_{T^*}$. Diaconis und Sturmfels (1998) beschreiben das zugrunde liegende statistische Modell anhand einer Varietät. Mit $\mathscr{Z}_t$ und insbesondere mit T^* legen sie eine polynomiale Parametrisierung $x = \mathcal{T}^{T^*(x)}$ dieser Varietät fest, wobei $\mathcal{T}$ die Menge der einzelnen Einträge der suffizienten Statistik ist. Das Ziel ist nun, diese polynomiale Parametrisierung in definierende Gleichungen für die kleinste affine Varietät zu übertragen, die diese Parametrisierung enthält. Damit ist der Diaconis-Sturmfels-Algorithmus als Spezialfall des Algorithmus für die implizite Darstellung einer polynomialen Parametrisierung der so genannten Eliminationstheorie der computergestützten Algebra entlehnt, siehe Cox et al. (1997), Kapitel 3.

Theorem 3.3.4 (Algorithmus von Diaconis-Sturmfels)

Es sei $\mathscr{H}$ eine endliche Menge und $\mathcal{T} = \{T_1, \ldots, T_d\}$ die Menge der einzelnen Einträge der suffizienten Statistik des interessierenden statistischen Modells. Ferner seien $T^* : \mathscr{H} \to \mathbb{N}^d$ mit $T^* = (T_1^*, \ldots, T_d^*)'$ und eine Monomordnung $\succ$ auf $\mathscr{H}$ gegeben. Diese Ordnung wird erweitert für $\mathscr{H} \cup \mathcal{T}$, so dass $T_i \succ x$, $x \in \mathscr{H}$, $T_i \in \mathcal{T}$, $i = 1, \ldots, d$. Definiere das Hilfsideal $\mathscr{I}_{hilf} = \{x - \mathcal{T}^{T^*(x)},\ x \in \mathscr{H}\} \subset K[\mathscr{H}, \mathcal{T}]$, wobei gilt $\mathcal{T}^{T^*(x)} := T_1^{T_1^*(x)} \cdot T_2^{T_2^*(x)} \cdot \ldots \cdot T_d^{T_d^*(x)}$. Dann ist $\mathscr{I}_{T^*} = \mathscr{I}_{hilf} \cap K[\mathscr{H}]$. Für die Bestimmung einer Gröbner-Basis $\mathscr{G}$ von $\mathscr{I}_{T^*}$ wird zunächst eine reduzierte Gröbner-Basis $\mathscr{G}_{hilf}$ von $\mathscr{I}_{hilf}$ berechnet. In $\mathscr{G}$ sind alle Polynome aus $\mathscr{G}_{hilf}$ enthalten, die ausschließlich Terme aus $\mathscr{H}$ beinhalten.

Diese Gröbner-Basis beschreibt eine geeignete Vorschlagsdichte im Metropolis-Hastings-Algorithmus für die Simulation aus der bedingten Verteilung einer diskreten

Exponentialfamilie mit beobachteter suffizienter Statistik.

In dieser Arbeit ist die Analyse struktureller Zusammenhänge multivariater kategorialer Zufallsvektoren von besonderem Interesse. Ausgehend vom Diaconis-Sturmfels-Algorithmus werden nachfolgend neue algebraische Tests für kategoriale Daten entwickelt. Diese weisen als bedingte Tests in einer k-parametrigen Exponentialfamilie Optimalitätseigenschaften auf, vergleiche Witting (1985), Kapitel 3.3, und bilden eine wertvolle Ergänzung zu traditionellen Testmethoden. Die Simulationsergebnisse der vorliegenden Arbeit basieren auf der folgenden Vorgehensweise:

Gemäß Korollar 3.3.1 und Theorem 3.3.4 sei eine Gröbner-Basis $\mathscr{G} = \{g_1, \dots, g_L\}$ für den Vorschlag eines neuen Zustandes der Markov-Kette gegeben. Mit Wahrscheinlichkeit $1/L$ wird eine Bewegung $g_u \in \mathscr{G}$ ausgewählt. Unabhängig davon wird mit Wahrscheinlichkeit $1/2$ eine Richtung dieser Bewegung $\epsilon = \pm 1$ festgelegt. Die Kette befinde sich im Zustand $z \in \mathscr{Z}_t$. Dann wird im nächsten Schritt $z' = z + \epsilon g_u \in \mathscr{Z}_t$ mit Wahrscheinlichkeit $\alpha = \min\left(\frac{H(z')}{H(z)}, 1\right) = \min\left(\frac{\prod\limits_{x \in \mathscr{H}} z(x)!}{\prod\limits_{x \in \mathscr{H}} (z(x)+\epsilon g_u(x))!}, 1\right)$ angenommen, wobei H die oben angegebene hypergeometrische Dichte bezeichnet. Ist ein Zelleintrag einer potenziellen neuen Kontingenztafel kleiner als Null, so ist dieser Zustand nicht Element von $\mathscr{Z}_t$. $H(z')$ und folglich α sind dann ebenfalls Null und die Kette verweilt in ihrem Zustand.

Enthält der betrachtete Datensatz strukturelle Nullzellen, so umfasst $\mathscr{Z}_t$ alle Datensätze mit denselben Nullzellen sowie demselben Wert der suffizienten Statistik. Die Gröbner-Basis für das interessierende Modell muss entsprechend der Position der Nullzellen modifiziert werden; es handelt sich demnach immer um Sonderfälle. Nachfolgend wird davon ausgegangen, dass die betrachteten Datensätze keine strukturellen Nullzellen aufweisen. Aoki and Takemura (2005) sowie Rapallo (2006) haben für verschiedene log-lineare Modelle und unter Berücksichtigung solcher Zellen Gröbner-Basen hergeleitet.

Gebräuchliche Verfahren zur Überprüfung der Anpassungsgüte eines Modells bei kategorialen Daten sind der χ^2- sowie der Likelihood-Quotienten-Test G. Liegt beispielsweise eine $I_1 \times I_2$-Kontingenztafel vor, so sind die jeweiligen Teststatistiken gegeben durch $\chi^2 = \sum_{i_1=1}^{I_1}\sum_{i_2=1}^{I_2} \frac{(N_{i_1i_2}-n\hat{\pi}_{i_1i_2})^2}{n\hat{\pi}_{i_1i_2}}$ bzw. $G = 2\sum_{i_1=1}^{I_1}\sum_{i_2=1}^{I_2} N_{i_1i_2}\log(\frac{N_{i_1i_2}}{n\hat{\pi}_{i_1i_2}})$, wobei $\hat{\pi}_{i_1i_2}$ die Maximum-Likelihood-Schätzer für die Zellwahrscheinlichkeiten $\pi_{i_1i_2}$, $i_1 = 1,\ldots,I_1$, $i_2 = 1,\ldots,I_2$, für das zu testende Modell sind. Beide Teststatistiken sind asymptotisch χ^2-verteilt, die Anzahl der Freiheitsgrade richtet sich nach den Parametern des jeweiligen Modells, vgl. Good (1973). Diese Approximation beruht auf der Annahme eines „ausreichend großen“ Stichprobenumfangs. Gängige Faustregeln hierfür stammen von Cochran (1954) und Conover (1971). So empfiehlt Cochran für χ^2-Approximationen mit mehr als einem Freiheitsgrad die Bedingung, dass alle erwarteten Zelleinträge größer als 1 und mindestens 80 % davon größer gleich 5 sein sollten. Conover sieht die Approximation gerechtfertigt, wenn fast alle erwarteten Zelleinträge von gleicher Größenordnung und größer gleich Eins sind. Bei geringem Stichprobenumfang bieten exakte Verfahren eine Alternative zu den approximativen Methoden. Grundlage hierfür ist $\mathscr{X}_t$, die Menge aller möglichen Datensätze mit beobachteter suffizienter Statistik t. Wie bereits erwähnt ist die Verteilung auf $\mathscr{X}_t$ hypergeometrisch. Für eine exakte Testentscheidung werden daher alle Werte der hypergeometrischen Dichte aufsummiert, deren zugehöriger Wert der Teststatistik größer oder gleich dem Wert der beobachteten Teststatistik ist. Jedoch wird $|\mathscr{X}_t|$ schnell sehr groß, so dass die Berechnung des exakten p-Wertes oft zu aufwändig ist.

Simulationsbasierte Tests nach Diaconis und Sturmfels (1998) bieten eine wichtige Ergänzung zu den traditionellen Testmethoden. Nachfolgend wird das Vorgehen für die Entwicklung neuer algebraischer Tests beschrieben. Wie in Korollar 3.3.1 bezeichne l die Länge der generierten Markov-Kette. Um für die weiteren Betrachtungen sowohl Abhängigkeiten zwischen dem zugrunde liegenden Datensatz und der Markov-Kette als auch zwischen den einzelnen Zuständen zu verhindern, werden die ersten b Datensätze in der so genannten Einschwingphase („burn-in-Phase“) vernachlässigt und nur jede $s$te Tafel berücksichtigt, d.h. s entspricht der Schrittlänge. Für jede der übrigen $\lfloor\frac{l-b}{s}\rfloor$ Zustände wird anschließend z.B. die interessierende Teststatistik berechnet und ihre Verteilung unter der Nullhypothese simuliert. Somit ist z. B. der algebraische p-Wert gegeben durch $p = \frac{1}{\lfloor\frac{l-b}{s}\rfloor}\sum_{i=1}^{\lfloor\frac{l-b}{s}\rfloor} \mathbb{1}_{\{\chi^2_{\text{beob}}\geq\chi^2_i\}}(i)$, wobei χ^2_{beob} den beobachteten Wert der χ^2-Teststatistik und χ^2_i die entsprechenden Werte der simulierten Datensätze bezeichnen.

Die Parameter l, b und s beeinflussen das Simulationsergebnis. Im Fokus dieser Arbeit liegen Anwendungen und Erweiterungen des Diaconis-Sturmfels-Algorithmus. Daher werden im Folgenden verschiedene Parameterwerte und ihr Einfluss auf das Simulationsergebnis untersucht, umfangreiche Simulationsstudien hierzu werden aber nicht präsentiert. Für eine entsprechende theoretische Betrachtung der Konvergenz von Markov-Ketten sei z. B. auf Diaconis und Sturmfels (1998) sowie Diaconis und Saloff-Coste (1995) verwiesen.

Kapitel 4

Algebraische Testprozeduren und Konfidenzintervalle

In diesem Kapitel werden neue Einsatzmöglichkeiten des Diaconis-Sturmfels-Verfahrens entwickelt. In verschiedenen Anwendungsgebieten der Statistik wie z. B. der Epidemiologie oder den Sozialwissenschaften liegen häufig so genannte gepaarte Beobachtungen (matched pairs data) vor. Für solche Daten ist die Analyse struktureller Übereinstimmungen von besonderem Interesse. Der Fokus liegt zunächst auf Symmetrieuntersuchungen gemäß dem Bowker-Test sowie zweier Modifikationen dieses Tests. Dabei wird überprüft, ob der zugrunde liegende Datensatz der Hypothese perfekt symmetrischer Zellwahrscheinlichkeiten widerspricht. Da perfekte Symmetrie sehr restriktiv ist, werden oftmals andere vorhandene Symmetriestrukturen nicht entdeckt. „Gewichtete" Symmetriemodelle wie das bedingte Symmetriemodell (conditional symmetry, triangular symmetry, siehe Bishop et al. (1995) und McCullagh (1978)), das diagonale Symmetriemodell von Goodman (1979) (diagonal symmetry) sowie das ordinale Quasi-Symmetriemodell (ordinal quasi symmetry, siehe Agresti (1983)) sind sinnvolle andere Möglichkeiten. Gemäß dem Verfahren von Diaconis und Sturmfels (1998) werden für diese Modelle neue algebraische Tests entwickelt. In einer Simulationsstudie werden anschließend die Eigenschaften der neuen Verfahren untersucht und mit den entsprechenden approximativen und exakten Methoden verglichen. Für die Analyse gepaarter Beobachtungen sind das Quasi-Symmetrie- und das Quasi-Unabhängigkeitsmodell bzw. Cohen's κ ebenfalls interessant, sie sind jedoch schwer zu interpretieren (siehe z. B. Agresti (2002), Kapitel 10, Chichetti und Feinstein (1990) sowie Feinstein und Cichetti (1990)). Rapallo (2005) betrachtet für diese

Modelle algebraische Testprozeduren gemäß dem Diaconis-Sturmfels-Algorithmus.

Oftmals ist für die Analyse kategorialer Daten nicht nur von Bedeutung, ob ein Zusammenhang zwischen den erhobenen Merkmalen besteht, sondern es interessiert zudem die Stärke dieses Zusammenhangs. So ist beispielsweise die Identifizierung so genannter Risikofaktoren eine wichtige Zielsetzung epidemiologischer Studien. Das relative Risiko und das Odds Ratio sind gängige Maße für die Beurteilung der Beziehung zwischen Krankheit und Exposition mit einem potentiellen Risikofaktor, vgl. Agresti (2002). Obwohl die Interpretation des relativen Risikos besonders intuitiv ist, wird es in der vorliegenden Arbeit nicht weiter erörtert, da es je nach Studientyp beliebig manipulierbare Ergebnisse liefern kann, siehe Agresti (1996), Kapitel 2.3.4, für ein Beispiel. Für 2×2-Kontingenztafeln wird nachfolgend das Odds Ratio betrachtet, siehe Kreienbrock und Schach (1995). Auf der Grundlage des in Kapitel 3 vorgestellten Algorithmus von Diaconis und Sturmfels wird ein algebraisches Konfidenzintervall für das Odds Ratio entwickelt. Die vorgestellte Vorgehensweise zur Bestimmung algebraischer Konfidenzintervalle kann auf weitere Parameter wie das Relative Risiko bei 2×2-Tafeln sowie das Odds Ratio bei $2 \times 2 \times K$-Datensätzen übertragen werden. In einer anschließenden Simulationsstudie werden die üblicherweise verwendeten approximativen und exakten Konfidenzintervalle mit den neuen Konfidenzintervallen verglichen.

4.1 Symmetriemodelle

Häufig liegen der statistischen Analyse kategorialer Daten so genannte gepaarte Beobachtungen zugrunde; d. h. der Datensatz besteht aus zwei voneinander abhängigen Stichproben, wobei jede Realisation der einen Stichprobe zu genau einer Beobachtung aus der anderen Stichprobe gehört. Eine solche Situation liegt z. B. in Gutachterverlässlichkeitsstudien vor, wenn zwei Gutachter, X_1 und X_2, unabhängig voneinander n Objekte (oder Subjekte) in I vorab definierte Kategorien einteilen. Die Beurteilungen der Gutachter werden in einer $I \times I$-Kontingenztafel zusammengefasst, $I \geq 2$, und es interessiert, ob eine systematische Übereinstimmung der Beurteilungen statistisch nachgewiesen werden kann. Es gibt verschiedene Vorschläge zur Analyse solcher Daten, vgl. Landis und Koch (1975), Bishop et al. (1995) und Agresti (1992, 2002).

Die einfachste Form der Symmetrie, *perfekte Symmetrie* (S), liegt vor, wenn die Wahrscheinlichkeiten für die Realisationen (i_1, i_2) identisch sind mit den Wahrscheinlichkeiten für (i_2, i_1), d. h. getestet wird $H_0 : \pi_{i_1 i_2} = \pi_{i_2 i_1}$ für alle i_1, $i_2 = 1, \ldots, I$, gegen $H_1 : \pi_{i_1 i_2} \neq \pi_{i_2 i_1}$ für mindestens ein Paar (i_1, i_2). Ist dies erfüllt, so gilt gleichzeitig die Homogenität der Randverteilungen. Die erwarteten Zellhäufigkeiten $m_{i_1 i_2}$ werden durch

$$\log(m_{i_1 i_2}) = \lambda + \lambda_{i_1}^{X_1} + \lambda_{i_2}^{X_1} + \lambda_{i_1 i_2}^{X_1 X_2}$$

repräsentiert, und es ist $\lambda_{i_1 i_2}^{X_1 X_2} = \lambda_{i_2 i_1}^{X_1 X_2}$. Die Eindeutigkeit der Parameter sei hier durch die Restriktionen $\sum\limits_{i_1} \lambda_{i_1}^{X_1} = \sum\limits_{i_2} \lambda_{i_2}^{X_2} = \sum\limits_{i_1} \lambda_{i_1 i_2}^{X_1 X_2} = 0$ gewährleistet. Der ML-Schätzer der erwarteten Zellhäufigkeiten im perfekten Symmetriemodell S ist gegeben durch $\hat{m}_{i_1 i_2}^{\mathrm{S}} = \frac{N_{i_1 i_2} + N_{i_2 i_1}}{2}$. Basierend auf dem χ^2-Anpassungstest hat McNemar (1947) einen Symmetrietest für 2×2-Kontingenztafeln entwickelt, der von Bowker (1948) für $I \times I$-Tafeln, $I \geq 2$, verallgemeinert wurde. Die zugehörige Teststatistik

$$\chi^2_{\text{Bowker}} = \sum_{i_1=1}^{I-1} \sum_{i_2=i_1+1}^{I} \frac{(N_{i_1 i_2} - N_{i_2 i_1})^2}{N_{i_1 i_2} + N_{i_2 i_1}};$$

ist unter Symmetrieannahme approximativ χ^2-verteilt mit $\frac{1}{2} I(I-1)$ Freiheitsgraden. Zur Verbesserung der Approximation schlägt Edwards (1948) eine stetigkeitskorrigierte Version des McNemar-Tests vor, die auf den Bowker-Test übertragen werden kann. Die resultierende Prüfgröße ist

$$\chi^2_{\text{Korr}} = \sum_{i_1=1}^{I-1} \sum_{i_2=i_1+1}^{I} \frac{(|N_{i_1 i_2} - N_{i_2 i_1}| - 1)^2}{N_{i_1 i_2} + N_{i_2 i_1}}.$$

Modifikationen dieses Typs werden allerdings in der Literatur kontrovers diskutiert, siehe z. B. Conover (1974), Grizzle (1967), Mantel und Greenhouse (1968) sowie Plackett (1964). May und Johnson (2001) entwickeln einen modifizierten Wald-Test als Alternative zum Bowker-Test. Dazu repräsentiere der Vektor $\delta = (\delta_{i_1 i_2})_{i_1, i_2 = 1, \ldots, I,\, i_1 < i_2}$ die Differenzen der Anteile unterschiedlicher Kategoriezuweisungen, d. h. $\delta_{i_1 i_2} := \pi_{i_1 i_2} - \pi_{i_2 i_1}$. Durch die Verwendung der modifizierten Kovarianz

$$V(\delta) = \frac{1}{n} \begin{pmatrix} \lambda_{12} - \delta_{12}^2 & 0 & \cdots & 0 \\ 0 & \lambda_{13} - \delta_{13}^2 & \cdots & \vdots \\ \vdots & \cdots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_{(I-1)I} - \delta_{(I-1)I}^2 \end{pmatrix},$$

mit $\lambda_{i_1i_2} = \pi_{i_1i_2} + \pi_{i_2i_1}$, $i_1 < i_2$, ist die Wald-Teststatistik darstellbar als

$$\begin{aligned}\chi^2_{\text{mod. Wald}} &= \hat{\delta}'\hat{V}(\hat{\delta})^{-1}\hat{\delta} \\ &= \sum_{i_1=1}^{I-1}\sum_{i_2=i_1+1}^{I} \frac{n\cdot(N_{i_1i_2} - N_{i_2i_1})^2}{n\cdot(N_{i_1i_2} + N_{i_2i_1}) - (N_{i_1i_2} - N_{i_2i_1})^2}.\end{aligned}$$

Die Testentscheidungen für χ^2_{Korr} und $\chi^2_{\text{mod. Wald}}$ erfolgen, wie beim Bowker-Test, gemäß der $\chi^2_{\frac{1}{2}I(I-1)}$-Verteilung.

Das perfekte Symmetriemodell ist sehr restriktiv und wird in der Regel nicht von den Daten unterstützt. Caussinus (1965) schlägt alternativ das Quasi-Symmetriemodell (QS) vor, in dem symmetrische Strukturen ausschließlich anhand der Interaktionsparameter $\lambda_{i_1i_2}$ im log-linearen Modell beschrieben werden und somit heterogene Randverteilungen zugelassen sind. Dieses Modell umfasst verschiedene andere Modelle wie z. B. das S-Modell als Spezialfälle. Ein wesentlicher Nachteil dieses Modells ist die schlechte Interpretierbarkeit. Im Folgenden werden verschiedene weitere intuitive Symmetriemodelle vorgestellt.

In dem *bedingten Symmetriemodell* (CS, conditional oder auch triangular symmetry) von McCullagh (1978) wird den gegenüberliegenden Zellwahrscheinlichkeiten eine konstant proportionale Beziehung unterstellt, d. h. $\pi_{i_1i_2} = c\pi_{i_2i_1}$ für alle $i_1 > i_2$, $i_1, i_2 = 1, \ldots, I$, $c \in \mathbb{R}^+$. Äquivalent dazu können die Zellwahrscheinlichkeiten dargestellt werden als

$$\pi_{i_1i_2} = \begin{cases} \tau\pi^S_{i_1i_2}, & i_1 > i_2 \\ \pi^S_{i_1i_2}, & i_1 = i_2 \\ (2-\tau)\pi^S_{i_1i_2}, & i_1 < i_2 \end{cases}$$

mit $\tau \in (0,2)$ und $c = \frac{\tau}{2-\tau}$. Dabei bezeichnet $\pi^S_{i_1i_2}$ die Zellwahrscheinlichkeit gemäß Modell S. Die erwarteten Zellhäufigkeiten $m_{i_1i_2}$ können durch

$$\log(m_{i_1i_2}) = \lambda + \lambda^{X_1}_{i_1} + \lambda^{X_2}_{i_2} + \lambda^{X_1X_2}_{i_1i_2} + \log(\tau)\mathbb{1}_{(i_1>i_2)}((i_1,i_2)) + \log(2-\tau)\mathbb{1}_{(i_1<i_2)}((i_1,i_2))$$

beschrieben werden, wobei $\lambda^{X_1X_2}_{i_1i_2} = \lambda^{X_1X_2}_{i_2i_1}$ gilt. Für die Eindeutigkeit der Parameter werden dieselben Restriktionen wie für das perfekte Symmetriemodell gefordert. Unter Verwendung des ML-Schätzers $\hat{\tau} = 2\cdot\frac{\sum_{i_1>i_2} N_{i_1i_2}}{\sum_{i_1\neq i_2} N_{i_1i_2}}$ für τ wird die erwartete Tafel geschätzt. Basierend auf üblichen Anpassungstests wie dem χ^2- oder dem Likelihood-Quotienten-Test G kann anschließend die Proportionalitätshypothese überprüft werden. Im Vergleich zum perfekten Symmetriemodell weist das CS-Modell einen zusätzlichen Parameter auf. Die Anzahl der Freiheitsgrade der χ^2-Verteilung beträgt daher hier $\frac{1}{2}(I+1)(I-2)$.

Liegen den betrachteten Datensätzen ordinale Merkmale zugrunde, so bieten sich Symmetriemodelle an, die diese zusätzliche Information beachten. Goodman (1979) beachtet für das *diagonale Symmetriemodell* (DS, diagonal symmetry) zusätzlich den Abstand zwischen den unterschiedlichen Kategorien. Die konstante Proportionalitätsbeziehung zwischen der oberen und der unteren Dreiecksmatrix des bedingten Symmetriemodells wird durch Proportionalitätsfaktoren $\alpha_k \in (0,2)$ ersetzt, die von der Entfernung $k := |i_1 - i_2| \in \{1,\ldots,I-1\}$ der betrachteten Kategorien abhängen, d. h.

$$\pi_{i_1 i_2} = \begin{cases} \alpha_k \pi^S_{i_1 i_2}, & i_1 > i_2 \\ \pi^S_{i_1 i_2}, & i_1 = i_2 \\ (2-\alpha_k)\pi^S_{i_1 i_2}, & i_1 < i_2 \end{cases} .$$

Das CS-Modell ist also ein Spezialfall des DS-Modells, wenn $\alpha_k = \tau$ für alle k gilt. Eine äquivalente Darstellung dieses Modells ist gegeben durch

$$\log(m_{i_1 i_2}) = \lambda + \lambda^{X_1}_{i_1} + \lambda^{X_2}_{i_2} + \lambda^{X_1 X_2}_{i_1 i_2} + \log(\alpha_k)\mathbb{1}_{(i_1 > i_2, |i_1 - i_2| = k)}((i_1, i_2)) + \log(2-\alpha_k)\mathbb{1}_{(i_1 < i_2, |i_1 - i_2| = k)}((i_1, i_2)),$$

und es ist $\lambda^{X_1 X_2}_{i_1 i_2} = \lambda^{X_1 X_2}_{i_2 i_1}$. Die Eindeutigkeit der Parameter wird wiederum entsprechend den Restriktionen für das S-Modell sichergestellt. Der ML-Schätzer für die Proportionalitätsfaktoren α_k ist $\hat{\alpha}_k = 2 \cdot \frac{\sum_{i_1 > i_2, |i_1 - i_2| = k} N_{i_1 i_2}}{\sum_{i_1 \neq i_2 | i_1 - i_2| = k} N_{i_1 i_2}}$. Damit können nachfolgend die jeweiligen Teststatistikwerte des χ^2-Anpassungstests sowie des Likelihood-Quotienten-Tests G bestimmt werden. Entsprechend der Parametrisierung des DS-Modells reduziert sich die Anzahl der Freiheitsgrade der zugrunde liegenden χ^2-Verteilung auf $\frac{1}{2}(I-1)(I-2)$.

In dem DS-Modell werden keinerlei Bedingungen an die Größe der einzelnen Proportionalitätsfaktoren α_k gestellt. Liegen ordinale Einteilungen zugrunde, wird häufig eine monotone Entwicklung der Parameter α_k, $k = |i_1 - i_2| \in \{1,\ldots,I-1\}$, erwartet. Agresti (1983) berücksichtigt diese zusätzliche Information und schlägt für das *ordinale*

Quasi-Symmetriemodell (OQS, ordinal quasi-symmetry) einen log-linearen Zusammenhang zwischen den Proportionalitätsfaktoren α_k und der jeweiligen Distanz k vor, d. h. es gilt $\alpha_k = \alpha^k$, $k \in \{1, \ldots, I-1\}$. Das OQS-Modell wird somit dargestellt durch $\pi_{i_1 i_2} = \pi_{i_2 i_1}\alpha^{i_1 - i_2}$, $i_1 > i_2$, und es ist ein weiterer Spezialfall des DS-Modells, vgl. Agresti (1983). Im Folgenden sei eine Menge $\{u_1, u_2, \ldots, u_I\}$ mit $u_1 \leq u_2 \leq \ldots \leq u_I$ sowie $u_1 < u_I$ gegeben und repräsentiere die Abstände der Kategorien anstelle der Indizes. Im Falle von äquidistanten u_{i_2} ist

$$\log(m_{i_1 i_2}) = \lambda + \lambda_{i_1}^{X_1} + \lambda_{i_2}^{X_1} + \lambda_{i_1 i_2}^{X_1 X_2} + \beta u_{i_2}$$

mit $\lambda_{i_1 i_2}^{X_1 X_2} = \lambda_{i_2 i_1}^{X_1 X_2}$ und den Restriktionen wie für das Symmetriemodell eine äquivalente Darstellung des ordinalen Quasi-Symmetriemodells. Für $\{u_{i_2} = i_2\}$ kann gezeigt werden, dass $\beta = -\log(\alpha)$ ist und die Zellwahrscheinlichkeiten können als

$$\pi_{i_1 i_2} = \frac{2\exp(-\beta u_{i_1})}{\exp(-\beta u_{i_1})\exp(-\beta u_{i_2})}\pi_{i_1 i_2}^{S}, \quad i_1, i_2 = 1, \ldots, I,\ \beta \in I\!R,$$

beschrieben werden, vgl. Kateri und Agresti (2007). Häufig wird o. B. d. A. zusätzlich $\sum\limits_{i_2} u_{i_2} = 0$ und $\sum\limits_{i_2} u_{i_2}^2 = 1$ gefordert, um eine verbesserte Interpretierbarkeit von β zu erreichen. Für $\beta = 0$ reduziert sich das OQS-Modell zu dem Symmetriemodell S. Andererseits ist das OQS-Modell ein Spezialfall des Quasi-Symmetriemodells mit $\lambda_{i_2}^{X_2} = \lambda_{i_2}^{X_1} + \beta u_{i_2}$. Die Eignung dieses Modells wird wiederum mit dem χ^2-Anpassungstest bzw. dem Likelihood-Quotienten-Test beurteilt; die zugrunde liegende χ^2-Verteilung besitzt $\frac{1}{2}(I+1)(I-2)$ Freiheitsgrade.

4.1.1 Neue Tests für Symmetriemodelle

Der Diaconis-Sturmfels-Algorithmus ermöglicht die Generierung einer Stichprobe aus der bedingten Verteilung einer diskreten Exponentialfamilie mit beobachteter suffizienter Statistik t, siehe Kapitel 3.3. Unter Berücksichtigung der Multinomial-Erhebungstechnik sowie der vier interessierenden Symmetriemodelle (S, CS, DS, OQS) erfüllt die jeweilige gemeinsame Dichte der Zelleinträge die Forderungen an eine diskrete Exponentialfamilie. Grundlage weiterer Betrachtungen sind ausschließlich Datensätze ohne strukturelle Nullzellen. Die suffiziente Statistik für die Parameter im S-Modell kann dargestellt werden als

$$T^{(\mathrm{S})}((i_1,i_2)) = (N_{i_1i_2},\ i_1=i_2=1,\ldots,I;\ (N_{i_1i_2}+N_{i_2i_1}),\ i_1,i_2=1,\ldots,I,\ i_1<i_2)'.$$

Anders als in Krampe und Kuhnt (2007) ist hier N_{II} zusätzlich in $T^{(S)}$ enthalten. Diese Modifizierung ist hilfreich für das später betrachtete OQS-Modell, ändert aber aufgrund der Redundanz von N_{II} nicht die Ergebnisse. Eine Erweiterung von $T^{(S)}((i_1,i_2))$ zu

$$\widetilde{T}^{(\mathrm{S})}((i_1,i_2)) = (N_{i_1i_1},\ i_1=i_2=1,\ldots,I;\ N_{i_1i_2}+N_{i_2i_1},$$

$$N_{i_2i_1}+N_{i_1i_2}, i_1,i_2=1,\ldots,I,\ i_1<i_2)'$$

erweist sich als zweckmäßig. Die suffizienten Statistiken für die Parameter des CS-, DS- und OQS-Modells sind entsprechend

$$\begin{aligned}
\widetilde{T}^{(\mathrm{CS})}((i_1,i_2)) &= (\widetilde{T}^{(\mathrm{S})}((i_1,i_2))',\ \sum_{i_1>i_2} N_{i_1i_2})',\\
\widetilde{T}^{(\mathrm{DS})}((i_1,i_2)) &= (\widetilde{T}^{(\mathrm{S})}((i_1,i_2))',\ \sum_{i_1-i_2=1} N_{i_1i_2},\ldots,\sum_{i_1-i_2=I-1} N_{i_1i_2})',\\
\widetilde{T}^{(\mathrm{OQS})}((i_1,i_2)) &= (\widetilde{T}^{(\mathrm{S})}((i_1,i_2))',\ \sum_{i_1=1}^{I} u_{i_1}N_{i_1+})'.
\end{aligned}$$

Gemäß Kapitel 3.3 ist aufgrund der getroffenen Symmetrieannahme die Menge $\mathscr{Z}_t := \{z : \mathscr{H} \to I\!N \mid \sum_{x\in\mathscr{H}} z(x)T^*(x) = t\}$ aller möglichen Datensätze mit beobachteter suffizienter Statistik und insbesondere T^* zu bestimmen. Die Anzahl der Elemente von T^* entspricht dabei der Anzahl der Elemente der zugrunde liegenden suffizienten Statistik; für das S-Modell beträgt diese $I + 2\sum_{v=1}^{I-1}(I-v) = I^2$. Bei bedingter und ordinaler Quasi-Symmetrie haben die entsprechenden T^* jeweils I^2+1 Einträge. Wird das diagonale Symmetriemodell betrachtet, so ist die Länge von T^* gegeben durch $(I-1)(I+2)+1$. Analog zu den suffizienten Statistiken wird im Weiteren das zugrunde liegende T^* mit dem dazu gehörenden Modell identifiziert.

Jedes $T^{*(S)}$ kann in I Teile aufgespalten werden; der erste Teil besteht aus I Einträgen und repräsentiert die Diagonalelemente $N_{i_1i_1}$, $i_1 \in \{1,\ldots,I\}$; insbesondere ist der $i_1$te Eintrag von $T^{*(S)}((i_1,i_1))$ gleich Eins, alle anderen sind Null. Für $i_1 \neq i_2$ sind die ersten I Einträge von $T^{*(\mathrm{S})}((i_1,i_2))$ Null; die übrigen $I-1$ Teile von $T^{*(\mathrm{S})}$ haben die Länge

$2(I-v)$, $v = 1, \ldots, I-1$, und repräsentieren durch zwei Einsen an der jeweiligen Stelle die $N_{i_1i_2} + N_{i_2i_1}$ und $N_{i_2i_1} + N_{i_1i_2}$, alle weiteren Einträge sind Null. Aufgrund der Symmetrieannahme gilt die Gleichheit von $T^{*(\mathrm{S})}((i_1, i_2))$ und $T^{*(\mathrm{S})}((i_2, i_1))$.

Beispiel 4.1.1

Als Erläuterung für die Zuordnung $T^{*(S)}((i_1, i_2))$ liege eine 4×4-Kontingenztafel vor. Die erweiterte suffiziente Statistik für die Parameter im S-Modell ist ein Vektor der Länge 16. Dieser ist gegeben als $\widetilde{T}^{(S)}((i_1, i_2)) = (N_{11}, \ldots, N_{44}, N_{12} + N_{21}, N_{21} + N_{12}, N_{13} + N_{31}, N_{31} + N_{13} \ldots, N_{34} + N_{43}, N_{43} + N_{34})'$. Damit hat $T^{*(S)}((i_1, i_2))$ ebenso 16 Einträge und es ist

$$
\begin{aligned}
T^{*(S)}((1,1)) &= (1,0,0,0,\ \ 0,0,0,0,0,0,\ \ 0,0,0,0,\ \ 0,0)',\\
T^{*(S)}((2,2)) &= (0,1,0,0,\ \ 0,0,0,0,0,0,\ \ 0,0,0,0,\ \ 0,0)',\\
T^{*(S)}((3,3)) &= (0,0,1,0,\ \ 0,0,0,0,0,0,\ \ 0,0,0,0,\ \ 0,0)',\\
T^{*(S)}((4,4)) &= (0,0,0,1,\ \ 0,0,0,0,0,0,\ \ 0,0,0,0,\ \ 0,0)',\\
T^{*(S)}((1,2)) &= (0,0,0,0,\ \ 1,1,0,0,0,0,\ \ 0,0,0,0,\ \ 0,0)',\\
T^{*(S)}((2,1)) &= (0,0,0,0,\ \ 1,1,0,0,0,0,\ \ 0,0,0,0,\ \ 0,0)',\\
T^{*(S)}((1,3)) &= (0,0,0,0,\ \ 0,0,1,1,0,0,\ \ 0,0,0,0,\ \ 0,0)',\\
T^{*(S)}((3,1)) &= (0,0,0,0,\ \ 0,0,1,1,0,0,\ \ 0,0,0,0,\ \ 0,0)',\\
&\ \ \vdots\\
T^{*(S)}((4,3)) &= (\underbrace{0,0,0,0,}_{\substack{\textit{entspricht}\\ N_{ii}, i=1,\ldots,4}}\ \ \underbrace{0,0,0,0,0,0,}_{\substack{(N_{1i_2}+N_{i_21}),\\ (N_{i_21}+N_{1i_2}),\ i_2>1}}\ \ \underbrace{0,0,0,0,}_{\substack{(N_{2i_2}+N_{i_22}),\\ (N_{i_22}+N_{2i_2}),\ i_2>2}}\ \ \underbrace{1,1}_{\substack{(N_{34}+N_{43}),\\ (N_{43}+N_{34})}}\ \ )'.
\end{aligned}
$$

Für das bedingte, das diagonale sowie das ordinale Quasi-Symmetriemodell werden die jeweiligen T^* aufbauend auf $T^{*(\mathrm{S})}$ dargestellt:

$$\begin{aligned}
T^{*(\mathrm{CS})}((i_1,i_2)) &= (T^{*(\mathrm{S})}((i_1,i_2))', \mathbb{1}_{\{i_1>i_2\}}((i_1,i_2)))',\\
T^{*(\mathrm{DS})}((i_1,i_2)) &= (T^{*(\mathrm{S})}((i_1,i_2))', \mathbb{1}_{\{i_1-i_2=1\}}((i_1,i_2)),\ldots,\mathbb{1}_{\{i_1-i_2=I-1\}}((i_1,i_2)))',\\
T^{*(\mathrm{OQS})}((i_1,i_2)) &= (T^{*(\mathrm{S})}((i_1,i_2))', \textstyle\sum_{l=1}^{I} \mathbb{1}_{\{l=i_1\}}(((i_1,i_2))u_l)'.
\end{aligned}$$

Wie in Theorem 3.3.4 erläutert, lassen sich diese T^* für die Bestimmung der Vorschlagsdichte im Metropolis-Hastings-Algorithmus nutzen.

Beispiel 4.1.2

Gegeben sei eine 4×4-Kontingenztafel. Gemäß Theorem 3.3.4 werden beispielsweise mit **`CoCoA`** *Gröbner-Basen der folgenden Hilfsideale berechnet:*

Für das perfekte Symmetriemodell:

$\mathscr{I}_{hilf}^{(S)} = \{x_{11} - N_{11},\ \ldots,\ x_{44} - N_{44},$

$x_{12} - (N_{12} + N_{21})(N_{21} + N_{12}), x_{21} - (N_{12} + N_{21})(N_{21} + N_{12}),\ \ldots,$

$x_{34} - (N_{34} + N_{43})(N_{43} + N_{34}),\ x_{43} - (N_{34} + N_{43})(N_{43} + N_{34})\}$;

Für das bedingte Symmetriemodell:

$\mathscr{I}_{hilf}^{(CS)} = \{x_{11} - N_{11},\ \ldots,\ x_{44} - N_{44},$

$x_{12} - (N_{12} + N_{21})(N_{21} + N_{12}),\ x_{21} - (N_{12} + N_{21})(N_{21} + N_{12}) \sum\limits_{i_1>i_2} N_{i_1 i_2},\ \ldots,$

$x_{34} - (N_{34} + N_{43})(N_{43} + N_{34}),\ x_{43} - (N_{34} + N_{43})(N_{43} + N_{34}) \sum\limits_{i_1>i_2} N_{i_1 i_2}\}$;

Für das diagonale Symmetriemodell:

$\mathscr{I}_{hilf}^{(DS)} = \{x_{11} - N_{11},\ \ldots,\ x_{44} - N_{44},$

$x_{12} - (N_{12} + N_{21})(N_{21} + N_{12}),\ x_{21} - (N_{12} + N_{21})(N_{21} + N_{12}) \sum\limits_{i_1-i_2=1} N_{i_1 i_2},$

$x_{13} - (N_{13} + N_{31})(N_{31} + N_{13}),\ x_{31} - (N_{13} + N_{31})(N_{31} + N_{13}) \sum\limits_{i_1-i_2=2} N_{i_1 i_2},\ \ldots,$

$x_{34} - (N_{34} + N_{43})(N_{43} + N_{34}),\ x_{43} - (N_{34} + N_{43})(N_{43} + N_{34}) \sum\limits_{i_1-i_2=1} N_{i_1 i_2}\}$;

Für das ordinale Quasi-Symmetriemodell:

$\mathscr{I}_{hilf}^{(OQS)} = \{x_{11} - N_{11},\ \ldots,\ x_{44} - N_{44},$

$$x_{12} - (N_{12}+N_{21})(N_{21}+N_{12})(\sum_{i_1=1}^{I} u_{i_1} N_{i_1+}),\ x_{21} - (N_{12}+N_{21})(N_{21}+N_{12})(\sum_{i_1=1}^{I} u_{i_1} N_{i_1+})^2,$$

$$x_{13} - (N_{13}+N_{31})(N_{31}+N_{13})(\sum_{i_1=1}^{I} u_{i_1} N_{i_1+}),\ x_{31} - (N_{13}+N_{31})(N_{31}+N_{13})(\sum_{i_1=1}^{I} u_{i_1} N_{i_1+})^3,$$

$$\ldots$$

$$x_{34} - (N_{34}+N_{43})(N_{43}+N_{34})(\sum_{i_1=1}^{I} u_{i_1} N_{i_1+})^3,\ x_{43} - (N_{34}+N_{43})(N_{43}+N_{34})(\sum_{i_1=1}^{I} u_{i_1} N_{i_1+})^4\}.$$

Die gesuchten Gröbner-Basen für die betrachteten Modelle enthalten alle Basispolynome der jeweiligen Hilfsgröbnerbasis $\mathscr{G}_{hilf}$, die ausschließlich Terme aus dem Stichprobenraum $\mathscr{H}$ enthalten.

Für die Symmetriemodelle sind insbesondere die Zellen außerhalb der Hauptdiagonalen, d. h. Zellen, in denen die zugrunde liegenden Objekte bzw. Subjekte unterschiedlich bewertet werden, von Interesse. Die Gröbner-Basis für das S-Modell besteht aus $\frac{1}{2}I(I-1)$ Polynomen, die ausschließlich Bewegungen zwischen Zelleinträgen mit entgegengesetzter Einteilung, d. h. $N_{i_1 i_2}$ und $N_{i_2 i_2}$, $i_1 \neq i_2$, zulassen. Die konstante Proportionalitätsbeziehung in dem bedingten Symmetriemodell erfordert eine Gröbner-Basis, die symmetrische Strukturen berücksichtigt sowie die Summe der Zelleinträge unter- und oberhalb der Hauptdiagonalen konstant hält. Insgesamt besteht diese Gröbner-Basis aus $\binom{\frac{1}{2}I(I-1)}{2}$ Polynomen. Für das diagonale Symmetriemodell wird zusätzlich die Entfernung $k = |i_1 - i_2|$ der Kategorien berücksichtigt. Damit ist die Gröbner-Basis des DS-Modells eine Teilmenge der Gröbner-Basis des CS-Modells.

Bemerkung 4.1.3

Zugrunde liege eine $I \times I$-Kontingenztafel. Die reduzierte Gröbner-Basis bezüglich der gradlexikographischen Monomordnung ist für das

- *S-Modell*
  $\mathscr{G}^{(S)} = \{x_{i_1 i_2} - x_{i_2 i_1}, 1 \leq i_1 < i_2 \leq I\},$
- *CS-Modell*
  $\mathscr{G}^{(CS)} = \{x_{i_1 i_2} x_{i'_1 i'_2} - x_{i_2 i_1} x_{i'_2 i'_1},\ 1 \leq i_1 < i_2 \leq I,\ \ 1 \leq i'_1 < i'_2 \leq I\},$

- *DS-Modell*
  $\mathscr{G}^{(DS)} = \{x_{i_1 i_2} x_{i'_2 i'_1} - x_{i_2 i_1} x_{i'_1 i'_2},\ |i_2 - i_1| = |i'_2 - i'_1| = k,\ i_1, i_2, i'_1, i'_2 \in \{1, \ldots, I\}\}$,

- *OQS-Modell*
  Die Gröbner-Basis für das ordinale Quasi-Symmetriemodell besteht aus verschiedenen „Bewegungstypen". Beispielsweise weist die Gröbner-Basis für eine 4×4-*Kontingenztafel insgesamt 45 Basispolynome auf, die sich wie folgt zusammensetzen*

  · *vier Basispolynome des Typs* $x_{12}x_{32} - x_{21}x_{23}$,

  · *sechs Basispolynome des Typs* $x_{13}x_{32}^2 - x_{23}^2 x_{31}$,

  · *zwölf Basispolynome des Typs* $x_{12}x_{23}x_{31} - x_{13}x_{21}x_{32}$,

  · *drei Basispolynome des Typs* $x_{14}x_{43}^3 - x_{34}^3 x_{41}$,

  · *zwölf Basispolynome des Typs* $x_{14}x_{32}x_{43}^2 - x_{23}x_{34}^2 x_{41}$,

  · *vier Basispolynome des Typs* $x_{12}x_{23}x_{34}x_{41} - x_{14}x_{21}x_{32}x_{43}$,

  · *zwei Basispolynome des Typs* $x_{14}^2 x_{42}^3 - x_{24}^3 x_{41}^2$,

  · *zwei Basispolynome des Typs* $x_{14}x_{31}x_{42}^2 - x_{41}x_{13}x_{24}^2$.

Die oben angegebenen Gröbner-Basen werden aufgrund der Eliminationstheorie bestimmt, vergleiche Kapitel 3.3. Die Überprüfung dieser Gröbner-Basen erfolgt daher anhand der Gröbner-Basis des jeweiligen Hilfsideals sowie der in Definition 3.2.6 angegebenen Eigenschaften einer Gröbner-Basis. Für die Untersuchung der verschiedenen Symmetriestrukturen werden entsprechend den Ausführungen in Kapitel 3.3 diese Gröbner-Basen für die Bestimmung neuer algebraischer Testmethoden verwendet. Aufgrund der Konstruktion einer Gröbner-Basis bzw. des dazugehörenden Ideals werden im Metropolis-Hastings-Algorithmus ausschließlich Datensätze generiert, die dieselbe realisierte suffiziente Statistik t aufweisen wie die beobachtete Kontingenztafel. Für ausgewählte Tafeln wird der Wert der interessierenden Teststatistik berechnet und somit die bedingte Verteilung der Teststatistik bei gegebenem t simuliert.

4.1.2 Simulationsstudie

Im Folgenden werden 4×4-Kontingenztafeln betrachtet. Basierend auf den Tests auf perfekte Symmetrie werden zunächst sinnvolle Werte für die „Parameter" der Markov-Kette, d. h. der Kettenlänge, Schrittlänge sowie der Einschwingphase, festgelegt. Im zweiten Teil der Simulationsstudie werden die Eigenschaften der approximativen, exakten und algebraischen Tests für das S-, CS-, DS- und OQS-Modell vergleichend analysiert.

Für die Simulationsstudie werden Modellwahrscheinlichkeiten gewählt, die einer realen Datensituation entsprechen. Ein Großteil der Beobachtungen ist typischerweise auf der Hauptdiagonalen zu finden; diese repräsentiert beispielsweise diejenigen Objekte, die von zwei Gutachtern übereinstimmend beurteilt wurden. Die gewählten Modellwahrscheinlichkeiten können Tabelle 4.1 (i) entnommen werden. Unter Beachtung der Multinomial-Erhebungstechnik müssen alle $\widehat{\pi}_{i_1 i_2} > 0$, $i_1 \neq i_2$, $i_1, i_2 = 1, \ldots, 4$, sein. Ist dies nicht erfüllt, so liegen die geschätzten Erwartungswerte nicht im zulässigen Parameterraum. Derartige Tafeln werden hier von der Untersuchung ausgeschlossen. Für eine entsprechende Simulation von 100 geeigneten Datensätzen aus dem dem S-Modell wurden bei $n = 25$ insgesamt 182 Tafeln zufällig generiert.

$$
\text{(i)} \begin{bmatrix} 0{,}0943 & 0{,}0660 & 0{,}0377 & 0{,}0283 \\ 0{,}0660 & 0{,}0943 & 0{,}0755 & 0{,}0472 \\ 0{,}0377 & 0{,}0755 & 0{,}0943 & 0{,}0566 \\ 0{,}0283 & 0{,}0472 & 0{,}0566 & 0{,}0943 \end{bmatrix}, \quad \text{(ii)} \begin{bmatrix} 0{,}0769 & 0{,}0192 & 0{,}0385 & 0{,}0192 \\ 0{,}0577 & 0{,}0769 & 0{,}0288 & 0{,}0385 \\ 0{,}1154 & 0{,}0865 & 0{,}0769 & 0{,}0288 \\ 0{,}0577 & 0{,}1154 & 0{,}0865 & 0{,}0769 \end{bmatrix},
$$

$$
\text{(iii)} \begin{bmatrix} 0{,}0990 & 0{,}0792 & 0{,}0792 & 0{,}0099 \\ 0{,}1188 & 0{,}0990 & 0{,}0198 & 0{,}0594 \\ 0{,}0396 & 0{,}0297 & 0{,}0990 & 0{,}0396 \\ 0{,}0396 & 0{,}0297 & 0{,}0594 & 0{,}0990 \end{bmatrix}, \text{(iv)} \begin{bmatrix} 0{,}1250 & 0{,}0656 & 0{,}0515 & 0{,}0180 \\ 0{,}0594 & 0{,}1250 & 0{,}0492 & 0{,}0258 \\ 0{,}0422 & 0{,}0445 & 0{,}1250 & 0{,}0574 \\ 0{,}0133 & 0{,}0211 & 0{,}0520 & 0{,}1250 \end{bmatrix}.
$$

Tabelle 4.1: Zellwahrscheinlichkeiten für die Simulationsmodelle bei S (i), CS (ii), DS (iii) und OQS (iv) Symmetrie.

Algebraische Tests basieren auf einer Markov-Kette der Länge l. Die ersten b Tafeln werden in der Einschwingphase vernachlässigt und von den übrigen Zuständen wird jede $s$te Tafel für die Berechnung des p-Wertes berücksichtigt. Nachfolgend wird der

Einfluss dieser Parameter auf die Konvergenz der Markov-Kette und damit auf das Ergebnis der vorgeschlagenen algebraischen Tests analysiert. Vorversuche zeigten keine Wechselwirkungen zwischen den betrachteten Parametern, so dass ihr Einfluss im Weiteren separat untersucht wird. Es wird also jeweils eine Parametereinstellung variiert, die beiden übrigen bleiben auf dem vorgegebenen Standardwert („Ein-Faktor-zur-Zeit-Methode", siehe z. B. Weihs und Jessenberger (1999)). Die betrachteten Längen der Markov-Ketten sind $l =$ 70.000, 80.000, 90.000, ..., 1.000.000; die Schrittlänge nimmt Werte zwischen $s =$ 10 und $s =$ 200, sukzessiv um 10 erhöht, an. Die Einschwingphasen umfassen $b = 0$, 5.000, 10.000, ..., 200.000 Zustände. Die verwendeten Standardwerte sind $l =$ 500.000, $s =$ 100 und $b =$ 50.000. Um die Testergebnisse aus den verschiedenen Einstellungen für s und b vergleichbar zu machen, werden die algebraischen p-Werte gemäß der geringsten Anzahl der betrachteten Datensätze adjustiert. Das bedeutet, dass in diesen Fällen nur $\frac{500.000-200.000}{100} = 3.000$ bzw. $\frac{500.000-50.000}{200} = 2.250$ Tafeln für die Berechnung des p-Wertes berücksichtigt werden. Stehen potenziell mehr Tafeln zur Verfügung, so werden die Datensätze zufällig ausgewählt.

Für die 100 generierten Kontingenztafeln und jede der Parametereinstellungen werden die algebraischen p-Werte des Bowker-Tests, des stetigkeitskorrigierten χ^2-Tests sowie des Tests von May und Johnson mit den entsprechenden exakten p-Werten verglichen. Da der exakte Test in dieser Situation als Standardverfahren angesehen wird, bieten die Differenzen Aufschluss über das Konvergenzverhalten der Markov-Kette. Es werden beispielhaft die Abbildungen für den Bowker-Tests präsentiert. Die Ergebnisse sind für den stetigkeitskorrigierten χ^2-Test sowie den Test von May und Johnson übertragbar und werden daher nicht zusätzlich diskutiert. Die entsprechenden Abbildungen können Anhang A.2 entnommen werden.

In Abbildung 4.1 sind die Differenzen der algebraischen und exakten p-Werte bei variierender Kettenlänge abgetragen. Besteht die Markov-Kette aus bis zu 150.000 generierten Zuständen, so ist ein deutlicher Unterschied in den p-Werten erkennbar; bei einer Kettenlänge von 70.000 beträgt die maximale absolute Differenz 0,087 (Bowker-Test), 0,086 (stetigkeitskorrigierter χ^2-Test) bzw. 0,087 (Test von May und Johnson). Bereits bei 250.000 generierten Zuständen erreichen das algebraische und exakte Verfahren vergleichbare Ergebnisse. Besonders im Hinblick auf größere Stichprobenumfänge wird eine Kettenlänge von 500.000 als Richtwert festgelegt. Sowohl die Schrittlänge als auch die Einschwingphase scheinen einen geringeren Einfluss auf die Konvergenz der

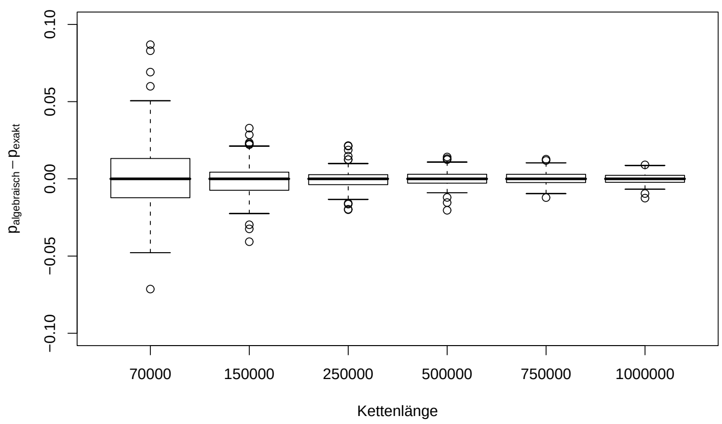

Abbildung 4.1: Boxplots der Differenzen der exakten und algebraischen p-Werte des Bowker-Tests für verschiedene Kettenlängen. Die Datengenerierung erfolgt gemäß dem S-Modell, siehe Tabelle 4.1 (i) mit $n = 25$.

Markov-Kette zu nehmen als die Kettenlänge, siehe Abbildung 4.2. Für die verschiedenen Schrittlängen sind die maximalen absoluten Differenzen zwischen den exakten und algebraischen p-Werten kleiner als 0,04 (0,036 Bowker-Test; 0,037 stetigkeitskorrigierter χ^2-Test bzw. 0,036 Test von May und Johnson). Ähnliche Ergebnisse können für variierende Einschwingphasen beobachtet werden (0,027 Bowker-Test; 0,039 stetigkeitskorrigierter χ^2-Test bzw. 0,027 Test von May und Johnson). Für diese beiden Parameter sind die Unterschiede in den algebraischen und exakten p-Werten gleichbleibend gering. Für die Festlegung der Standardwerte erscheint eine Schrittlänge von $s = 100$ bzw. eine Einschwingphase von $b = 50.000$ Zuständen gerechtfertigt.

Im zweiten Teil dieser Simulationsstudie werden die Eigenschaften der vorgeschlagenen algebraischen Symmetrietests mit den Eigenschaften traditioneller Symmetrietests verglichen. Dazu werden von dem perfekten, bedingten, diagonalen und ordinalen Quasi-Symmetriemodell zufällig 1.000 Datensätze mit insgesamt $n = 25$ und $n = 100$ Beobachtungen generiert. Die jeweiligen Modellwahrscheinlichkeiten sind in Tabelle 4.1 angegeben. Ebenso wie für das S-Modell werden für die weiteren Symmetriemodelle viele Beobachtungen auf der Hauptdiagonalen erwartet. Diese Elemente haben auf das Ergebnis der Symmetrieanalyse keinerlei Einfluss und werden daher gleichgesetzt. Ferner ist zu beachten, dass gemäß der Multinomial-Erhebungsmethode $\widehat{\pi}_{i_1 i_2} > 0$, $i_1 \neq i_2$, sein muss. Für alle simulierten Datensätze werden das perfekte, bedingte, diagonale und or-

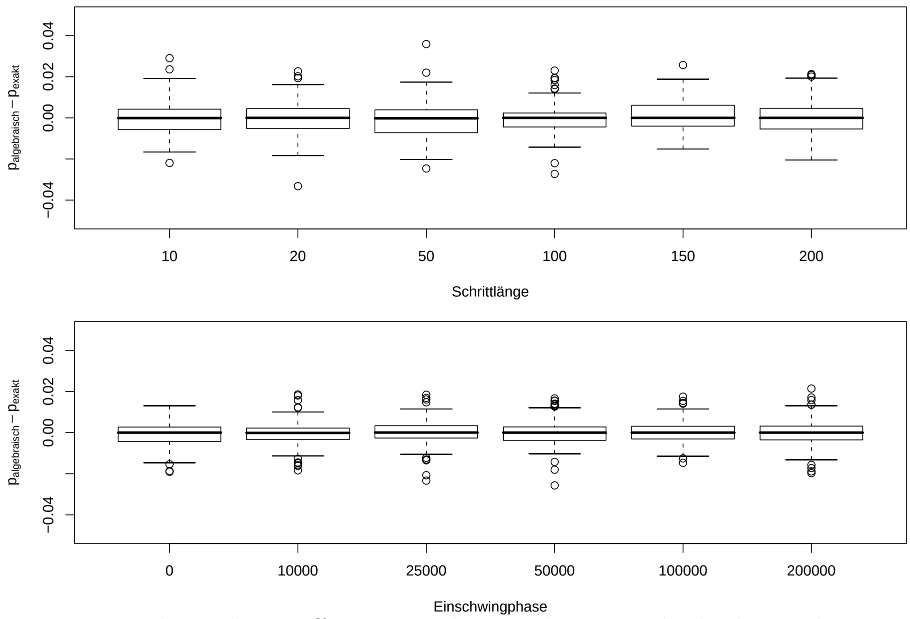

Abbildung 4.2: Boxplots der Differenzen der exakten und algebraischen p-Werte des Bowker-Tests für verschiedene Schrittlängen (oben) bzw. Einschwingphasen (unten). Die Datengenerierung erfolgt gemäß dem S-Modell, siehe Tabelle 4.1 (i) mit $n = 25$.

dinale Quasi-Symmetriemodell anhand der vorgestellten approximativen, exakten sowie der neuen algebraischen Tests überprüft. Die speziellen Modifikationen des Bowker-Tests werden an dieser Stelle nicht gesondert betrachtet, da die Ergebnisse zu denen des Bowker-Tests ähnlich sind.

Es folgt eine Analyse der Kontingenztafeln mit Stichprobenumfang $n = 25$. Für die Simulation von 1.000 geeigneten Datensätzen wurden für das S-Modell 1.890 und für das CS-Modell 1.510 Kontingenztafeln generiert; für das DS Modell waren insgesamt 2.554 und für das OQS-Modell 3.702 Tafeln nötig. Aufgrund des geringen Stichprobenumfangs ist die Eignung asymptotischer Methoden in Frage gestellt, siehe dazu Kapitel 3.3. Beispielhaft werden zunächst die Datensätze betrachtet, die gemäß dem OQS-Modell generiert wurden. Für eine Symmetrieanalyse dieser Tafeln werden die Ergebnisse der approximativen, exakten und algebraischen χ^2-Anpassungstests auf perfekte, bedingte, diagonale und ordinale Quasi-Symmetrie zum Niveau $\alpha = 0,05$ diskutiert. Die entsprechenden Testergebnisse des Likelihood-Quotienten-Tests können Anhang A.2 entnommen werden. Ein qualitativer Vergleich der Testergebnisse erfolgt anhand von Streudiagrammen der p-Werte. Die algebraischen und approximativen p-Werte sind in Abbildung 4.3 zusammenfassend dargestellt. Zwei interessante Beobachtungen seien hier beschrieben. Einerseits ist eine große Streuung der p-Wert-Paare deutlich erkennbar, d. h. die betrachteten p-Werte weichen zum Teil sehr stark voneinander ab. So ist die maximale absolute Differenz der approximativen und algebraischen p-Werte für einen χ^2-Anpassungstest aller untersuchten Symmetriemodelle größer als 0,6 (0,623 für das S-, 0,608 für das CS-, 0,749 für das DS- und 0,749 für das OQS-Modell). Zum anderen sind die algebraischen p-Werte überwiegend größer als die entsprechenden approximativen. Dies kann insbesondere für höhere p-Werte beobachtet werden. Von 1.000 gemäß dem OQS-Modell generierten Kontingenztafeln führt ein Vergleich der algebraischen und approximativen Testentscheidungen zu insgesamt nur 18 (S), 20 (TS), 11 (DS) bzw. 38 (OQS) unterschiedlichen Beurteilungen der Anpassungsgüte.

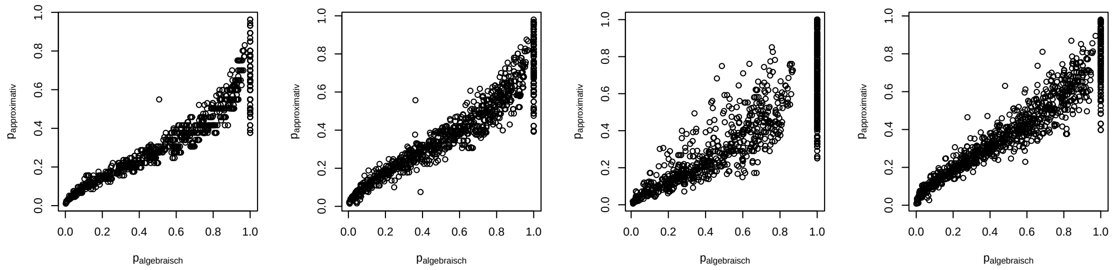

Abbildung 4.3: Vergleich der approximativen und algebraischen p-Werte der χ^2-Anpassungstests. Das zu testende Modell wird repräsentiert durch die Spalten (S-, CS-, DS-, OQS-Modell, von links nach rechts). Die Datengenerierung erfolgt gemäß dem OQS-Modell, siehe Tabelle 4.1 (iv) mit $n = 25$.

Die algebraischen und exakten Verfahren bringen sehr ähnliche Ergebnisse hervor, denn die p-Werte sind auf der jeweiligen Diagonalen angeordnet, siehe Abbildung 4.4. Insgesamt wird nur bei einer Kontingenztafel die Hypothese der perfekten Symmetrie mit dem exakten Test abgelehnt, während der algebraische p-Wert größer als das vorgegebene Niveau $\alpha = 0,05$ ist. Allerdings ist die Differenz der beiden p-Werte sehr gering ($p_{\text{algebraisch}} = 0,046$, $p_{\text{exakt}} = 0,052$). Umgekehrt lehnt der algebraische Test nur bei einem Datensatz die perfekte ($p_{\text{algebraisch}} = 0,051$, $p_{\text{exakt}} = 0,049$), bei zwei

Datensätzen die bedingte ($p_{\text{algebraisch}} = 0,049$, $p_{\text{exakt}} = 0,051$; $p_{\text{algebraisch}} = 0,048$, $p_{\text{exakt}} = 0.051$) und bei drei Datensätzen die ordinale Quasi-Symmetriehypothese ab ($p_{\text{algebraisch}} = 0,047$, $p_{\text{exakt}} = 0,052$; $p_{\text{algebraisch}} = 0,047$, $p_{\text{exakt}} = 0,051$; $p_{\text{algebraisch}} = 0,048$, $p_{\text{exakt}} = 0,056$), während der entsprechende exakte Test keine Abweichung von der Nullhypothese erkennt.

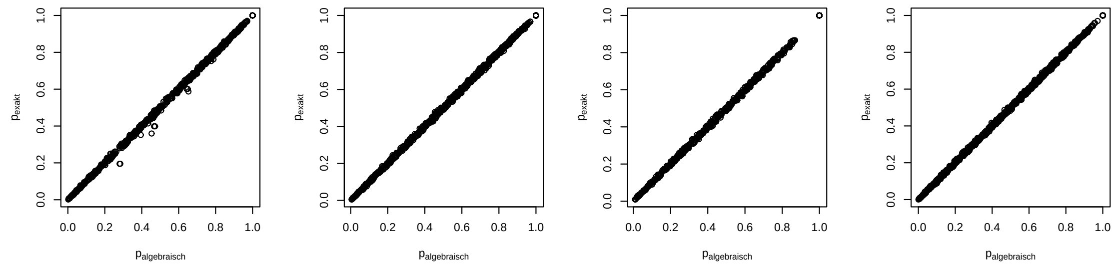

Abbildung 4.4: Vergleich der exakten und algebraischen p-Werte der χ^2-Tests. Das zu testende Modell wird repräsentiert durch die Spalten (S-, CS-, DS-, OQS-Modell, von links nach rechts). Die Datengenerierung erfolgt gemäß dem OQS-Modell, siehe Tabelle 4.1 (iv) mit $n = 25$.

Wie erwartet ist die Anwendung exakter Tests bei Datensätzen mit geringem Stichprobenumfang einem approximativen Test vorzuziehen. Daher interessiert nachfolgend insbesondere der Vergleich der algebraischen Symmetrietests mit den jeweiligen exakten Verfahren. Die Testergebnisse des χ^2-Anpassungstests für die aus dem S-, CS- und DS-Modell generierten Daten sind in Abbildung 4.5 abgetragen. Die Spalten dieser Abbildungsmatrix entsprechen den durchgeführten Anpassungstests für perfekte, bedingte, diagonale und ordinale Quasi-Symmetrie; die Zeilen repräsentieren das zugrunde liegende Datengenerierungsmodell (S, CS, DS). Insgesamt sind die p-Werte auf der Diagonalen angeordnet, was eine gute Übereinstimmung der Testergebnisse beider Verfahren anzeigt.

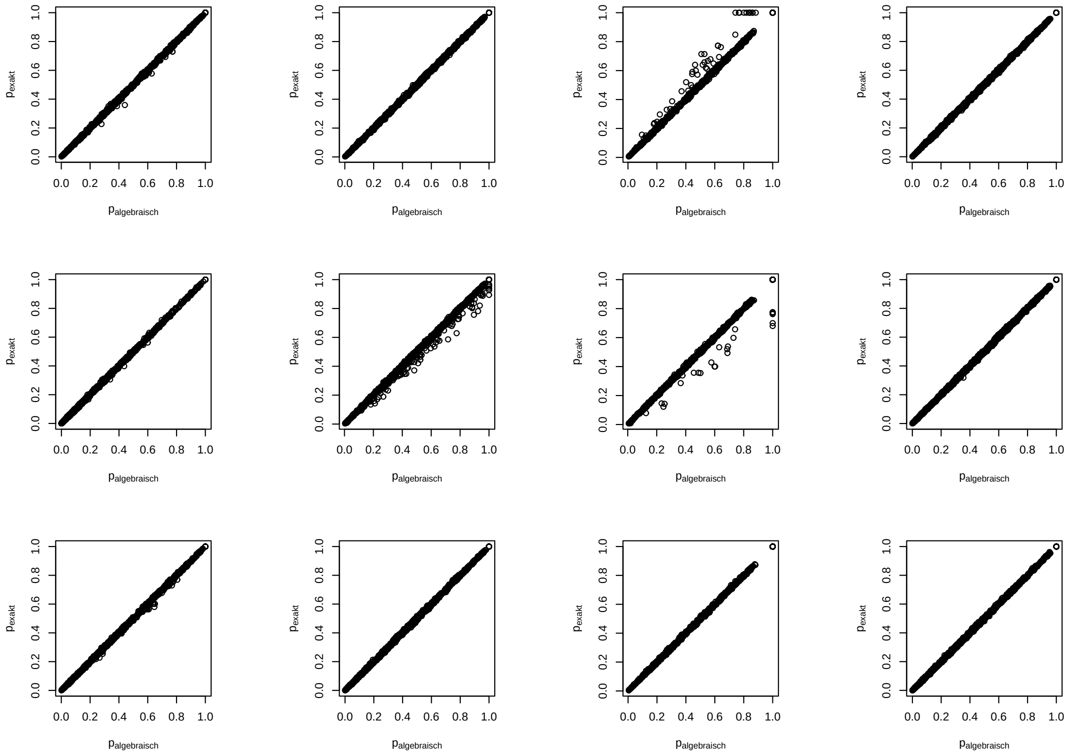

Abbildung 4.5: Vergleich der exakten und algebraischen p-Werte der χ^2-Tests. Das zu testende Modell wird repräsentiert durch die Spalten (S-, CS-, DS-, OQS-Modell, von links nach rechts). Die Datengenerierung wird durch die Zeilen wiedergegeben (S-, CS-, DS-Modell, von oben nach unten, siehe Tabelle 4.1 (i)-(iii) mit $n = 25$).

Die Approximation der Verteilung der Teststatistiken verbessert sich naturgemäß mit wachsendem Stichprobenumfang. Im Anschluss werden daher für $n = 100$ die approximativen und algebraischen Testergebnisse des χ^2-Anpassungstests analysiert. In Abbildung 4.6 wird wiederum das Datengenerierungsmodell (S, CD, DS, OQS) durch die Zeilen symbolisiert; die Spalten zeigen das getestete Modell (S, CS, DS, OQS) an. Die Ergebnisse des entsprechenden Likelihood-Quotienten-Tests sind im Anhang A.2 zu finden.

Beide χ^2-Anpassungstests liefern für einen Test auf das bedingte und das ordinale Quasi-Symmetriemodell sehr ähnliche Ergebnisse; die gepaarten p-Werte sind überwiegend auf der Hauptdiagonalen angeordnet. Wird das perfekte Symmetriemodell überprüft, so sind die p-Werte in einem kleinen Bogen abgetragen. Die algebraischen Tests ergeben überwiegend etwas größere p-Werte als die approximativen Pendants.

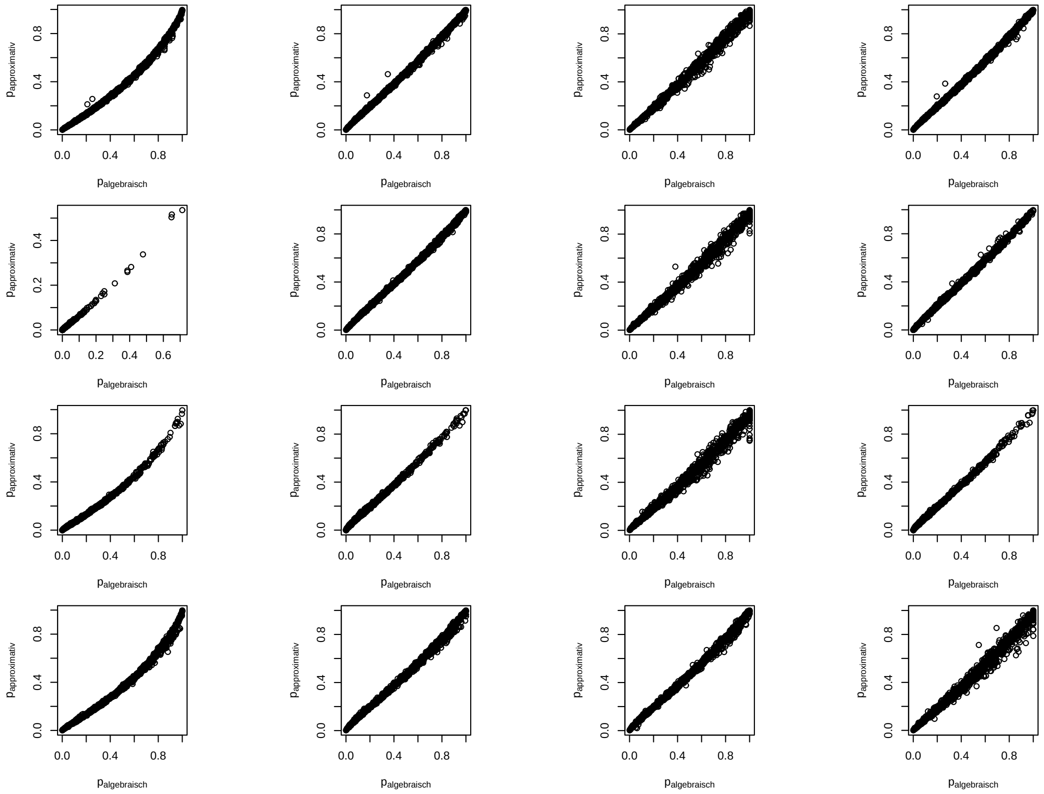

Abbildung 4.6: Vergleich der approximativen und algebraischen p-Werte der χ^2-Tests. Das zu testende Modell wird repräsentiert durch die Spalten (S-, CS-, DS-, OQS-Modell, von links nach rechts). Die Datengenerierung wird durch die Zeilen wiedergegeben (S-, CS-, DS-, OQS-Modell, von oben nach unten, siehe Tabelle 4.1 (i)-(iv) mit $n = 100$).

Die Ergebnisse der Tests auf diagonale Symmetrie weichen stärker voneinander ab als bei den anderen untersuchten Modellen; die p-Werte streuen um die Hauptdiagonale. Für eine Symmetrieanalyse wird in der Praxis üblicherweise eine Kontingenztafel mit großen Einträgen auf der Hauptdiagonalen erwartet. Die Simulation möglichst authentischer Datensätze wurde durch die Datengenerierungsmodelle, siehe Tabelle 4.1, gewährleistet. Dieses Design hat zur Folge, dass Datensätze selbst bei einem Stichprobenumfang von $n = 100$ nur wenige Realisationen außerhalb der Hauptdiagonalen aufweisen können. Für solche Datensätze ist die Approximation der Verteilung der Teststatistik eventuell noch nicht gerechtfertigt, siehe Kapitel 3.3.

4.1.3 Datenbeispiele

Zur Verdeutlichung der vorgestellten algebraischen Tests werden in diesem Abschnitt die Symmetriestrukturen zweier realer Datensätze analysiert. Für alle betrachteten Modelle wird der χ^2-Anpassungstest durchgeführt, die algebraischen p-Werte basieren auf den Standardeinstellungen wie in Kapitel 4.1.2 beschrieben.

Zur Illustration ihrer Symmetriemodelle untersuchen McCullagh (1978), Goodman (1979) und Agresti (1983) die Sehkraft von insgesamt 7477 britischen Frauen, die von 1943 bis 1946 in königlichen Militärmaterial-Fabriken arbeiteten. Insbesondere interessierte die Kurzsichtigkeit der Arbeiterinnen. Getrennt voneinander wurde die Sehkraft des linken und rechten Auges ohne Sehhilfe in vier Kategorien eingeteilt, wobei 1 die beste und 4 die schlechteste Sehkraft kodiert. Der Datensatz ist in Tabelle 4.2 angegeben.

	Sehkraft linkes Auge			
Sehkraft rechtes Auge	1	2	3	4
1	1520	266	124	66
2	234	1512	432	78
3	117	362	1772	205
4	36	82	179	492

Tabelle 4.2: Datensatz 4.1. Beobachtete Häufigkeiten der Kurzsichtigkeit des linken und rechten Auges von 7477 britischen Arbeiterinnen.

Aufgrund der großen Zelleinträge ist zu erwarten, dass die asymptotischen Testergebnisse adäquat sind und mit den simulierten gut übereinstimmen. Die Maximum-Likelihood-Schätzer der Modellparameter sind $\hat{\tau} = 0.926$ (CS), $\hat{\alpha_1} = 0.924$, $\hat{\alpha_2} = 0.993$, $\hat{\alpha_3} = 0.706$ (DS) und $\hat{\beta} = -0,10706$ (OQS). Weil $\hat{\tau}$ und $\hat{\alpha_k}$, $k = 1, \ldots, 3$, kleiner als Eins sind und $\hat{\beta}$ kleiner als Null ist, weisen das CS-, DS- und OQS-Modell auf eine höhere Wahrscheinlichkeit für eine bessere Sehkraft beim rechten als beim linken Auge hin. Speziell wird für das CS-Modell die Wahrscheinlichkeit, dass die Sehstärke des rechten Auges k Kategorien besser ist als die des linken, geschätzt durch 1,16 mal die Wahrscheinlichkeit, dass die Sehstärke des rechten Auges k Kategorien schlechter ist als die des linken. Für das diagonale Symmetriemodell wird $\frac{\pi_{i+k,i}}{\pi_{i,i+k}}$ geschätzt durch 1,16;

1,01 bzw. 1,83; $k = 1, 2, 3$. Unter Annahme des OQS-Modells werden diese Quotienten geschätzt als 1,11; 1,24 und 1,38, denn $\frac{\hat{\pi}_{i+k,i}}{\hat{\pi}_{i,i+k}} = e^{-0{,}10706\cdot(u_i - u_{i+k})} = e^{0{,}10706\cdot k}$, $k = 1, 2, 3$.

In Tabelle 4.3 sind die p-Werte des χ^2- und des Likelihood-Quotienten-Tests für die perfekte, bedingte, diagonale und ordinale Quasi-Symmetrie angegeben. Das perfekte Symmetriemodell scheint ungeeignet, Datensatz 4.1 zu beschreiben; das CS- und das OQS-Modell weisen eine ähnlich gute Anpassung auf. Das diagonale Symmetriemodell scheint Datensatz 4.1 am besten zu repräsentieren.

	p-Wert	
	approximativ	algebraisch
S	0,004	0,002
	0,004	0,002
CS	0,202	0,209
	0,196	0,205
DS	0,919	0,917
	0,919	0,917
OQS	0,201	0,205
	0,201	0,204

Tabelle 4.3: p-Werte des χ^2-Anpassungstests (1. Wert) und des Likelihood-Quotienten-Tests (2. Wert) für perfekte (S), bedingte (CS), diagonale (DS) und ordinale Quasi-Symmetrie (OQS) für Datensatz 4.1.

Der zweite Datensatz entstammt einer retrospektiven Studie über Speiseröhrenkrebs bei chinesischen Männern. Anhand von verschiedenen Kriterien, wie zum Beispiel das Alter, wurden jedem der insgesamt 80 Speiseröhrenkrebs-Patienten vier Kontrollpersonen zugewiesen. Beide Gruppen bekamen sehr heiße Getränke angeboten. In Tabelle 4.4 ist die Anzahl der getrunkenen Becher für die Patienten und der ersten Gruppe von Kontrollpersonen zusammengefasst, siehe Agresti (1983) und Breslow (1982).

Datensatz 4.2 weist sowohl Nullzellen als auch sehr geringe Zelleinträge auf, so dass die Approximation der Verteilung der Teststatistik vielleicht nicht gerechtfertigt ist. Neben den approximativen und algebraischen Tests werden auch die jeweiligen exakten Tests durchgeführt. Für das S-Modell umfasst $\mathscr{Z}_t$ insgesamt 30.240 Kontingenztafeln. Unter Annahme des CS-Modells gibt es 588, für das DS-Modell 12 und für das OQS-Modell 273 Tafeln mit derselben realisierten suffizienten Statistik wie Datensatz 4.2. Der Parameter τ des CS-Modells wird durch die Maximum-Likelihood-Methode als

	Kontrollpersonen			
Patienten	0	1	2	3
0	31	5	5	0
1	12	1	0	0
2	14	1	2	1
3	6	1	1	0

Tabelle 4.4: Datensatz 4.2. Beobachtete Anzahl der getrunkenen Heißgetränke von Speiseröhrenkrebs-Patienten und Kontrollpersonen.

$\hat{\tau} = 1,522$ geschätzt. Für die Parameter des DS-Modells ergeben sich als ML-Schätzer $\hat{\alpha_1} = 1,4$, $\hat{\alpha_2} = 1,5$ sowie $\hat{\alpha_3} = 2,0$, und für das OQS-Modell ist $\hat{\beta} = 1,617$. In Tabelle 4.4 sind die p-Werte der χ^2-Anpassungstests sowie der Likelihood-Quotienten-Tests aufgeführt. Wie erwartet weichen die approximativen p-Werte von den exakten und algebraischen p-Werten etwas ab. Alle durchgeführten Tests kommen aber zum Niveau $\alpha = 0,05$ zu denselben Testergebnissen. Das diagonale und das ordinale Quasi-Symmetriemodell scheinen die Daten am besten darzustellen.

	p-Wert		
	approx.	exakt	algebraisch
S	0,019	0,005	0,002
	0,005	0,004	0,002
CS	0,608	0,669	0,670
	0,376	0,461	0,458
DS	0,762	1,000	1,000
	0,646	1,000	1,000
OQS	0,790	0,829	0,822
	0,631	0,807	0,796

Tabelle 4.5: p-Werte des χ^2-Anpassungstests (1. Wert) und des Likelihood-Quotienten-Tests (2. Wert) für perfekte (S), bedingte (CS), diagonale (DS) und ordinale Quasi-Symmetrie (OQS) für Datensatz 4.2.

Wird das OQS-Modell angenommen, ist die geschätzte Wahrscheinlichkeit, dass ein Patient k Heißgetränke mehr als eine Kontrollperson trinkt $2,061^k$ mal die Wahrscheinlichkeit, dass ein Patient k Heißgetränke weniger als eine Kontrollperson trinkt, $k = 1,2,3$. Im DS-Modell wird $\frac{\hat{\pi}_{i+k,i}}{\hat{\pi}_{i,i+k}}$ durch 2,33 bzw. 3 für $k = 1,\ 2$ geschätzt.

Für $k = 3$ ist $\hat{\pi}_{14}$ gleich Null und der Quotient somit nicht definiert.

4.2 Identifikation von Risikofaktoren

Ein wichtiges Anwendungsfeld der Statistik ist die Epidemiologie. Dort ist es häufig von Interesse, so genannte Risikofaktoren für eine Erkrankung zu identifizieren sowie deren Wirkungsgrad zu quantifizieren. Das Odds Ratio (OR) sowie das zugehörige Konfidenzintervall geben hierzu beispielsweise Aufschluss. Nachfolgend wird für den Spezialfall einer 2×2-Kontingenztafel ein Schätzer für das Odds Ratio eingeführt sowie approximative und exakte Konfidenzintervalle für das betrachtete Odds Ratio beschrieben. Auf der Grundlage des Diaconis-Sturmfels-Algorithmus werden anschließend neue algebraische Konfidenzintervalle als Alternative zu den herkömmlichen Verfahren entwickelt. Dieses Vorgehen kann ebenfalls für die Konstruktion eines algebraischen Konfidenzintervalls für das Relative Risiko sowie für das Odds Ratio einer $2 \times 2 \times K$-Kontingenztafel erweitert werden.

Zunächst sei angenommen, dass der zugrunde liegende Datensatz die Beobachtungen zweier binärer Merkmale X_1 und X_2 umfasst. Analog zu Kapitel 2 bezeichne $n_{i_1 i_2}$ die beobachteten Zellhäufigkeiten der multinomial verteilten Zufallsvariablen $N_{i_1 i_2}$, $i_1,\ i_2 = 1, 2$, mit $\sum_{i_1=1}^{2} \sum_{i_2=1}^{2} n_{i_1 i_2} = n$ und $n_{i_1 i_2} \in \{0, \ldots, n\}$. Für die Analyse des möglichen Zusammenhangs der erhobenen Zufallsvariablen wird in der vorliegenden Arbeit das Odds Ratio verwendet. Da seine Definition auf Wahrscheinlichkeiten basiert, wird die entsprechende Notation kurz anhand der folgenden Tafel erläutert:

		X_2: $X_2 = 1$	$X_2 = 2$
X_1	$X_1 = 1$	$\pi_{11} = P(X_1 = 1 \mid X_2 = 1)$	$\pi_{12} = P(X_1 = 1 \mid X_2 = 2)$
	$X_1 = 2$	$\pi_{21} = P(X_1 = 2 \mid X_2 = 1)$	$\pi_{22} = P(X_1 = 2 \mid X_2 = 2)$

Das Odds Ratio (abkürzend auch OR genannt) wird erklärt durch den Quotienten $OR = \frac{\pi_{11}}{\pi_{21}} / \frac{\pi_{12}}{\pi_{22}} = \frac{\pi_{11} \cdot \pi_{22}}{\pi_{21} \cdot \pi_{12}}$ und vergleicht somit die so genannte „Chance“ (engl.: odds), mit der $X_1 = 1$ eintritt, wenn $X_2 = 1$ realisiert ist, mit der Chance für $X_1 = 1$, wenn $X_2 = 2$ beobachtet wird. Das Odds Ratio kann Werte zwischen Null und Unendlich annehmen. Seine Interpretation erfolgt abhängig vom Parameterwert. Der Maximum-Likelihood-Schätzer für das Odds Ratio ist $\widehat{OR} = \frac{N_{11} \cdot N_{22}}{N_{12} \cdot N_{21}}$. Damit existiert kein ML-Schätzwert für $\widehat{OR}$, wenn n_{12} bzw. n_{21} Null sind. Um dies zu umgehen, kann ein modifizierter Schätzer $\widehat{OR}_{\text{mod}} = \frac{(N_{11}+0{,}5) \cdot (N_{22}+0{,}5)}{(N_{12}+0{,}5) \cdot (N_{21}+0{,}5)}$ verwendet werden, vergleiche Haldane (1955) und Fleiss et al. (2003), Kapitel 6. Ein weiterer alternativer Schätzer ist der bedingte ML-Schätzer, siehe z. B. Agresti (2002), Kapitel 3. Im Folgenden wird jedoch stets der oben beschriebene ML-Schätzer $\widehat{OR}$ für das Odds Ratio verwendet.

Neben dem Wert des Punktschätzers interessiert ein Konfidenzintervall für den betrachteten Parameter. Die Konstruktion des approximativen Konfidenzintervalls nach Woolf für das Odds Ratio für 2×2-Tafeln beruht auf der asymptotischen Normalverteilung von $\log(\widehat{OR})$ mit Erwartungswert $\log(OR)$. Die Varianz wird geschätzt durch $\widehat{\text{var}}(\log(\widehat{OR})) = \frac{1}{N_{11}} + \frac{1}{N_{12}} + \frac{1}{N_{21}} + \frac{1}{N_{22}}$, vergleiche z. B. Fleiss et al. (2003), Kapitel 6. Damit gilt insbesondere

$$P\left(-u_{1-\frac{\alpha}{2}} \leq \frac{\log(\widehat{OR}) - \log(OR)}{\sqrt{\widehat{var}(\log(\widehat{OR}))}} \leq u_{1-\frac{\alpha}{2}}\right) \approx 1 - \alpha,$$

wobei $u_{1-\frac{\alpha}{2}}$ das $(1 - \frac{\alpha}{2})$-Quantil der Standardnormalverteilung bezeichnet. Nach Umformen ergibt sich als approximatives $(1 - \alpha) \cdot 100\%$-Konfidenzintervall für das Odds Ratio einer 2×2-Tafel

$$KI(OR) := \left[\widehat{OR} \cdot \exp\left\{\pm u_{1-\frac{\alpha}{2}} \cdot \sqrt{\widehat{\text{var}}(\log(\widehat{OR}))}\right\}\right].$$

Es existieren alternative asymptotische Konfidenzintervalle für das Odds Ratio, vgl. Kreienbrock und Schach (1995), Kapitel 6.2.3, die in der vorliegenden Arbeit nicht weiter betrachtet werden. Die benötigte Verteilung für die Bestimmung dieses Konfidenzintervalls ist approximativ und damit nur für „hinreichend große“ Stichproben gerechtfertigt. Gängige Faustregeln hierzu werden in Kapitel 3.3 der vorliegenden Arbeit beschrieben.

Für ein exaktes Konfidenzintervall bei 2×2-Tafeln wird ausgenutzt, dass das OR zweier erhobener Merkmale X_1 und X_2 genau dann gleich Eins ist, wenn X_1 und

X_2 stochastisch unabhängig voneinander sind. Analog zum exakten Test von Fisher werden die weiteren Betrachtungen auf die zugrunde liegende suffiziente Statistik $T = (N_{1+}, N_{2+}, N_{+1}, N_{+2})'$ für die Parameter im Unabhängigkeitsmodell bedingt. Für den vorliegenden Fall multinomialverteilter Zelleinträge $N_{i_1 i_2}$, $i_1, i_2 = 1, 2$, hängt die bedingte Verteilung von N_{11}, gegeben die realisierte suffiziente Statistik t, nur noch vom Odds Ratio ab, die entsprechende Dichte f ist die nichtzentrale hypergeometrische Dichte

$$f(N_{11} = x | \, t, OR) = \frac{\binom{n_{1+}}{x} \cdot \binom{n_{2+}}{n_{+1}-x} \cdot (OR)^x}{\sum\limits_{u=m_-}^{m^+} \binom{n_{1+}}{u} \cdot \binom{n_{2+}}{n_{+1}-u} \cdot (OR)^u}, \quad m_- \leq x \leq m^+;$$

für die untere und obere Summationsgrenze gilt: $m_- = \max(0, n_{1+} + n_{+1} - n)$ bzw. $m^+ = \min(n_{1+}, n_{+1})$ (siehe Fleiss et al. (2003), Kapitel 6.4, Metha et al. (1985) und Zelen (1971)). Damit wird die untere Grenze OR_* des exakten $(1 - \alpha) \cdot 100\%$-Konfidenzintervalls für das Odds Ratio wie folgt bestimmt

$$\begin{aligned} OR_* = 0 \quad &, \quad \text{falls } \; n_{11} = m_-, \\ \sum_{x=n_{11}}^{m^+} f(x| \, t, OR_*) = \frac{\alpha}{2} \quad &, \quad \text{falls } \; m_- < n_{11} \leq m^+. \end{aligned}$$

Die obere Grenze des Konfidenzintervalls OR^* erfüllt die Bedingung:

$$\begin{aligned} \sum_{x=m_-}^{n_{11}} f(x| \, t, OR^*(n_{21})) = \frac{\alpha}{2} \quad &, \quad \text{falls } \; m_- \leq n_{11} < m^+, \\ OR^* = \infty \quad &, \quad \text{falls } \; n_{11} = m^+. \end{aligned}$$

4.2.1 Ein neues Konfidenzintervall für das Odds Ratio

Gemäß den Ausführungen in Kapitel 3.3 wird nachfolgend ein neues algebraisches Konfidenzintervall für das Odds Ratio als Alternative zu den bekannten asymptotischen und exakten Methoden entwickelt. Basierend auf dem Diaconis-Sturmfels-Algorithmus und MCMC-Simulation wird dafür eine Stichprobe aus der bedingten Verteilung einer diskreten Exponentialfamilie mit beobachteter suffizienter Statistik generiert.

Das Konstruktionsprinzip des neuen Konfidenzintervalls für das Odds Ratio stützt sich sowohl auf das exakte als auch auf das approximative Verfahren. Zunächst wird das zugrunde liegende statistische Modell algebraisch dargestellt. Wie beim exakten Konfidenzintervall für das Odds Ratio bei 2×2-Tafeln wird auch hier die Äquivalenz zwischen dem Odds Ratio-Wert von Eins und der stochastischen Unabhängigkeit der beiden betrachteten Merkmale ausgenutzt. Die gemeinsame Verteilung der multinomialverteilten $N_{i_1 i_2}$, $i_1, i_2 = 1, 2$, ist eine 4-parametrige Exponentialfamilie; eine suffiziente Statistik T für die Parameter im Unabhängigkeitsmodell sind die Zeilen- und Spaltensummen, d. h. $T = (N_{1+}, N_{2+}, N_{+1}, N_{+2})'$. Für eine geeignete algebraische Darstellung des statistischen Modells und damit für die Anwendung des Diaconis-Sturmfels-Algorithmus ist wesentlich, die Menge aller Datensätze mit beobachteter suffizienter Statistik t, d. h. $\mathscr{Z}_t = \{z : \mathscr{H} \to I\!N | \sum_{x \in \mathscr{H}} z(x) T^*(x) = t\}$ und damit auch T^* darzustellen, wobei der Stichprobenraum $\mathscr{H}$ eine endliche Menge ist. Aufgrund des Unabhängigkeitsmodells ist mit der suffizienten Statistik T auch T^* festgelegt: $T^*((i_1, i_2))$ ist ein Vektor mit derselben Länge wie T, d. h. hier 4. Zwei Einträge dieses Vektors sind gleich Eins: an der Stelle i_1 sowie an der Stelle $2 + i_2$, die übrigen Einträge sind Null, vergleiche Diaconis und Sturmfels (1998). Für eine 2×2-Tafel gilt also im Unabhängigkeitsmodell

$$T^*((1,1)) = (1,0,1,0)', \qquad T^*((1,2)) = (1,0,0,1)',$$
$$T^*((2,1)) = (0,1,1,0)', \qquad T^*((2,2)) = (0,1,0,1)'.$$

Entsprechend Theorem 3.3.4 werden diese T^* für die Bestimmung der reduzierten Gröbner-Basis verwendet. D.h. es wird zunächst die Gröbner-Basis $\mathscr{G}_{hilf}$ des Hilfsideals $\mathscr{I}_{hilf}$ mit

$$\mathscr{I}_{hilf} = < x_{11} - N_{1+}N_{+1}, x_{12} - N_{1+}N_{+2}, x_{21} - N_{2+}N_{+1}, x_{22} - N_{2+}N_{+2} >$$

berechnet, vergleiche Diaconis und Sturmfels (1998). Die Gröbner-Basis für das Unabhängigkeitsmodell enthält alle Basispolynome aus $\mathscr{G}_{hilf}$, die ausschließlich Terme aus dem Stichprobenraum $\mathscr{H}$ enthalten.

Bemerkung 4.2.1

Zugrunde liege eine 2×2-Kontingenztafel. Unter Beachtung der gradlexikographischen Monomordnung besteht die reduzierte Gröbner-Basis für das Unabhängigkeitsmodell aus einem Basispolynom, und es ist $\mathscr{G} = \{x_{11} \cdot x_{22} - x_{12} \cdot x_{21}\}$.

Die zulässige Bewegung im Metropolis-Hastings-Algorithmus ist intuitiv $\begin{pmatrix} + & - \\ - & + \end{pmatrix}$ und dient der Generierung von Datensätzen mit denselben realisierten Zeilen- und Spaltensummen wie die zugrunde liegende Kontingenztafel, vergleiche Korollar 3.3.1 und Theorem 3.3.4. Entsprechend den Ausführungen in Kapitel 3.3 wird nun eine Markov-Kette der Länge l generiert, die ersten b Datensätze werden in der so genannten Einschwingphase vernachlässigt und anschließend nur jede $s$te Tafel berücksichtigt. Für jede der verbliebenen $\lfloor \frac{l-b}{s} \rfloor$ Kontingenztafeln wird der geschätzte Wert des logarithmierten Odds Ratios $\log(\widehat{OR})$ berechnet und somit die bedingte Verteilung von $\log(\widehat{OR})$ unter der Nullhypothese, d. h. stochastischer Unabhängigkeit von X_1 und X_2, bei beobachteter suffizienter Statistik simuliert. Für die Bestimmung des simulierten Konfidenzintervalls wird diese Verteilung standardisiert, d. h. für jede der ausgewählten Kontingenztafeln wird $\frac{\log(\widehat{OR})-\log(OR)}{\sqrt{\widehat{var}(\log(\widehat{OR}))}}$ berechnet. Aufgrund der Unabhängigkeitsannahme ist der Erwartungswert $\log(OR) = 0$. Entsprechend dem approximativen Konfidenzintervall nach Woolf wird für das simulierte Konfidenzintervall ausgenutzt, dass gilt:

$$P\left(q_{\text{sim},\,\frac{\alpha}{2}} \leq \frac{\log(\widehat{OR}) - \log(OR)}{\sqrt{\widehat{var}(\log(\widehat{OR}))}} \leq q_{\text{sim},1-\frac{\alpha}{2}}\right) \approx 1-\alpha,$$

wobei die Quantile der simulierten standardisierten Verteilung q_{sim} die approximativ gültigen Standardnormalverteilungsquantile ersetzen. Das algebraische $(1-\alpha)\cdot 100\%$-Konfidenzintervall für das Odds Ratio ist also gegeben durch

$$\begin{aligned} KI(OR) := \Bigg[&\widehat{OR}\cdot\exp\left\{-q_{\text{sim},\,1-\frac{\alpha}{2}}\cdot\sqrt{\widehat{\text{var}}(\log(\widehat{OR}))}\right\}; \\ &\widehat{OR}\cdot\exp\left\{-q_{\text{sim},\,\frac{\alpha}{2}}\cdot\sqrt{\widehat{\text{var}}(\log(\widehat{OR}))}\right\}\Bigg]. \end{aligned}$$

Ein algebraisches Konfidenzintervall für das Relative Risiko (RR) wird analog zum Konfidenzintervall für das OR bestimmt und daher nicht vorgeführt.

Der Mantel-Haenszel-Schätzer (Mantel und Haenszel (1959)) wird häufig für die Schätzung des gemeinsamen Odds Ratios für $2 \times 2 \times K$-Tafeln verwendet. Ein approximatives und exaktes Konfidenzintervall für das OR_{MH} einer $2 \times 2 \times K$-Tafel werden dem 2×2-Fall entsprechend bestimmt, vergleiche z. B. Fleiss (2003), Kapitel

10.3 und Kapitel 10.5. Ein algebraisches Konfidenzintervall für das OR_{MH} basiert auf demselben Vorgehen wie für das Odds Ratio einer 2×2-Tafel und wird daher ebenfalls nicht weiter ausgeführt.

4.2.2 Simulationsstudie

In dieser Simulationsstudie werden das vorgestellte algebraische und die traditionellen 95%-Konfidenzintervalle miteinander verglichen. Die algebraischen Konfidenzgrenzen basieren auf den Standardeinstellungen wie in Kapitel 4.1.2 beschrieben; die resultierenden Markov-Ketten konvergieren. Die Daten für diese Simulationsstudie werden aus zwei möglichst realen Datensituationen generiert. Im ersten Modell wird angenommen, dass es keinen Zusammenhang zwischen den erhobenen Merkmalen gibt. Im zweiten Szenario wird eine Abhängigkeit unterstellt; im epidemiologischen Kontext wird beispielsweise ein schädigender Einfluss des potenziellen Risikofaktors auf die interessierende Krankheit betrachtet. Die gewählten Modellwahrscheinlichkeiten sind in Tabelle 4.6 angegeben.

$$\text{(i)} \begin{bmatrix} 0{,}278 & 0{,}278 \\ 0{,}222 & 0{,}222 \end{bmatrix}, \quad \text{(ii)} \begin{bmatrix} 0{,}308 & 0{,}154 \\ 0{,}154 & 0{,}385 \end{bmatrix}.$$

Tabelle 4.6: Zellwahrscheinlichkeiten für die Simulationsmodelle bei einem geschätzten Wert von $\widehat{OR} = 1$ (i) und $\widehat{OR} = 5$ (ii).

Unter Beachtung der Multinomial-Erhebungstechnik müssen alle $\hat{\pi}_{i_1 i_2} > 0$ sein, $i_1, i_2 = 1, 2$. Ist dies nicht erfüllt, so liegen die geschätzten Erwartungswerte nicht im zulässigen Parameterraum. Ferner sei angenommen, dass die realisierten Einträge $n_{i_1 i_2}$, i_1, $i_2 = 1, 2$, stets größer als Null sind und damit ein ML-Schätzwert für $\widehat{OR}$ bzw. $\widehat{var}(\log(\widehat{OR}))$ existiert.

Unter Beachtung des Unabhängigkeitsmodells (i) werden bei einem Stichprobenumfang von $n = 15$ für eine Simulation von 100 Datensätzen 103 Tafeln simuliert; für das Modell (ii) werden hierfür 158 Kontingenztafeln generiert. Für jede dieser Kontingenztafeln werden die asymptotischen, exakten und algebraischen 95%-Konfidenzintervalle für das Odds Ratio bestimmt. Die Approximation der Verteilung von $\log(\widehat{OR})$ ist aufgrund des geringen Stichprobenumfangs vielleicht noch nicht gerechtfertigt. Für die Erstellung der Grafiken werden die generierten Odds Ratios der Größe nach sortiert und die entsprechenden Konfidenzintervalle abgetragen. Zur Verbesserug der Übersichtlichkeit werden jeweils 25 Datensätze in einer Grafik betrachtet. Es wird eine Auswahl an Grafiken gezeigt, die übrigen können Anhang A.2 entnommen werden.

In Abbildung 4.7 sind die kleinsten 25 Werte des geschätzten Odds Ratios mit den zugehörigen Konfidenzintervallen aus dem Unabhängigkeitsmodell (i) und Modell (ii) zusammengefasst.

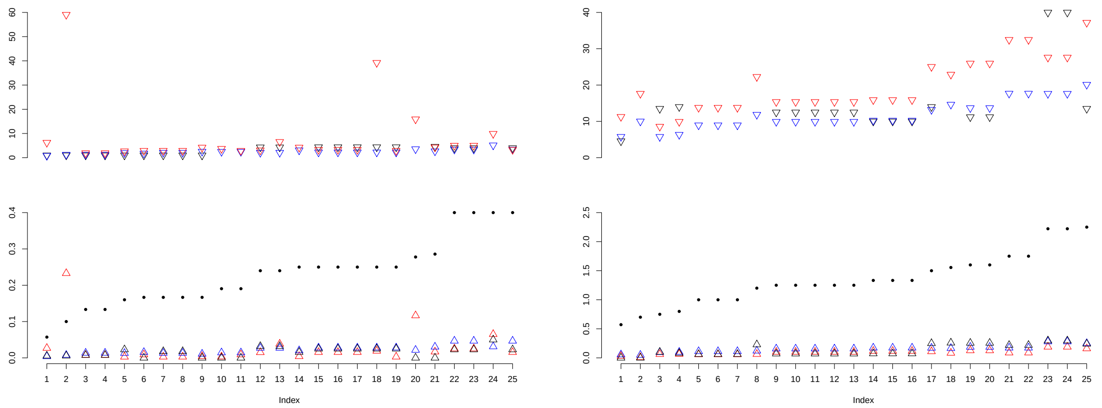

Abbildung 4.7: Vergleich der approximativen, exakten und algebraischen 95% Konfidenzgrenzen (obere Grenze: Bild oben; untere Grenze: Bild unten) für die 25 kleinsten ML-Schätzwerte für das Odds Ratio aus Modell (i) links, bzw. Modell (ii) rechts, aus Tabelle 4.6 mit $n = 15$. Der Wert des geschätzten Odds Ratios wird durch Punkte repräsentiert. Die Konfidenzintervalle werden farblich unterschieden (approximativ: blau, exakt: rot, algebraisch: schwarz).

Es kann beobachtet werden, dass die unteren Konfidenzgrenzen für das Odds Ratio bei allen angewendeten Prozeduren überwiegend gut übereinstimmen. Die oberen Grenzen hingegen weichen zum Teil stark voneinander ab. Insbesondere ist die obere Grenze des exakten Konfidenzintervalls oftmals deutlich größer als die entsprechende Grenze des approximativen Konfidenzintervalls; die exakte Prozedur ist wie erwartet konservativ. Die obere algebraische Konfidenzgrenze nimmt in wenigen Fällen den Wert Unendlich an (beispielsweise Datensatznr. 10, Abbildung 4.7, links). Ursächlich hierfür ist, dass „zu viele" Datensätze mit Nullen in den Zellen n_{12} und n_{21} generiert werden. Liegt ein Datensatz mit einem geringen Stichprobenumfang vor, so könnte der modifizierte Schätzer $\widehat{OR}_{\text{mod}}$ für das Odds Ratio das algebraische Konfidenzintervall verbessern.

Bei einem Stichprobenumfang von $n = 50$ werden 100 Datensätze gemäß dem Unabhängigkeitsmodell (i) sowie gemäß Modell (ii) zufällig generiert. Beispielhaft werden wiederum aus jedem Szenario 25 Datensätze analysiert, weitere Grafiken befinden sich im Anhang A.2.

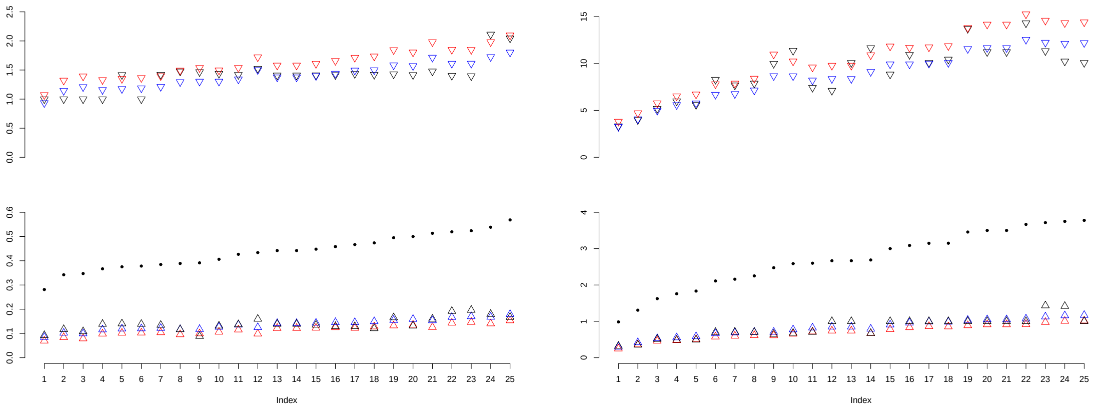

Abbildung 4.8: Vergleich der approximativen, exakten und algebraischen 95% Konfidenzgrenzen (obere Grenze: Bild oben; untere Grenze: Bild unten) für die 25 kleinsten ML-Schätzwerte für das Odds Ratio aus Modell (i) links, bzw. Modell (ii) rechts, aus Tabelle 4.6 mit $n = 50$. Der Wert des geschätzten Odds Ratios wird durch Punkte repräsentiert. Die Konfidenzintervalle werden farblich unterschieden (approximativ: blau, exakt: rot, algebraisch: schwarz).

Die approximativen, exakten und algebraischen unteren Konfidenzgrenzen sind ungefähr identisch. Die oberen Konfidenzgrenzen weichen stärker voneinander ab. Das exakte Verfahren liefert erwartungsgemäß häufig das größte und damit konservativste Konfidenzintervall. Die algebraischen und approximativen Intervallgrenzen unterscheiden sich zum Teil nur geringfügig, wobei das algebraische Konfidenzintervall bei vielen der betrachteten Datensätze am kleinsten war. Werte von Unendlich als obere algebraische Konfidenzgrenze werden hier nicht angenommen. Die vorgeschlagene algebraische Prozedur bietet demnach eine wertvolle Ergänzung zu den traditionellen Konfidenzintervallen für das Odds Ratio.

4.2.3 Datenbeispiele

In dem folgenden Kapitel werden für reale Datensätze potenzielle Risikofaktoren für eine interessierende Krankheit anhand des algebraischen sowie traditioneller Konfidenzintervalle für das Odds Ratio untersucht.

Zunächst interessiert eine Fall-Kontroll-Studie, die zwischen Juni 1985 und Dezember 1988 in Birmingham und Alabama durchgeführt wurde. In dieser Studie wurde die Entstehung eines Endometriumkarzinoms untersucht; speziell wurde der Zusammenhang zwischen dieser Krankheit und den Ernährungsgewohnheiten der Frauen analysiert. Dazu haben 103 Patientinnen und 236 Kontrollen einen Fragebogen ausgefüllt. In der vorliegenden Arbeit wird ein Teildatensatz dieser Fall-Kontroll-Studie ausgewertet, der in Tabelle 4.7 zusammengefasst ist, siehe Barbarone et al. (1993) sowie Kreienbrock und Schach (1995), S. 199.

	Verzehr von Milchprodukten	
	Ja	Nein
Fall	61	42
Kontrolle	162	74

Tabelle 4.7: Datensatz 4.3. Beobachtete Anzahl von Endometriumkarzinom-Fällen und Kontrollen bei regelmäßigem Genuss von Milchprodukten.

Das geschätzte Odds Ratio beträgt 0,663, d. h. es wird zunächst ein präventiver Einfluss von Milchprodukten auf die Entstehung eines Endometriumkarzinoms vermutet. Alle drei 95%-Konfidenzintervalle überdecken jedoch die Eins (approximativ: 0,411;

1,072; algebraisch: 0,389; 1,062; exakt: 0,400; 1.106), so dass zum Niveau $\alpha = 0,05$ ein Zusammenhang nicht nachgewiesen werden kann. Die Konfidenzgrenzen aller drei verwendeten Verfahren sind sehr ähnlich.

Der zweite Datensatz entstammt einer retrospektiven Studie. 41 Patienten mit Kehlkopfkrebs unterzogen sich einer Operation bzw. einer Strahlentherapie. Zwei Jahre nach ihrer Behandlung wurde die Anzahl der Rezidive festgehalten, siehe Agresti (2002), S. 107. Dabei interessiert, ob die gewählte Therapieform (Operation bzw. Strahlentherapie) und der Krankheitsstatus bei Kehlkopfkrebs voneinander abhängen. Die Daten können Tabelle 4.8 entnommen werden.

	Therapieform	
	Strahlentherapie	Operation
Rezidiv	3	2
kein Rezidiv	15	21

Tabelle 4.8: Datensatz 4.4. Beobachtete Anzahl von Rezidivfällen bzw. Nicht-Rezidivfällen bei Kehlkopfkarzinom nach einer Operation und nach einer Strahlentherapie.

Der Wert des geschätzten Odds Ratios beträgt 2,1. Das bedeutet, dass die Chance, mit der ein Rezidiv nach erfolgter Strahlentherapie eintritt, 2,1 mal so hoch ist wie ein Rezidiv nach einer Operation zu erleiden. Dieser Zusammenhang kann wiederum nicht statistisch nachgewiesen werden, denn die 95%-Konfidenzintervalle überdecken die Eins (approximativ: 0,312; 14,152; algebraisch: 0,334; ∞; exakt: 0,209; 27,552). Die hohe obere Grenze des algebraischen Konfidenzintervalls ist bedingt durch die kleinen Einträge in der ersten Zeile von Datensatz 4.4 im Vergleich zur zweiten Zeile; im Metropolis-Hastings-Algorithmus werden häufig Datensätze mit Nullzellen generiert. Dies bereitet Schwierigkeiten für eine Schätzung des jeweiligen Odds Ratios sowie der zugehörigen Standardabweichung. Abhilfe könnte hier der in Kapitel 4.2 beschriebene modifizierte Schätzer $\widehat{OR}_{\text{mod}}$ schaffen.

Kapitel 5

Algebraische Modellselektion

Eine wichtige und grundlegende Aufgabe in der Statistik ist es, Abhängigkeitsstrukturen in einem Datensatz zu erkennen. So dienen die in Kapitel 2 vorgestellten log-linearen und graphischen Modelle der Beschreibung verschiedener Unabhängigkeitsbeziehungen. Häufig ist von Interesse, aus einer gegebenen Menge möglicher Modelle dasjenige auszuwählen, das die beobachteten Daten gemäß einem bestimmten Kriterium am besten beschreibt. Da ein Modell aufgrund der zu seiner Bildung herangezogenen Daten beurteilt wird, ist eine Modellwahl prinzipiell explorativ zu bewerten, siehe Chatfield (1995) und Edwards (2000), Kapitel 6.

Es existieren verschiedene Ansätze zur Beurteilung der Anpassungsgüte von Modellen an Daten. Akaike entwickelt beispielsweise einen Schätzer für die erwartete relative Kullback Leibler-Information basierend auf der maximierten log-Likelihood-Funktion und einem Korrekturterm für die Verzerrung (Burnham und Anderson (2002), Kapitel 2 und 7). In dieser Arbeit werden Strategien betrachtet, die (üblicherweise approximative oder exakte) Anpassungstests zur Beurteilung der Modelle nutzen. Wie bereits gezeigt, bietet der Algorithmus von Diaconis und Sturmfels eine wertvolle Ergänzung zu diesen traditionellen Testmethoden. Die algebraischen Tests basieren jedoch auf eine separate Simulation für jedes betrachtete Modell, weshalb sie sehr aufwändig und daher für diese Fragestellung in der Praxis ungeeignet erscheinen.

Weisen die interessierenden Modelle eine hierarchische Struktur auf, so kann der erforderliche Simulationsaufwand deutlich verringert werden. Das vorgeschlagene Verfahren wird am Beispiel der Modellselektion für graphische Modelle sowie der in Kapitel 4.1 vorgestellten Symmetrietests vorgeführt. Die so generierten algebraischen p-Werte ersetzen anschließend die traditionellen asymptotischen bzw. exakten p-Werte.

In einer Simulationsstudie werden traditionelle Modellselektionsmethoden für graphische Modelle mit dem herkömmlichen Diaconis-Sturmfels-Verfahren sowie der neu eingeführten Methode verglichen.

5.1 Methoden zur Modellselektion

In diesem Abschnitt werden verschiedene gängige Modellwahlstrategien für graphische Modelle vorgestellt. Die jeweiligen Anpassungen werden mit dem χ^2-Test oder dem Likelihood-Quotienten-Test beurteilt. Die weiteren Ausführungen stützen sich im Wesentlichen auf Edwards (2000).

Bei der *schrittweisen Modellwahl* wie der Rückwärtsselektion werden ausgehend von dem saturierten Modell sukzessiv die Kanten mit dem größten nicht-signifikanten p-Wert des durchgeführten Anpassungstests entfernt. Dabei kann das von Gabriel (1969) eingeführte Prinzip der Kohärenz (coherence) beachtet werden. Wird beispielsweise die Entfernung einer Ecke in einem Schritt abgelehnt, so kann sie auch in den folgenden Schritten nicht entfernt werden. Sind alle p-Werte signifikant, so stoppt die Prozedur. Die Vorwärtsselektion startet hingegen mit dem Unabhängigkeitsmodell und fügt in jedem Schritt die Kante mit dem kleinsten p-Wert des Anpassungstests hinzu. Auch hier kann das Kohärenzprinzip beachtet werden. Die Modellwahl ist abgeschlossen, wenn alle p-Werte nicht-signifikant sind. Die Rückwärtsselektion betrachtet im Gegensatz zur Vorwärtsselektion ausschließlich Modelle, die von den Daten unterstützt werden und wird daher üblicherweise vorgezogen. Beide Modellwahlprozeduren bieten die Möglichkeit, Variablen zu fixieren. Bei Durchführung der Rückwärtsselektion stehen diese Variablen nicht für eine sukzessive Entfernung zur Verfügung. Häufig wird die interessierende Modellklasse durch die Forderung nach Zerlegbarkeit eingeschränkt. Wichtige theoretische Ergebnisse erleichtern die Handhabung dieser Modelle. So stehen beispielsweise exakte Tests nur für zerlegbare graphische Modelle zur Verfügung.

Die Modellselektion gemäß der *Edwards und Havránek Prozedur* (EH-Prozedur), Edwards und Havránek (1985, 1987), basiert auf einen globalen Suchalgorithmus, der eine Menge von Modellen auswählt, die mit den Daten 'konsistent' sind. Zur Vereinfachung wird nachfolgend anstelle von „das Modell konnte zum Niveau α nicht verworfen werden“ ersetzt durch „das Modell ist zum Niveau α akzeptabel“ (accepted). Edwards und Havránek (1985, 1987) reduzieren die Anzahl der interessierenden Modelle gemäß dem Kohärenzprinzip von Gabriel (1969) wie folgt: Gilt ein Modell als akzeptabel, so werden die auf diesem Modell aufbauenden weiteren Modelle als schwach akzeptabel (weakly accepted) angenommen. Wird umgekehrt ein Modell abgelehnt, so sind auch die Untermodelle schwach abgelehnt (weakly rejected).

Das Ziel der EH-Prozedur ist, für die interessierende Menge möglicher Modelle $\mathcal{M}$ eine Menge akzeptabler Modelle $\mathcal{A}$ sowie eine Menge abgelehnter Modelle $\mathcal{R}$ mit $\mathcal{A},\ \mathcal{R} \subseteq \mathcal{M}$ zu finden, so dass jedes weitere Modell als schwach akzeptabel bzw. schwach abgelehnt eingeordnet werden kann. Dabei ist zu beachten, dass $\mathcal{A}$ und $\mathcal{R}$ nicht notwendigerweise eindeutig bestimmt sind, siehe Edwards und Havránek (1985). Im Weiteren wird sowohl eine Teilmenge als auch eine bestehende Hierarchie, wie z. B. ein Untermodell, durch „$\subseteq$“ gekennzeichnet.

M_1 sei ein Untermodell von M_2, $M_1 \subseteq M_2$. Für jede Teilmenge von Modellen $S \subseteq \mathcal{M}$ werden die komplexesten Modelle (maximal models) aus S bestimmt durch $\max(S) = \{M_1 \in S : M_1 \subset M_2 \Rightarrow M_2 \notin S\}$. Analog dazu sind die einfachsten Modelle (minimal models) aus S dargestellt als $\min(S) = \{M_1 \in S : M_2 \subset M_1 \Rightarrow M_2 \notin S\}$. Eine Menge von Modellen S heißt nicht vergleichbar (incomparable), wenn keine Modelle $M_1,\ M_2 \in S$ existieren, so dass $M_1 \subset M_2$ gilt.

Die Mengen $\mathcal{A}$ und $\mathcal{R}$ werden anhand der so genannten r- und a-Duale bestimmt. Dazu sei $S = \{M_1, \dots M_p\},\ M_i \in \mathcal{M},\ i = 1, \dots, p$, eine Menge akzeptabler Modelle. Dann bezeichne $I_a(S) = \{M \in \mathcal{M} : M_i \subseteq M$ für ein $M_i \in S\}$ die Menge der schwach akzeptablen Modelle aus $\mathcal{M}$; die übrigen Modelle werden in $I_a^c(S) = \{M \in \mathcal{M} : M_i \nsubseteq M$ für ein $M_i \in S\}$ zusammengefasst. Die Menge der maximalen (komplexesten) Modelle aus $I_a^c(S)$ werden als r-Dual, $D_r(S)$, von S bezeichnet. Nun sei $S = \{M_1, \dots M_p\},\ M_i \in \mathcal{M},\ i = 1, \dots, p$, eine Menge abgelehnter Modelle. Die Menge der schwach abgelehnten Modelle aus $\mathcal{M}$ wird beschrieben durch $I_r(S) = \{M \in \mathcal{M} : M \subseteq M_i$ für ein $M_i \in S\}$. Dann ist das a-Dual von S, $D_a(S)$, die Menge der einfachsten Modelle aus $I_r^c(S) = \{M \in \mathcal{M} : M \nsubseteq M_i$ für ein $M_i \in S\}$. Insgesamt muss $I_a(\mathcal{A}) \cap I_r(\mathcal{R}) = \emptyset$ gelten.

Die Edwards-Havránek-Prozedur ermöglicht die Beurteilung der Anpassung aller Modelle aus der zugrunde liegenden Modellfamilie $\mathcal{M}$ in endlich vielen Schritten.

Schritt 1:
Teste die Anpassung der Modelle einer Anfangsmenge S_0 und ordne sie entsprechend der Testergebnisse in die Mengen $\mathcal{A}$ und $\mathcal{R}$ ein.

Schritt 2:
Ist die Menge der abgelehnten Modelle $\mathcal{R}$ leer, so gehe zu Schritt 2a. Ist umgekehrt die Menge der akzeptablen Modelle $\mathcal{A}$ leer, so führe Schritt 2b aus. Sind weder $\mathcal{A}$ noch $\mathcal{R}$ leer, so kann beliebig Schritt 2a oder 2b als nächstes gewählt werden.

> Schritt 2a:
> Teste die Anpassung der Modelle aus $D_r(\mathcal{A}) \setminus \mathcal{R}$. Die Prozedur stoppt, wenn jedes dieser Modelle abgelehnt werden kann. Ansonsten aktualisiere die Mengen $\mathcal{A}$ und $\mathcal{R}$, und gehe zurück zu Schritt 2.
> Schritt 2b:
> Teste die Anpassung der Modelle aus $D_a(\mathcal{R}) \setminus \mathcal{A}$. Die Prozedur stoppt, wenn jedes dieser Modelle als akzeptabel angenommen werden kann. Ansonsten aktualisiere die Mengen $\mathcal{A}$ und $\mathcal{R}$, und gehe zurück zu Schritt 2.

Edwards und Havránek (1985, 1987) beschreiben Kriterien für die Festlegung der Anfangsmenge S_0, der „Anpassungsrichtung“ sowie des durchzuführenden Anpassungstests.

5.2 Neue algebraische Modellselektionsprozeduren

Die vorgestellten Modellselektionsstrategien beurteilen die Modellanpassungen anhand approximativer bzw. exakter Tests. Oftmals sind diese Annäherungen nicht gerechtfertigt oder eine exakte Berechnung zu aufwändig. Algebraische Test gemäß dem Diaconis-Sturmfels-Algorithmus basieren auf Einzelsimulationen für jedes betrachtete Modell; für eine Modellwahl sind sie daher ebenfalls in der Regel ungeeignet. Am Beispiel graphischer Modelle wird nachfolgend ein algebraisches Verfahren zur Modellselektion entwickelt. Dazu werden der Diaconis-Sturmfels-Algorithmus sowie die hierarchische Struktur graphischer Modelle ausgenutzt. Speziell werden algebraische p-Werte bereit gestellt, die die traditionellen asymptotischen oder exakten p-Werte in der gewählten Modellselektionsprozedur ersetzen. Dieses neue Verfahren kann allgemein verwendet werden, falls die interessierende Menge möglicher log-linearer Modelle eine hierarchische Struktur aufweist. Als weiteres Beispiel werden die in Kapitel 4.1 eingeführten Symmetriemodelle betrachtet.

Anhand einer drei-dimensionalen Kontingenztafel wird zunächst die Modellselektion gemäß „herkömmlicher" algebraischer Tests beschrieben. Dazu seien drei kategoriale Variablen X_1, X_2 und X_3 mit I_1, I_2 bzw. I_3 möglichen Realisationen gegeben. Die Menge der interessierenden graphischen Modelle ist in Abbildung 5.1 dargestellt. Das saturierte Modell wird nicht betrachtet, da Anpassungstests für dieses Modell immer eine Anpassungsgüte von Eins ergeben. Die entsprechenden log-linearen Modelle mit zugehörigen suffizienten Statistiken können Tabelle 5.1 entnommen werden.

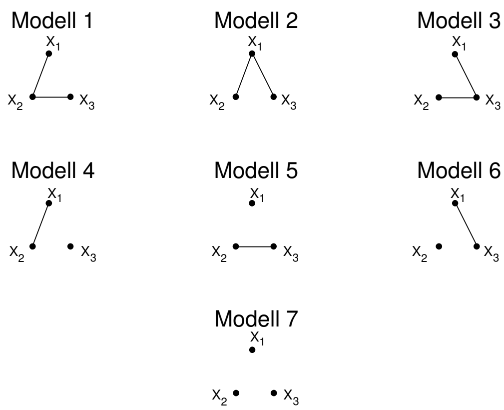

Abbildung 5.1: Graphische Modelle für 3-dimensionale Kontingenztafeln.

Bemerkung 5.2.1

Für eine $I_1 \times I_2 \times I_3$-Kontingenztafel werden die graphischen Modelle anhand log-linearer Modelle, wie in Tabelle 5.1 gegeben, beschrieben. Die Anwendung des Diaconis-Sturmfels-Algorithmus erfordert die Spezifizierung der jeweiligen $T^{*(i)}(X)$. Diese sind Vektoren derselben Länge wie die zugehörige suffiziente Statistik $T^{(i)}(X)$, wobei der Index i, $i = 1, \ldots, 7$, das betrachtete Modell kennzeichnet. Gemäß dem zugrunde liegenden $T^{(i)}(X)$ kann $T^{*(i)}(X)$ in verschiedene Teile aufgespalten werden. So setzt sich beispiels-

weise $T^{*(7)}(X)$ aus drei Teilen der Längen I_1, I_2 und I_3 zusammen. $T^{*(7)}((i_1, i_2, i_3))$ enthält an den Stellen i, $I_1 + i_2$ und $I_1 + I_2 + i_3$ jeweils eine Eins, ansonsten Nullen. Die Herleitungen der übrigen $T^{*(i)}$, $i = 1, \ldots, 6$, sind analog, siehe Anhang A.3 für ein Beispiel.

Modell 1	$\log(m_{i_1i_2i_3}) = \lambda + \lambda_{i_1}^{X_1} + \lambda_{i_2}^{X_2} + \lambda_{i_3}^{X_3} + \lambda_{i_1i_2}^{X_1X_2} + \lambda_{i_2i_3}^{X_2X_3}$
	$T^{(1)} = (N_{i_1i_2+}, i_1 = 1, \ldots, I_1,\ i_2 = 1, \ldots, I_2,\ N_{+i_2i_3}, i_2 = 1, \ldots, I_2,\ i_3 = 1, \ldots, I_3)'$
Modell 2	$\log(m_{i_1i_2i_3}) = \lambda + \lambda_{i_1}^{X_1} + \lambda_{i_2}^{X_2} + \lambda_{i_3}^{X_3} + \lambda_{i_1i_2}^{X_1X_2} + \lambda_{i_1i_3}^{X_1X_3}$
	$T^{(2)} = (N_{i_1i_2+}, i_1 = 1, \ldots, I_1,\ i_2 = 1, \ldots, I_2,\ N_{i_1+i_3}, i_1 = 1, \ldots, I_1,\ i_3 = 1, \ldots, I_3)'$
Modell 3	$\log(m_{i_1i_2i_3}) = \lambda + \lambda_{i_1}^{X_1} + \lambda_{i_2}^{X_2} + \lambda_{i_3}^{X_3} + \lambda_{i_1i_3}^{X_1X_3} + \lambda_{i_2i_3}^{X_2X_3}$
	$T^{(3)} = (N_{i_1+i_3}, i_1 = 1, \ldots, I_1,\ i_3 = 1, \ldots, I_3,\ N_{+i_2i_3}, i_2 = 1, \ldots, I_2,\ i_3 = 1, \ldots, I_3)'$
Modell 4	$\log(m_{i_1i_2i_3}) = \lambda + \lambda_{i_1}^{X_1} + \lambda_{i_2}^{X_2} + \lambda_{i_3}^{X_3} + \lambda_{i_1i_2}^{X_1X_2}$
	$T^{(4)} = (N_{i_1i_2+}, i_1 = 1, \ldots, I_1,\ i_2 = 1, \ldots, I_2,\ N_{++i_3}, i_3 = 1, \ldots, I_3)'$
Modell 5	$\log(m_{i_1i_2i_3}) = \lambda + \lambda_{i_1}^{X_1} + \lambda_{i_2}^{X_2} + \lambda_{i_3}^{X_3} + \lambda_{i_2i_3}^{X_2X_3}$
	$T^{(5)} = (N_{i_1++}, i_1 = 1, \ldots, I_1,\ N_{+i_2i_3}, i_2 = 1, \ldots, I_2,\ i_3 = 1, \ldots, I_3)'$
Modell 6	$\log(m_{i_1i_2i_3}) = \lambda + \lambda_{i_1}^{X_1} + \lambda_{i_2}^{X_2} + \lambda_{i_3}^{X_3} + \lambda_{i_1i_3}^{X_1X_3}$
	$T^{(6)} = (N_{+i_2+}, i_2 = 1, \ldots, I_2,\ N_{i_1+i_3}, i = 1, \ldots, I_1,\ i_3 = 1, \ldots, I_3)'$
Modell 7	$\log(m_{i_1i_2i_3}) = \lambda + \lambda_{i_1}^{X_1} + \lambda_{i_2}^{X_2} + \lambda_{i_3}^{X_3}$
	$T^{(7)} = (N_{i_1++}, i_1 = 1, \ldots, I_1,\ N_{+i_2+}, i_2 = 1, \ldots, I_2,\ N_{++i_3}, i_3 = 1, \ldots, I_3)'$
	$i_1 = 1, \ldots, I_1,\ i_2 = 1, \ldots, I_2,\ i_3 = 1, \ldots, I_3$

Tabelle 5.1: Hierarchische log-lineare Modelle und zugehörige suffiziente Statistiken für eine drei-dimensionale Kontingenztafel.

Die Gröbner-Basen bezüglich der gradlexikographischen Monomordnung für die Modelle 1–7 sind in Tabelle 5.2 beschrieben. Diese Darstellung ist nicht minimal, d.h. Basispolynome können doppelt vorkommen; diese (doppelten) sind nicht zu berücksichtigen. Zur Verifikation dieser Gröbner-Basen werden die Eigenschaften der Gröbner-Basen des entsprechenden Hilfsideals nachgewiesen, siehe Definition 3.2.6.

$$
\begin{array}{lll}
\text{Modell 1: } \mathscr{G}^{(1)} & x_{i_1i_2i_3}\,x_{i'_1i_2i'_3} - x_{i'_1i_2i_3}\,x_{i_1i_2i'_3}, & \text{für } i_1 = i_3, \\[2ex]
\text{Modell 2: } \mathscr{G}^{(2)} & x_{i_1i_2i_3}\,x_{i_1i'_2i'_3} - x_{i_1i'_2i_3}\,x_{i_1i_2i'_3}, & \text{für } i_2 = i_3, \\[2ex]
\text{Modell 3: } \mathscr{G}^{(3)} & x_{i_1i_2i_3}\,x_{i'_1i'_2i_3} - x_{i_1i'_2i_3}\,x_{i'_1i_2i_3}, & \text{für } i_1 = i_2, \\[2ex]
\text{Modell 4: } \mathscr{G}^{(4)} & x_{i_1i_2i_3}\,x_{i'_1i_2i'_3} - x_{i'_1i_2i_3}\,x_{i_1i_2i'_3}, & \text{für } i_1 = i_3, \\
 & x_{i_1i_2i_3}\,x_{i_1i'_2i'_3} - x_{i_1i'_2i_3}\,x_{i_1i_2i'_3}, & \text{für } i_2 = i_3, \\
 & x_{i_1i_2i_3}\,x_{i'_1i'_2i'_3} - x_{i_1i_2i'_3}\,x_{i'_1i'_2i_3}, & \text{für } i_2 = i_2, \\[2ex]
\text{Modell 5: } \mathscr{G}^{(5)} & x_{i_1i_2i_3}\,x_{i'_1i'_2i_3} - x_{i_1i'_2i_3}\,x_{i'_1i_2i_3}, & \text{für } i_1 = i_2, \\
 & x_{i_1i_2i_3}\,x_{i'_1i_2i'_3} - x_{i'_1i_2i_3}\,x_{i_1i_2i'_3}, & \text{für } i_1 = i_3, \\
 & x_{i_1i_2i_3}\,x_{i'_1i'_2i'_3} - x_{i_1i'_2i'_3}\,x_{i'_1i_2i_3}, & \text{für } i_1 = i_2, \\[2ex]
\text{Modell 6: } \mathscr{G}^{(6)} & x_{i_1i_2i_3}\,x_{i'_1i'_2i_3} - x_{i_1i'_2i_3}\,x_{i'_1i_2i_3}, & \text{für } i_1 = i_2, \\
 & x_{i_1i_2i_3}\,x_{i_1i'_2i'_3} - x_{i_1i'_2i_3}\,x_{i_1i_2i'_3}, & \text{für } i_2 = i_3, \\
 & x_{i_1i_2i_3}\,x_{i'_1i'_2i'_3} - x_{i_1i'_2i_3}\,x_{i'_1i_2i'_3}, & \text{für } i_1 = i_2, \\[2ex]
\text{Modell 7: } \mathscr{G}^{(7)} & x_{i_1i_2i_3}\,x_{i'_1i'_2i_3} - x_{i_1i'_2i_3}\,x_{i'_1i_2i_3}, & \text{für } i_1 = i_2, \\
 & x_{i_1i_2i_3}\,x_{i'_1i_2i'_3} - x_{i'_1i_2i_3}\,x_{i_1i_2i'_3}, & \text{für } i_1 = i_3, \\
 & x_{i_1i_2i_3}\,x_{i'_1i'_2i'_3} - x_{i_1i'_2i'_3}\,x_{i'_1i_2i_3}, & \text{für } i_1 = i_2,\ i_2 < i'_2,\ i_3 < i'_3, \\
 & & \qquad i_1 = i_3,\ i_2 < i'_2,\ i_3 < i'_3, \\
 & & \qquad i_2 = i_3,\ i_1 < i'_1,\ i_2 < i'_2,\ i_3 < i'_3, \\
 & x_{i_1i_2i_3}\,x_{i_1i'_2i'_3} - x_{i_1i'_2i_3}\,x_{i_1i_2i'_3}, & \text{für } i_2 = i_3, \\
 & x_{i_1i_2i_3}\,x_{i'_1i'_2i'_3} - x_{i_1i'_2i_3}\,x_{i'_1i_2i'_3}, & \text{für } i_1 = i_2,\ i_1 < i'_1,\ i_3 < i'_3, \\
 & & \qquad i_1 = i_3,\ i_1 < i'_1,\ i_2 < i'_2,\ i_3 < i'_3, \\
 & & \qquad i_2 = i_3,\ i_1 < i'_1,\ i_2 < i'_2,\ i_3 < i'_3, \\
 & x_{i_1i_2i_3}\,x_{i'_1i'_2i'_3} - x_{i_1i_2i'_3}\,x_{i'_1i'_2i_3}, & \text{für } i_1 = i_2,\ i_1 < i'_1,\ i_2 < i'_2,\ i_3 < i'_3, \\
 & & \qquad i_1 = i_3,\ i_1 < i'_1,\ i_2 < i'_2, \\
 & & \qquad i_2 = i_3,\ i_1 < i'_1,\ i_2 < i'_2. \\
 & \multicolumn{2}{l}{i_1 = 1, \ldots, I_1,\ i_2 = 1, \ldots, I_2 \text{ und } i_3 = 1, \ldots, I_3.}
\end{array}
$$

Tabelle 5.2: Reduzierte Gröbner-Basen für die graphischen Modelle 1–7 einer $I_1 \times I_2 \times I_3$-Kontingenztafel unter Berücksichtigung der gradlexikographischen Monomordnung.

Um die approximativen bzw. exakten Anpassungstests durch algebraische Tests zu ersetzen, wird der Diaconis-Sturmfels-Algorithmus für jedes der betrachteten graphischen Modelle einzeln angewendet. Gemäß den Ausführungen in Kapitel 3.3 wird anschließend für jedes graphische Modell (gegeben in Tabelle 5.1) jeweils eine Markov-Kette der Länge l generiert. Das weitere Vorgehen entspricht dem üblicher MCMC-Prozeduren: Die ersten b Tafeln der Markov-Kette werden ignoriert, anschließend wird für jeden $s$ten Datensatz die Teststatistik des χ^2- bzw. des Likelihood-Quotienten-Tests berechnet. Die resultierenden Testergebnisse ersetzen anschließend die exakten bzw. approximativen p-Werte. Die jeweilige Anzahl der Freiheitsgrade ist die Differenz der Dimension des saturierten und zu testenden Modells. Eine herkömmliche algebraische Modellselektion ist insgesamt simulationsintensiv und mit zunehmender Anzahl interessierender Modelle nicht mehr durchführbar. Nachfolgend wird gezeigt, wie der nötige Simulationsaufwand reduziert werden kann. Auf dieser Basis können anschließend neue algebraische Modellselektionsprozeduren vorgeschlagen werden.

Graphische Modelle für kategoriale Daten werden oftmals als hierarchische log-lineare Modelle dargestellt. Das saturierte log-lineare Modell entspricht einem kompletten graphischen Modell. Es enthält alle weiteren hierarchischen log-linearen Modelle als Spezialfälle, vgl. Kapitel 2. Diese Ordnung kann entsprechend fortgesetzt werden; so schließt beispielsweise Modell 1 aus Abbildung 5.1 die Modelle 4, 5 und 7 ein. Eine solche Hierarchie kann ebenfalls bei den jeweiligen suffizienten Statistiken T und damit auch bei $\mathscr{Z}_t$, den zugehörigen Mengen aller möglichen Datensätze mit Wert t der suffizienten Statistik beobachtet werden. Im Folgenden werden zwei verschiedene (graphische) Modelle $M1$ und $M2$ für denselben Datensatz miteinander verglichen. $M1$ sei ein Untermodell von $M2$, $M1 \subseteq M2$, d.h. die Kantenmenge von $M1$ ist eine Teilmenge der Kantenmenge von $M2$. Daraus folgt, dass die suffiziente Statistik $T^{(M2)}$ für die Parameter des Modells $M2$ die suffiziente Statistik $T^{(M1)}$ enthält. Aufgrund dieser zusätzlichen Restriktionen für $M2$ enthält $\mathscr{Z}_{t(M1)}$ alle Datensätze aus $\mathscr{Z}_{t(M2)}$, d.h. $\mathscr{Z}_{t(M2)} \subseteq \mathscr{Z}_{t(M1)}$. Das Unabhängigkeitsmodell, repräsentiert durch einen Graphen ohne Kanten, ist ein Untermodell von jedem möglichen graphischen Modell. Damit enthält $\mathscr{Z}_{t(7)}$ die entsprechenden Mengen für alle weiteren graphischen Modelle. Diese Hierarchie besteht ebenfalls bei den zugehörigen Gröbner-Basen.

Theorem 5.2.2

Es seien $M1$ und $M2$ zwei verschiedene hierarchische log-lineare Modelle mit $M1 \subseteq M2$. Unter Anwendung des Diaconis-Sturmfels-Algorithmus werden die interessierenden Ideale $\mathscr{I}^{(M1)}$ und $\mathscr{I}^{(M2)}$ bestimmt, dabei gilt $\mathscr{I}^{(M1)} \supseteq \mathscr{I}^{(M2)}$.

Beweis:

Die Menge der Einträge der suffizienten Statistik für die Parameter der Modelle $M1$ und $M2$ seien gemäß Kapitel 3.3 mit $\mathscr{T}^{(M1)}$ und $\mathscr{T}^{(M2)}$ bezeichnet. Aufgrund der hierarchischen Struktur der Modelle ist $\mathscr{T}^{(M1)}$ vollständig durch $\mathscr{T}^{(M2)}$ bestimmt, d. h. $\mathscr{T}^{(M1)} \subseteq \mathscr{T}^{(M2)}$. Aus der Definition 3.1.2 einer Varietät folgt weiter, dass $V^{(M2)}$ in $V^{(M1)}$ enthalten ist, $V^{(M1)} \supseteq V^{(M2)}$, wobei $V^{(M1)}$ und $V^{(M2)}$ die durch den Diaconis-Sturmfels-Algorithmus festgelegten Varietäten der Modelle $M1$ und $M2$ sind (vergleiche Theorem 3.3.4). Dies impliziert $\mathscr{I}(V^{(M1)}) \supseteq \mathscr{I}(V^{(M2)})$, siehe Cox et al. (1997), Proposition 8, S. 34.

□

Für die betrachteten Modelle (gegeben in Tabelle 5.1) gilt $\{\mathscr{T}^{(1)}, \mathscr{T}^{(2)}, \mathscr{T}^{(3)}\} \supseteq \{\mathscr{T}^{(4)}, \mathscr{T}^{(5)}, \mathscr{T}^{(6)}\} \supseteq \mathscr{T}^{(7)}$. Gemäß Theorem 5.2.2 folgt daraus, dass $\mathscr{I}^{(7)} \supseteq \{\mathscr{I}^{(6)}, \mathscr{I}^{(5)}, \mathscr{I}^{(4)}\} \supseteq \{\mathscr{I}^{(3)}, \mathscr{I}^{(2)}, \mathscr{I}^{(1)}\}$. Die Gröbner-Basen für die zugrunde liegenden Modelle sind in Tabelle 5.2 aufgeführt. Die oben beschriebene Anordnung der Ideale äußert sich für die zugehörigen Gröbner-Basen wie folgt: Jede Gröbner-Basis der Modelle 1–6 ist entweder in $\mathscr{G}^{(7)}$ enthalten oder kann durch eine Linearkombination der Basispolynome aus $\mathscr{G}^{(7)}$ dargestellt werden.

Beispiel 5.2.3

Gegeben sei eine $2 \times 2 \times 2$ Kontingenztafel. Die Gröbner-Basen bezüglich der gradlexikographischen Monomordnung sind für die zugrunde liegenden Modelle (vgl. Tabelle 5.1) gegeben durch:

Modell 1: $\mathscr{G}^{(1)} = \{g_1^{(1)}, g_2^{(1)}\}$ *mit*

$g_1^{(1)} = x_{121}\, x_{222} - x_{122}\, x_{221}, \quad g_2^{(1)} = x_{111}\, x_{212} - x_{112}\, x_{211};$

Modell 2: $\mathscr{G}^{(2)} = \{g_1^{(2)}, g_2^{(2)}\}$ *mit*

$g_1^{(2)} = x_{211}\, x_{222} - x_{212}\, x_{221}, \quad g_2^{(2)} = x_{111}\, x_{122} - x_{112}\, x_{121};$

Modell 3: $\mathscr{G}^{(3)} = \{g_1^{(3)}, g_2^{(3)}\}$ *mit*

$g_1^{(3)} = x_{112}\, x_{222} - x_{122}\, x_{212}, \quad g_2^{(3)} = x_{111}\, x_{221} - x_{121}\, x_{211};$

Modell 4: $\mathscr{G}^{(4)} = \{g_1^{(4)}, g_2^{(4)}, g_3^{(4)}, g_4^{(4)}, g_5^{(4)}, g_6^{(4)}\}$ *mit*

$g_1^{(4)} = x_{211}\, x_{222} - x_{212}\, x_{221}, \quad g_2^{(4)} = x_{121}\, x_{222} - x_{122}\, x_{221}, \quad g_3^{(4)} = x_{111}\, x_{222} - x_{112}\, x_{221},$

$g_4^{(4)} = x_{121}\, x_{212} - x_{122}\, x_{211}, \quad g_5^{(4)} = x_{111}\, x_{212} - x_{112}\, x_{211}, \quad g_6^{(4)} = x_{111}\, x_{122} - x_{112}\, x_{121};$

Modell 5: $\mathscr{G}^{(5)} = \{g_1^{(5)}, g_2^{(5)}, g_3^{(5)}, g_4^{(5)}, g_5^{(5)}, g_6^{(5)}\}$ *mit*

$g_1^{(5)} = x_{121}\, x_{222} - x_{122}\, x_{221}, \quad g_2^{(5)} = x_{112}\, x_{222} - x_{122}\, x_{212}, \quad g_3^{(5)} = x_{112}\, x_{221} - x_{121}\, x_{212},$

$g_4^{(5)} = x_{111}\, x_{222} - x_{122}\, x_{211}, \quad g_5^{(5)} = x_{111}\, x_{221} - x_{121}\, x_{211}, \quad g_6^{(5)} = x_{111}\, x_{212} - x_{112}\, x_{211};$

Modell 6: $\mathscr{G}^{(6)} = \{g_1^{(6)}, g_2^{(6)}, g_3^{(6)}, g_4^{(6)}, g_5^{(6)}, g_6^{(6)}\}$ *mit*

$g_1^{(6)} = x_{211}\, x_{222} - x_{212}\, x_{221}, \quad g_2^{(6)} = x_{112}\, x_{222} - x_{122}\, x_{212}, \quad g_3^{(6)} = x_{112}\, x_{221} - x_{122}\, x_{211},$

$g_4^{(6)} = x_{111}\, x_{222} - x_{121}\, x_{212}, \quad g_5^{(6)} = x_{111}\, x_{221} - x_{121}\, x_{211}, \quad g_6^{(6)} = x_{111}\, x_{122} - x_{112}\, x_{121};$

Modell 7: $\mathscr{G}^{(7)} = \{g_1^{(7)}, g_2^{(7)}, g_3^{(7)}, g_4^{(7)}, g_5^{(7)}, g_6^{(7)}, g_7^{(7)}, g_8^{(7)}, g_9^{(7)}\}$ *mit*

$g_1^{(7)} = x_{121}\, x_{222} - x_{122}\, x_{221}, \quad g_2^{(7)} = x_{112}\, x_{222} - x_{122}\, x_{212}, \quad g_3^{(7)} = x_{111}\, x_{222} - x_{122}\, x_{211},$

$g_4^{(7)} = x_{211}\, x_{222} - x_{212}\, x_{221}, \quad g_5^{(7)} = x_{111}\, x_{222} - x_{121}\, x_{212}, \quad g_6^{(7)} = x_{111}\, x_{222} - x_{112}\, x_{221},$

$g_7^{(7)} = x_{111}\, x_{212} - x_{112}\, x_{211}, \quad g_8^{(7)} = x_{111}\, x_{122} - x_{112}\, x_{121}, \quad g_9^{(7)} = x_{111}\, x_{221} - x_{121}\, x_{211}.$

Alle Basispolynome von $\mathscr{G}^{(1)}, \mathscr{G}^{(2)}$ *und* $\mathscr{G}^{(3)}$ *sind Element von* $\mathscr{G}^{(7)}$*; die Gröbner-Basen der Modelle 4–6 enthalten jeweils ein Polynom, das nicht direkt in* $\mathscr{G}^{(7)}$ *enthalten ist, aber durch* $\mathscr{G}^{(7)}$ *dargestellt werden kann:*

$g_4^{(4)} = x_{121}\, x_{212} - x_{122}\, x_{211} = x_{111}\, x_{222} - x_{122}\, x_{211} - (x_{111}\, x_{222} - x_{121}\, x_{212}) = g_3^{(7)} - g_5^{(7)},$

$g_3^{(5)} = x_{112}\, x_{221} - x_{121}\, x_{212} = x_{111}\, x_{222} - x_{121}\, x_{212} - (x_{211}\, x_{222} - x_{212}\, x_{221}) = g_5^{(7)} - g_4^{(7)},$

$g_3^{(6)} = x_{112}\,x_{221} - \, x_{122}\,x_{211} = x_{111}\,x_{222} - x_{122}\,x_{211} - (x_{211}\,x_{222} - \, x_{212}\,x_{221}) = g_3^{(7)} - g_4^{(7)}.$

Eine gemäß dem Diaconis-Sturmfels-Algorithmus konstruierte Markov-Kette für das Modell 7 kann also jeden möglichen Zustand aller weiteren Modelle erreichen. Damit können für jedes betrachtete Modell einzelne Ketten aus der Markov-Kette von Modell 7 extrahiert werden. Die Selektion erfolgt gemäß der realisierten suffizienten Statistik des betrachteten Datensatzes für die Parameter des interessierenden Modells. Stimmt diese mit dem jeweiligen Wert der suffizienten Statistik einer generierten Kontingenztafel überein, so ist dieser Zustand der Markov-Kette Element der „selektierten Kette“. Unter der Annahme, dass die hypergeometrische Verteilung auf $\mathscr{Z}_{t^{(7)}}$ durch die Simulation gemäß Modell 7 geeignet angenähert wird, geben die selektierten Ketten eine passende Approximation der entsprechenden bedingten Verteilungen wieder, d.h.

$$P((N_x)_{x\in\mathscr{H}} = (n_x)_{x\in\mathscr{H}} | T^{(7)} = t^{(7)}) \approx \frac{|\text{simulierte Tafeln gleich } (n_x)_{x\in\mathscr{H}}|}{|\text{simulierte Tafeln}|}$$

für alle realisierten Datensätze $(n_x)_{x\in\mathscr{H}} \in \{(n_x)_{x\in\mathscr{H}} | n_x \geq 0, \sum_{x\in\mathscr{H}} n_x = n\}$. Für die Modelle $i = 1, \ldots, 6$ gilt $\{(n_x)_{x\in\mathscr{H}} | T^{(i)} = t^{(i)}\} \subseteq \{(n_x)_{x\in\mathscr{H}} | T^{(7)} = t^{(7)}\}$, wobei $t^{(7)}$ und $t^{(i)}$ von derselben beobachteten Kontingenztafel berechnet werden. Damit gilt für alle $i = 1, \ldots, 6$:

$$\begin{aligned}
& P((N_x)_{x\in\mathscr{H}} = (n_x)_{x\in\mathscr{H}} | T^{(i)} = t^{(i)}) \\
= {} & P((N_x)_{x\in\mathscr{H}} = (n_x)_{x\in\mathscr{H}} | T^{(i)} = t^{(i)} \cap T^{(7)} = t^{(7)}) \\
= {} & \frac{P((N_x)_{x\in\mathscr{H}} = (n_x)_{x\in\mathscr{H}} \cap T^{(i)} = t^{(i)} | T^{(7)} = t^{(7)})}{P(T^{(i)} = t^{(i)} | T^{(7)} = t^{(7)})} \\
\approx {} & \frac{|\text{simulierte Tafeln gleich } (n_x)_{x\in\mathscr{H}} \text{ und mit } T^{(i)} = t^{(i)}|}{|\text{simulierte Tafeln mit } T^{(i)} = t^{(i)}|}.
\end{aligned}$$

Die Simulation einer Markov-Kette gemäß dem Unabhängigkeitsmodell (Modell 7, vgl. Tabelle 5.1) ist ausreichend für die Analyse der Abhängigkeitsstrukturen graphischer Modelle bei kategorialen Daten. Selektierte Ketten ersetzen nun die einzeln simulierten Markov-Ketten für alle betrachteten Modelle. Allerdings hängt die Approximation der bedingten Verteilungen und damit auch der Vorteil des neuen Modellselektionsverfahrens von der Anzahl der generierten Tafeln mit $T^{(i)} = t^{(i)}$, $i = 1, \ldots, 6$, ab. Bei zunehmendem Stichprobenumfang sinkt diese Anzahl. Um die Annäherung dann zu gewährleisten, muss die Kettenlänge der generierten Markov-Kette gemäß Modell 7 angepasst werden.

Korollar 5.2.4 (Neue algebraische Modellselektion)
Es interessiere die Untersuchung der Abhängigkeitsstrukturen für kategoriale Daten, das Modellwahlkriterium sei ein Anpassungstest. Können die interessierenden log-linearen Modelle wie die betrachteten graphischen Modelle hierarchisch angeordnet werden, so wird das folgende Vorgehen vorgeschlagen:

Schritt 1 *Generiere eine Markov-Kette unter Verwendung der Gröbner-Basis des Spezialfalls aller zugrunde liegenden Modelle. Bei den graphischen Modellen aus Tabelle 5.1 ist dies das Unabhängigkeitsmodell.*

Schritt 2 *Extrahiere für die weiteren Modelle die möglichen Datensätze anhand der Werte der realisierten suffizienten Statistiken. Die so entstandenen Ketten werden „selektierte Ketten“ genannt.*

Schritt 3 *Berechne die algebraischen p-Werte anhand der selektierten Ketten.*

Schritt 4 *Verwende die p-Werte aus Schritt 3 anstelle der approximativen oder exakten p-Werte in den Modellwahlstrategien basierend auf Anpassungstests, vgl. z. B. Kapitel 5.1.*

Die in Kapitel 4.1 beschriebenen Symmetriemodelle können ebenfalls hierarchisch angeordnet werden. Insbesondere ist das perfekte Symmetriemodell ein Spezialfall des bedingten, diagonalen sowie des ordinalen quasi-Symmetriemodells. Als solches ist die suffiziente Statistik $T^{(S)}$ für die Parameter des S Modells in den suffizienten Statistiken $T^{(CS)}$, $T^{(DS)}$ und $T^{(OQS)}$ enthalten; daher gilt $\mathscr{X}_{t(S)} \supseteq \{\mathscr{X}_{t(CS)}, \mathscr{X}_{t(DS)}, \mathscr{X}_{t(OQS)}\}$. Die Argumentation zur Entwicklung einer neuen algebraischen Modellselektionsprozedur für graphische Modelle gilt folglich auch für die betrachteten Symmetriemodelle. Gemäß Korollar 5.2.4 wird die Gröbner-Basis des perfekten Symmetriemodells für die Generierung einer Markov-Kette verwendet. Mit den beobachteten suffizienten Statistiken der weiteren Modelle werden anschließend die extrahierten Ketten erstellt. Die Berechnung der neuen algebraischen p-Werte erfolgt anhand dieser Ketten. Die resultierenden Symmetrietests werden an verschiedenen Datenbeispielen vorgeführt, siehe Kapitel 5.4.

5.3 Simulationsstudie

In diesem Abschnitt wird insbesondere die herkömmliche Diaconis-Sturmfels-Prozedur für die Abhängigkeitsanalyse graphischer Modelle kategorialer Daten mit dem vorgeschlagenen algebraischen Modellselektionsverfahren verglichen. Die gewählten Einstellungen des Simulationsdesigns zeigen gutes Konvergenzverhalten der Markov-Kette; eine Diskussion der Konvergenzrate kann z. B. in Diaconis und Sturmfels (1998) nachgelesen werden. Für eine Anwendung der Diaconis-Sturmfels-Prozedur wird eine Markov-Kette der Länge 500.000 generiert; die Einschwingphase umfasst 50.000 Datensätze. Von den übrigen Tafeln wird jede 100ste Tafel für die Berechnung des algebraischen p-Wertes berücksichtigt. Für das neue Modellselektionsverfahren wird folgendes Simulationsdesign gewählt: Gemäß dem Unabhängigkeitsmodell (Modell 7, Tabelle 5.1) wird eine Markov-Kette mit 1.000.000 Zuständen simuliert, anschließend werden für die interessierenden Modelle die selektierten Ketten extrahiert. Die Anordnung der Zustände in einer selektierten Kette ist üblicherweise verschieden von der Anordnung der Zustände in der ursprünglichen Markov-Kette; daher werden nur noch die ersten 10 Kontingenztafeln der selektierten Kette vernachlässigt, von den übrigen Datensätzen wird nun jede 10te Tafel weiter berücksichtigt. Für den Vergleich beider Methoden ist ferner zu beachten, dass die Länge einer selektierten Kette vorher nicht festgelegt ist.

Im ersten Teil dieser Simulationsstudie werden die gemäß Theorem 3.3.4 sowie Korollar 5.2.4 gewonnenen (neuen) algebraischen Wahrscheinlichkeiten mit den entsprechenden theoretischen hypergeometrischen Wahrscheinlichkeiten aller Datensätze aus $\mathscr{Z}_{t(i)}$, $i = 1, \ldots, 7$, verglichen. Dazu wird Datensatz 5.1, gegeben in Tabelle 5.3, betrachtet.

	$i_2 = 1$	$i_2 = 2$
$i_1 = 1$	2	1
$i_1 = 2$	2	3

$i_3 = 1$

	$i_2 = 1$	$i_2 = 2$
$i_1 = 1$	1	2
$i_1 = 2$	0	3

$i_3 = 2$

Tabelle 5.3: Datensatz 5.1.

Die Mengen aller möglichen Datensätze mit gleicher suffizienter Statistik wie Datensatz 5.1 umfassen $|\mathscr{Z}_{t(1)}| = 8$, $|\mathscr{Z}_{t(2)}| = 12$, $|\mathscr{Z}_{t(3)}| = 8$, $|\mathscr{Z}_{t(4)}| = 44$, $|\mathscr{Z}_{t(5)}| = 40$, $|\mathscr{Z}_{t(6)}| = 44$ sowie $|\mathscr{Z}_{t(7)}| = 194$ Kontingenztafeln. Da die neue (d. h. „selektierte“) algebraische Verteilung auf $\mathscr{Z}_{t(i)}$, $i = 1, \ldots, 6$, mit der entsprechenden exakten hypergeometrischen Verteilung überwiegend übereinstimmt, wird die Qualität des neuen Modellselektionsverfahrens als gut bewertet. Für die Modelle 2–6 werden erst ab der dritten Nachkommastelle Unterschiede beobachtet. Für das erste Modell beträgt die maximale absolute Differenz zwischen exakter und neuer algebraischer Wahrscheinlichkeit 0,017. Ein Vergleich der herkömmlichen algebraischen Wahrscheinlichkeiten mit den exakten zeigt, dass auch hier der maximale Unterschied 0,010 beträgt. Dieser Wert ist einerseits noch „klein genug“, andererseits erweist er sich nicht als Besonderheit des neuen Simulationsvorgehens, so dass zusammenfassend die vorgeschlagene Modellselektionsprozedur die hypergeometrischen Verteilungen auf $\mathscr{Z}_{t(i)}$, $i = 1, \ldots, 6$ adäquat simuliert. In Tabelle 5.4 sind die exakten, die algebraischen nach Diaconis und Sturmfels sowie die neuen algebraischen Wahrscheinlichkeiten beispielhaft für das Modell 2 aufgeführt. Die Ergebnisse der weiteren Modelle sind in Anhang A.3 angegeben.

Wahrscheinlichkeit		
exakt	algebraisch	neues Verfahren
0,241	0,231	0,240
0,161	0,171	0,155
0,241	0,232	0,242
0,161	0,164	0,162
0,048	0,051	0,052
0,027	0,022	0,027
0,048	0,049	0,049
0,018	0,018	0,019
0,027	0,030	0,029
0,018	0,022	0,017
0,005	0,006	0,005
0,005	0,004	0,004

Tabelle 5.4: Exakte hypergeometrische und algebraische Wahrscheinlichkeiten für alle Elemente von $\mathscr{Z}_{t(2)}$ des Datensatzes 5.1.

Im zweiten Teil der Simulationsstudie interessiert insbesondere ein Vergleich der herkömmlichen mit den neuen algebraischen p-Werten. Dazu werden für den $2 \times 2 \times 2$-Fall sowohl von Modell 3 als auch von Modell 4 jeweils 100 Datensätze zufällig generiert. Die erwarteten Zellwahrscheinlichkeiten können Tabelle 5.5 entnommen werden; der betrachtete Stichprobenumfang beträgt jeweils $n = 35$.

a)

$i_3 = 1$

	$i_2 = 1$	$i_2 = 2$
$i_1 = 1$	0,14286	0,25397
$i_1 = 2$	0,11429	0,20317

$i_3 = 2$

	$i_2 = 1$	$i_2 = 2$
$i_1 = 1$	0,20000	0,20000
$i_1 = 2$	0,08571	0,08571

b)

$i_3 = 1$

	$i_2 = 1$	$i_2 = 2$
$i_1 = 1$	0,14423	0,22857
$i_1 = 2$	0,14423	0,14423

$i_3 = 2$

	$i_2 = 1$	$i_2 = 2$
$i_1 = 1$	0,08571	0.,17143
$i_1 = 2$	0,08571	0,08571

Tabelle 5.5: Erwartete Zellwahrscheinlichkeiten für die interessierenden Simulationsmodelle. a) Modell 3, b) Modell 4.

Am Beispiel der Datensätze des Simulationsmodells 3 werden die Eigenschaften der vorgeschlagenen Modellselektionsprozedur untersucht. Dabei dienen die Ergebnisse des herkömmlichen algebraischen Tests nach Theorem 3.3.4 als Vergleichskriterium. Die Resultate für die Datensätze des Simulationsmodells 4 sind analog und werden daher nicht weiter beschrieben, vergleiche Krampe und Kuhnt (2008) und Anhang A.3.

Für jeden der 100 zufällig generierten Datensätze wird entsprechend dem Unabhängigkeitsmodell eine Markov-Kette der Länge 1.000.000 generiert. Aus diesen werden anschließend die selektierten Ketten für die Modelle 1–6 extrahiert, vergleiche Korollar 5.2.4. Die resultierende Anzahl von Tafeln mit gleicher beobachteter suffizienter Statistik ist in Abbildung 5.2 dargestellt.

Auffällig ist, dass für die Modelle 4–6 überwiegend mehr Datensätze selektiert werden als für die Modelle 1–3. Dies ist mit den jeweiligen suffizienten Statistiken zu begründen, denn diese dienen als Selektionskriterium. Für eine $2 \times 2 \times 2$-Kontingenztafel müssen für die Modelle 1–3 insgesamt acht Werte mit denen des beobachteten Datensatzes übereinstimmen; für die Modelle 4–6 sind es hingegen nur sechs, vergleiche Tabelle 5.1. Außerdem ist die Variabilität der Anzahl selektierter Datensätze für die Modelle 1–3 sehr viel geringer als für die Modelle 4–6. Die Ursache hierfür scheinen die in $\mathscr{G}^{(7)}$ fehlenden Basispolynome $g_4^{(4)}$, $g_3^{(5)}$ und $g_3^{(6)}$ zu sein, vgl. Beispiel 5.2.3. Ist die Kettenlänge der gemäß Modell 7 generierten Markov-Kette adäquat gewählt, so werden „ausreichend viele“ Datensätze für die zugrunde liegenden Modelle selektiert. Die Güte des vorgeschlagenen Modellselektionsverfahrens ist damit trotz der größeren

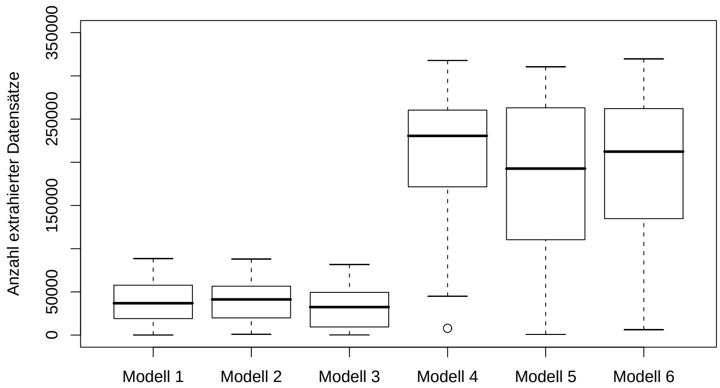

Abbildung 5.2: Boxplots der Anzahl selektierter Datensätze der Modelle 1–6 bei zugrunde liegendem Simulationsmodell 3.

Variabilität in der Anzahl extrahierter Kontingenztafeln sichergestellt.

Anschließend werden die Abhängigkeitsstrukturen der 100 zufällig aus Modell 3 generierten Datensätze untersucht. Dazu werden die Anpassungstests gemäß Diaconis und Sturmfels als auch die vorgeschlagenen algebraischen Tests durchgeführt. In Abbildung 5.3 sind die p-Werte beider Prozeduren als Wertepaare abgetragen. Diese ordnen sich auf oder sehr nahe der Winkelhalbierenden an. Beide algebraische Testmethoden liefern also zumeist fast identische Ergebnisse. So stimmen alle Testentscheidungen beider Tests zum Niveau $\alpha = 0,05$ überein. Damit erscheinen die neuen algebraischen p-Werte geeignet, in einem gewählten Modellselektionsverfahren die Anpassungsgüte eines Modells adäquat zu beurteilen.

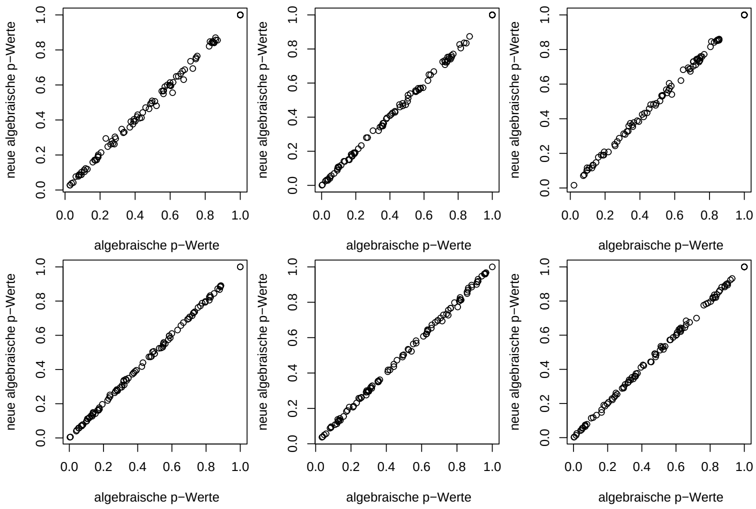

Abbildung 5.3: Streudiagramme der p-Werte des χ^2-Anpassungstests (algebraische p-Werte nach Diaconis und Sturmfels und dem neuen Verfahren) für die Modelle 1–6 (beginnend oben von links nach rechts) bei zugrunde liegendem Simulationsmodell 3.

5.4 Datenbeispiele

In diesem Abschnitt werden zunächst die Abhängigkeitsstrukturen zweier Datensätze anhand graphischer Modelle analysiert. Anschließend wird die algebraische Modellselektion auf die in Kapitel 4.1 vorgestellten Symmetrietests angewendet. Entscheidungsgrundlage für die durchgeführten Anpassungstests sind die algebraischen, die neuen algebraischen und die approximativen p-Werte. Das Signifikanzniveau sei auf $\alpha = 0,10$ festgelegt.

In der Zeitung The New York Times wurde 1991 eine Studie veröffentlicht, die den Effekt einer Antiretroviralen Medizin (Azidohymidin, AZT) auf die Entwicklung von AIDS-Symptomen untersucht. Die Daten umfassen $n = 338$ Probanden, deren Immunsysteme erste Symptome von AIDS aufweisen. Diese Patienten wurden zufällig in zwei Gruppen unterteilt: Die erste Gruppe erhielt unverzüglich Arzneimittel, bei der zweiten Gruppe wartete man mit der Medikation, bis das Immunsystem von dem HI-Virus angegriffen wurde, vgl. Agresti (2002), S. 184. Die Behandlung wird durch $X_1 = i_1$ repräsentiert mit $i_1 = 1$ sofortiger Einsatz von AZT, $i_1 = 2$ sonst. Der Krankheitsstatus eines Probanden wird durch $X_2 = i_2$ dargestellt, und es ist $i_2 = 1$, falls AIDS-Symptome entwickelt sind, und $i_2 = 2$, falls keine AIDS-Symptome erkennbar sind. Zudem wurde die Hautfarbe des Patienten $X_3 = i_3$ erhoben; $i_3 = 1$ bedeutet weiß und $i_3 = 2$ schwarz. Die Daten sind in Tabelle 5.6 zusammengefasst.

	$i_2 = 1$	$i_2 = 2$
$i_1 = 1$	14	93
$i_1 = 2$	32	81

$i_3 = 1$

	$i_2 = 1$	$i_2 = 2$
$i_1 = 1$	11	52
$i_1 = 2$	12	43

$i_3 = 2$

Tabelle 5.6: Datensatz 5.2: AIDS-Datensatz, vergleiche Agresti (2002).

Der Stichprobenumfang ist mit $n = 338$ Probanden im Vergleich zu den vorher betrachteten Datensätzen groß. Die Mächtigkeiten der $\mathscr{Z}_{t(i)}$, $i = 1, \ldots, 7$, der Mengen aller möglichen Datensätze mit derselben suffizienten Statistik wie die beobachtete Tafel sind entsprechend umfangreich. Um sicherzustellen, dass die Markov-Kette konvergiert, werden die Simulationsparameter wie folgt angepasst: Für die Diaconis-Sturmfels-Prozedur werden pro Modell 800.000 Zustände generiert; für das neue Verfahren wird gemäß Modell 7 eine Markov-Kette mit 1.600.000 Tafeln simuliert. Die Einstellungen für die Einschwingphase sowie für die Schrittlänge bleiben unverändert, d. h., es werden die ersten 50.000 (herkömmliches Diaconis-Sturmfels-Verfahren) bzw. 10 (neues Verfahren) Zustände vernachlässigt und anschließend jede 100ste bzw. jede 10te Tafel weiter berücksichtigt. Die vorgeschlagene algebraische Modellselektion basiert für Modell 1 auf 4.113, für Modell 2 auf 2.545 und für Modell 3 auf 83.018 Kontingenztafeln. Weiter werden 40.222, 1.298.307 bzw. 767.278 Datensätze für die Modelle 4–6 selektiert.

Zur Illustration sind in Tabelle 5.7 die p-Werte für alle betrachteten Modelle aufgeführt. Es zeigt sich, dass die Testergebnisse der algebraischen Anpassungstests mit den Testergebnissen der neuen Prozedur fast übereinstimmen. Die Modellselektion anhand der χ^2-Approximation scheint für diesen Datensatz ebenfalls geeignet zu sein.

Beispielhaft wird für diesen Datensatz die Edwards-Havránek-Prozedur durchgeführt. Als Startmodell wird das Haupteffekt-Modell, Modell 7 aus Tabelle 5.1, gewählt. Im ersten Schritt der EH-Prozedur wird festgestellt, dass das Modell 7 zum Niveau $\alpha = 0,10$ abgelehnt werden kann. Anschließend werden die Modelle mit jeweils einer Wechselwirkung, Modelle 4, 5, 6, getestet (Schritt 2b). Es zeigt sich, dass das

	p-Werte		
	approximativ	algebraisch	neues Verfahren
Modell 1	0,359	0,365	0,361
Modell 2	0,493	0,496	0,490
Modell 3	0,018	0,021	0,017
Modell 4	0,552	0,497	0,550
Modell 5	0,033	0,033	0,031
Modell 6	0,040	0,037	0,039
Modell 7	0,060	0,058	0,059

Tabelle 5.7: Approximative und algebraische (nach Diaconis-Sturmfels und dem neuen Verfahren) Ergebnisse des χ^2-Anpassungstest für den AIDS-Datensatz.

Modell 4 akzeptabel ist, die beiden übrigen jedoch abgelehnt werden können. Im nächsten Schritt (Schritt 2a) ist Modell 3 zu testen, welches zum gewählten Niveau $\alpha = 0,10$ abgelehnt werden kann; damit bricht die Modellselektions-Prozedur ab. Die approximative, algebraische und neue algebraische EH-Prozedur schlägt somit Modell 4 für die Beschreibung der Abhängigkeitsstrukturen in dem AIDS-Datensatz vor, die Modelle 1 und 2 sind schwach akzeptabel. Die Medikation mit AZT scheint demnach abhängig von dem jeweiligen Krankheitsstatus zu sein. Ein Zusammenhang zwischen der Behandlung mit AZT und der Hautfarbe des Patienten konnte nicht eindeutig festgestellt werden.

In dem zweiten Datensatz werden an $n = 97$ zehnjährigen Schulkindern insgesamt drei Merkmale beobachtet. Lehrer beurteilen das Verhalten der Kinder im Klassenzimmer $X_1 = i_1$, wobei $i_1 = 1$ normales und $i_1 = 2$ auffälliges Verhalten kodiert. Die häuslichen Umstände der Kinder werden durch $X_2 = i_2$ charakterisiert. Anhand verschiedener Faktoren wie beispielsweise der Größe der Familie werden die häuslichen Umstände klassifiziert als $i_2 = 1$ nicht gefährdend und $i_2 = 2$ gefährdend. Die schulischen Umstände $X_3 = i_3$ werden aufgrund der Anzahl freier Mahlzeiten, der Fluktuation in der Schülerzahl, etc. bewertet als $i_3 = 1$ gut, $i_3 = 2$ mittel und $i_3 = 3$ schlecht. Die Kontingenztafel ist in Tabelle 5.8 angegeben.

Für eine $2 \times 2 \times 3$-Kontingenztafel enthält die Gröbner-Basis $\mathscr{G}^{(7)}$ mehr mögliche Bewegungen als für die bisher betrachteten $2 \times 2 \times 2$-Tafeln. Daher wird das Simulationsdesign wie folgt angepasst: Für die Diaconis-Sturmfels-Prozedur wird eine Markov-Kette der

	$i_2 = 1$	$i_2 = 2$
$i_1 = 1$	16	7
$i_1 = 2$	1	1

$i_3 = 1$

	$i_2 = 1$	$i_2 = 2$
$i_1 = 1$	15	34
$i_1 = 2$	3	8

$i_3 = 2$

	$i_2 = 1$	$i_2 = 2$
$i_1 = 1$	5	3
$i_1 = 2$	1	3

$i_3 = 3$

Tabelle 5.8: Datensatz 5.3: Schulkinder-Datensatz, vergleiche Everitt (1977), S. 67.

Länge 800.000 generiert. Für das vorgeschlagene Verfahren wird eine Markov-Kette mit 4.000.000 Zuständen generiert. Für Modell 1 werden insgesamt 113, für Modell 2 hingegen 5.973 und für Modell 3 nur 21 Kontingenztafeln selektiert. Für die Modelle 4–6 werden 443.558, 969 bzw. 54.807 Datensätze extrahiert. Insgesamt ist auffällig, dass die Anzahl selektierter Datensätze stark variiert, insbesondere werden nur sehr wenige Kontingenztafeln generiert, die aus $\mathscr{Z}_{t^{(3)}}$ stammen. Ursächlich hierfür sind zum einen die suffizienten Statistiken als Selektionskriterium; für die Modelle 1–6 ist $T^{(3)}$ mit 12 Einträgen der längste Vektor. Entsprechend besteht die Gröbner-Basis $\mathscr{G}^{(3)}$ aus nur drei Basispolynomen, vgl. Tabelle 5.2. Die Anzahl vorgeschlagener Bewegungen ist damit für Modell 3 am geringsten. Zum anderen sind diese Bewegungsmöglichkeiten zusätzlich durch geringe Zelleinträge des betrachteten Schulkinder-Datensatzes beeinträchtigt. Aufgrund der z. T. geringen Länge der extrahierten Markov-Kette wird hier für die selektierte Simulation die Schrittlänge auf Eins gesetzt. Für einen Vergleich der verwendeten Methoden sind in Tabelle 5.9 die p-Werte aller interessierenden Modelle zusammengefasst. Wie erwartet scheint das neue Modellselektionsverfahren für Modell 3 ungeeignet zu sein; die algebraischen p-Werte der übrigen Modelle stimmen recht gut mit den neuen algebraischen p-Werten überein. Alle drei Verfahren kommen zum Niveau $\alpha = 0,10$ zu denselben Testergebnissen; dennoch empfiehlt sich eine Modifikation der Simulations-Parameter wie z. B. eine Verlängerung der Kette.

Die Abhängigkeitsstrukturen in dem Schulkinder-Datensatz werden ebenfalls exemplarisch mit der EH-Prozedur analysiert. Das Haupteffekt-Modell, Modell 7 aus Tabelle 5.1, sei das Startmodell. Im ersten Schritt ist dieses Modell zum Niveau $\alpha = 0,10$ abzulehnen. Von den anschließend zu testenden Modellen, Modell 4, 5, 6, ist das Modell 5 akzeptabel; die beiden übrigen Modelle können abgelehnt werden (Schritt 2b). Wird anschließend Schritt 2a der EH-Prozedur ausgeführt, so ist das Modell 2 zu testen. Dies kann zum Niveau $\alpha = 0,10$ abgelehnt werden und die EH-Prozedur bricht ab. Die asymptotische, algebraische und neue algebraische Edwards-Havránek-Prozedur

	p-Werte		
	approximativ	algebraisch	neues Verfahren
Modell 1	0,341	0,351	0,495
Modell 2	0,025	0,006	0,024
Modell 3	0,600	0,724	0,273
Modell 4	0,020	0,020	0,022
Modell 5	0,288	0,293	0,330
Modell 6	0,027	0,018	0,019
Modell 7	0,016	0,019	0,019

Tabelle 5.9: Approximative und algebraische (nach Diaconis-Sturmfels und dem neuen Verfahren) Ergebnisse des χ^2-Anpassungstest für den Schulkinder-Datensatz.

wählt jeweils Modell 5 für eine Beschreibung des Schulkinder-Datensatzes aus. Als schwach akzeptabel werden die Modelle 1 und 3 angenommen. D. h. die häuslichen und schulischen Umstände scheinen einander zu beeinflussen, eine direkte Beziehung zu dem Klassenzimmerverhalten der Kinder kann nicht eindeutig nachgewiesen werden.

Für die algebraische Untersuchung der Symmetriestrukturen in einem Datensatz kann ausgenutzt werden, dass das perfekte Symmetriemodell S ein Spezialfall aller betrachteten Symmetriemodelle, bedingte, diagonale und ordinale Quasi-Symmetrie (CS, DS bzw. OQS), ist, vergleiche Kapitel 4.1. Nach Theorem 5.2.2 ist also eine Simulation gemäß S ausreichend, um auch die Anpassungsgüte der weiteren Modelle zu testen; die Gröbner-Basis von S umfasst nach Bemerkung 4.1.3 die Gröbner-Basen von CS, DS und OQS.

$$\begin{bmatrix} 1 & 0 & 0 & 1 \\ 2 & 1 & 4 & 4 \\ 2 & 1 & 3 & 1 \\ 0 & 1 & 2 & 2 \end{bmatrix}$$

Tabelle 5.10: Datensatz 5.4.

Nachfolgend wird zunächst eine gemäß Tabelle 4.1 (i) simulierte Kontingenztafel mit $n = 25$ betrachtet; diese ist in Tabelle 5.10 angegeben. Für das herkömmliche Diaconis-Sturmfels-Verfahren wird eine Markov-Kette der Länge 500.000 generiert, die Einschwingphase wird auf 50.000 festgesetzt und anschließend wird jede 100ste Tafel für die Berechnung des p-Wertes berücksichtigt. Das neue Verfahren stützt sich auf eine

Kettenlänge von 1.000.000. Insgesamt konnten für das CS Modell 167.472 Kontingenztafel extrahiert werden, für DS waren es 33.822 und für QOS wurden 89.910 Datensätze extrahiert. Im Weiteren werden die ersten 10 Zustände der selektierten Ketten vernachlässigt und jede 10te Tafel weiter verwendet. Ein Vergleich der exakten, asymptotischen und der beiden algebraischen p-Werte des χ^2-Anpassungstests ist in Tabelle 5.11 angegeben.

	p-Werte			
	approximativ	exakt	algebraisch	neues Verfahren
Modell S	0,177	0,131	0,131	0,131
Modell CS	0,116	0,086	0,088	0,083
Modell DS	0,048	0,048	0,048	0,049
Modell OQS	0,093	0,105	0,093	0,103

Tabelle 5.11: Approximative, exakte und algebraische (nach Diaconis-Sturmfels und dem neuen Verfahren) Ergebnisse des χ^2-Anpassungstest für Datensatz 5.4.

Es zeigt sich eine gute Übereinstimmung der exakten und der beiden algebraischen p-Werte. Zum gewählten Niveau $\alpha = 0,10$ können weder der approximative, der exakte noch die algebraischen Anpassungstests das perfekte Symmetriemodell für Datensatz 5.4 ablehnen.

Im Weiteren wird der in Kapitel 4.1 bereits betrachtete Datensatz 4.2 erneut untersucht. Das Vorgehen bei der Diaconis-Sturmfels-Prozedur entspricht dem oben beschriebenen Verfahren. Für die vorgeschlagene Modellselektion wird gemäß dem S-Modell eine Markov-Kette der Länge 2.200.000 generiert. Aus dieser Kette entsprechen insgesamt 318 Tafeln dem CS-, 17 dem DS- und 82 dem OQS-Modell. Die Einschwingphase ist hier wieder auf 10 Tafeln festgesetzt. Aufgrund der zum Teil geringen Anzahl selektierter Datensätze wird analog zu dem Schulkinder-Datensatz anschließend jede der übrigen Tafeln für die p-Wert Bestimmung berücksichtigt. Zur besseren Übersichtlichkeit werden die approximativen, exakten und algebraischen (nach Diaconis und Sturmfels) p-Werte erneut angegeben.

	p-Werte			
	approximativ	exakt	algebraisch	neues Verfahren
Modell S	0,019	0,005	0,002	0,002
Modell CS	0,608	0,669	0,670	0,682
Modell DS	0,762	1,000	1,000	1,000
Modell OQS	0,790	0,829	0,822	0,958

Tabelle 5.12: Approximative, exakte und algebraische (nach Diaconis-Sturmfels und dem neuen Verfahren) Ergebnisse des Pearson Anpassungstest für Datensatz 4.2.

Das neue Verfahren identifiziert dieselben Symmetriestrukturen wie der exakte und der herkömmliche algebraische Test, vergleiche Tabelle 5.12. Insbesondere werden das DS- sowie das OQS-Modell von den Daten unterstützt. Für das vorgeschlagene algebraische Verfahren konnten, trotz der angepassten Länge der generierten Markov-Kette, nur wenige Datensätze mit derselben realisierten suffizienten Statistik extrahiert werden, so dass die Ergebnisse nicht notwendigerweise verlässlich sind. Ursächlich hierfür scheinen die großen Unterschiede in der Mächtigkeit der jeweiligen $\mathscr{X}_{t(i)}$, $i = S, CS, DS, OQS$, zu sein, siehe Kapitel 4.1.3. So umfasst $\mathscr{X}_{t(S)}$ insgesamt 30.240 Kontingenztafeln, $\mathscr{X}_{t(DS)}$ besteht jedoch nur aus 12 Datensätzen.

Kapitel 6

Ausblick

Die computergestützte Algebra liefert vielversprechende Lösungsansätze für verschiedene statistische Fragestellungen. Die meisten Untersuchungen zur Verwendung der algebraischen Statistik setzen jedoch diskrete Variablen voraus. Eine Erweiterung möglicher Anwendungsgebiete steht daher im Fokus aktueller Forschung. Drton (2006) nutzt algebraische Methoden zur Untersuchung stetiger Variablen. Speziell definiert er den Begriff des „algebraischen Gauß-Modells“ und führt an ausgewählten Beispielen die Verwendung der algebraischen Statistik für eine Analyse normalverteilter Daten vor. In Drton et al. (2007) wird eine algebraische Faktoranalyse eingeführt. Matúš (2005) beschreibt bedingte Unabhängigkeiten zweier Variablen eines normalverteilten Zufallsvektors anhand von Idealen. Lněicka und Matúš (2007) erweitern diesen Ansatz für gleichzeitig auftretende bedingte Unabhängigkeiten von Teilvektoren eines multivariat normalverteilten Zufallsvektors.

Zentraler Bestandteil der vorliegenden Arbeit ist die Anwendung des Diaconis-Sturmfels-Algorithmus in der Statistik. Dieser erweist sich für die Analyse kategorialer Daten als wertvolle Ergänzung zur traditionellen Asymptotik und zu exakten Berechnungen. In diesem Kapitel interessiert speziell die Untersuchung gemischt stetiger-diskreter Variablen unter Verwendung des Diaconis-Sturmfels-Algorithmus. Nachfolgend werden verschiedene Lösungsansätze vorgestellt und die damit verbundenen Probleme erläutert.

Zugrunde liege ein Zufallsvektor $X = (X'_\Delta, X'_\Gamma)'$, wobei $\Delta = \{1, \ldots, q\}$ die Indexmenge der diskreten und $\Gamma = \{1, \ldots, r\}$ die der stetigen Merkmale bezeichne, vergleiche Kapitel 2.2. Graphische Modelle bieten eine geeignete Möglichkeit für die Darstellung der Abhängigkeitsstrukturen zwischen den Komponenten von X. Die gemeinsame Verteilung von X sei die CG-Verteilung, siehe Definition 2.2.5, die hier als homogen angenommen wird. Mit der kanonischen Parametrisierung kann gezeigt werden, dass die CG-Verteilung aus einer Exponentialfamilie stammt, siehe Lauritzen und Wermuth (1989). Die Identifizierung von Unabhängigkeitsbeziehungen erfolgt anhand der „ausgedehnten“ kanonischen Parametrisierung. Damit ist die Dichte der CG-Verteilung darstellbar als

$$f_{(X_\Delta,X_\Gamma)}(i,y) = \exp\left\{ \sum_{\{a:a\subseteq\Delta\}} \lambda^a_{i_a} + \sum_{\{a:a\subseteq\Delta\}} \sum_{\gamma\in\Gamma} \eta^{\gamma;a}_{i_a} y_\gamma - \frac{1}{2} \sum_{\{a:a\subseteq\Delta\}} \sum_{\gamma,\,\mu\in\Gamma} \psi^{\gamma,\mu;a}_{i_a} y_\gamma y_\mu \right\},$$

siehe Kapitel 2. Eine homogene CG-Verteilung weist keine gemischten quadratischen Effekte auf, d.h. $\psi^{\gamma,\mu;a}_{i_a}$ ist gleich Null, wenn $a \subseteq \Delta \neq 0$, siehe Lauritzen (1998), Kapitel 6.24. Weiter sind zwei Merkmale genau dann voneinander bedingt unabhängig, gegeben die übrigen Variablen, wenn alle Interaktionseffekte, die diese beiden Variablen beinhalten, gleich Null sind. Die Bestimmung der suffizienten Statistik für die Parameter des zugrunde liegenden Modells erfolgt gemäß Lauritzen und Wermuth (1989):
Der Zufallsvektor $X = (X'_\Delta, X'_\Gamma)'$ habe N Realisationen $x^{(1)},\dots,x^{(N)}$ mit $x^{(\nu)} = (i^{(\nu)\prime}, y^{(\nu)\prime})' = (i^{(\nu)\prime}_\delta, \delta\in\Delta,\, y^{(\nu)\prime}_\gamma, \gamma\in\Gamma)'$, wobei $i^{(\nu)}_\delta \in \mathcal{I}_\delta$ und $y^{(\nu)}_\gamma \in I\!R$ ist, vgl. Kapitel 2. Weiter bezeichne

$\mathscr{C}_\Delta$	die Menge der Cliquen in $\mathcal{G} = (\Delta, E_\Delta)$;
$\mathscr{C}_\Delta(\gamma)$	die Menge der Untermengen d von Δ, so dass $d\cup\{\gamma\}$ eine Clique in $\mathcal{G} = (\Delta\cup\{\gamma\}, E_{\Delta\cup\{\gamma\}})$ ist;
$\mathscr{C}_\Delta(\gamma,\mu)$	die Menge der Untermengen d von Δ, so dass $d\cup\{\gamma\}\cup\{\mu\}$ eine Clique in $\mathcal{G} = (\Delta\cup\{\gamma\}\cup\{\mu\}, E_{\Delta\cup\{\gamma\}\cup\{\mu\}})$ ist.

Damit ist die beobachtete suffiziente Statistik für die Parameter des interessierenden Modells gegeben durch

$$\begin{aligned}
&n(i_d),\ d\in\mathscr{C}_\Delta,\\
&s(i_d)_\gamma,\ sp(i_d)_{\gamma\gamma},\ \gamma\in\Gamma,\ d\in\mathscr{C}_\Delta(\gamma),\\
&sp(i_d)_{\gamma\mu},\ \{\gamma,\mu\}\in E_\Gamma,\ d\in\mathscr{C}_\Delta(\gamma,\mu),
\end{aligned}$$

mit $n(i) = \sum_{\nu:i^{(\nu)}=i} 1$, $s(i) = \sum_{\nu:i^{(\nu)}=i} y^{(\nu)}$, $sp(i) = \sum_{\nu:i^{(\nu)}=i} y^{(\nu)} y^{(\nu)\prime}$. Wie üblich werden für die entsprechenden Zufallsvariablen Großbuchstaben verwendet, $N(i_d)$, $S(i_d)_\gamma$, $SP(i_d)_{\gamma\mu}$.

Unter Verwendung von MCMC-Verfahren ermöglicht der Diaconis-Sturmfels-Algorithmus einen „exakten“, bedingten Test in einer k-parametrigen diskreten Exponentialfamilie. Speziell liefert er eine Gröbner-Basis zur Bestimmung der jeweiligen Vorschlagsdichte im Metropolis-Hastings-Algorithmus. Die Elemente einer Gröbner-Basis bewirken eine Veränderung der zugrunde liegenden Kontingenztafel. Beispielsweise ist das Bewegungsmuster für einen Test auf Unabhängigkeit zweier diskreter Merkmale gegeben durch $\begin{pmatrix} +1 & -1 \\ -1 & +1 \end{pmatrix}$. Die Schritte sind diskret und nicht direkt auf den stetigen Fall übertragbar.

Ein möglicher Lösungsansatz könnte jeweils ein separater Vorschlag für die Beobachtungen der diskreten und der stetigen Variablen sein. Speziell könnte der Diaconis-Sturmfels-Algorithmus zunächst für den Vorschlag eines neuen Zustandes der diskreten Beobachtungen verwendet werden. Für die entsprechenden stetigen Ausprägungen werden dann alternative Prozeduren für den Vorschlag eines möglichen Datensatzes mit derselben beobachteten suffizienten Statistik gesucht. Beide Verfahren könnten anschließend im Metropolis-Hastings-Algorithmus zusammengeführt werden.

Im Folgenden werden zerlegbare graphische Modelle betrachtet, denn diese weisen für die Herleitung bedingter Tests in einer k-parametrigen Exponentialfamilie günstige Eigenschaften auf. Lauritzen (1998), Kapitel 6.3.3, beschreibt exakte Test auf bedingte Unabhängigkeit zweier diskreter, zweier stetiger Variablen bzw. einer stetigen und einer diskreten Variable, gegeben die übrigen Variablen. Für die Konstruktion solcher Tests ist es erforderlich, dass der k-dimensionale Parametervektor in einen eindimensionalen zu testenden Parameter und $k-1$ Störparameter affin transformiert werden kann. Frydenberg und Lauritzen (1989) nutzen Eigenschaften zerlegbarer graphischer Modelle aus, um das interessierende Testproblem zu vereinfachen. Dazu sei angenommen, dass beide zugrunde liegenden graphischen Modelle $M(\mathcal{G})$ und $M(\mathcal{G}')$ zerlegbar sind und $M(\mathcal{G}')$ eine Kante weniger aufweist als $M(\mathcal{G})$. Es bezeichne C^* die eindeutige Clique des Graphen $\mathcal{G}$, die diese Kante enthält. Dann ist der Likelihood-Quotient der Modelle $M(\mathcal{G})$ und $M(\mathcal{G}')$ gleich dem Likelihood-Quotienten der durch C^* induzierten

Modelle $M(\mathcal{G}_{C^*})$ und $M(\mathcal{G}'_{C^*})$.

Nachfolgend werden zwei binäre und zwei stetige Variablen betrachtet, $\Delta = \{1, 2\}$ und $\Gamma = \{3, 4\}$. In Abbildung 6.1 ist eine Auswahl zerlegbarer Graphen angegeben, die sich in dieser Reihenfolge nur durch eine Kante von ihrem Nachbarn unterscheiden.

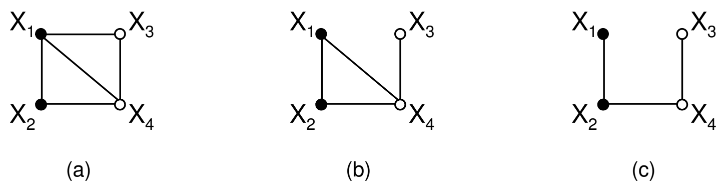

Abbildung 6.1: Beispiele für zerlegbare Graphen mit $V = \{\Delta \cup \Gamma\}$, wobei $\Delta = \{1, 2\}$ und $\Gamma = \{3, 4\}$.

Zugrunde liege das bedingte Unabhängigkeitsmodell $X_2 \perp\!\!\!\perp X_3 | \{X_1, X_4\}$ aus Abbildung 6.1 (a). Um die Eindeutigkeit der Parameter zu gewährleisten, sei $i^* = (1, 1)$ als Referenzzelle festgelegt. Im Folgenden interessiere beispielhaft ein Test für Modell (a) gegen Modell (b). Das Testproblem kann somit ausgedrückt werden als

$$H_0 : \eta_{i_1}^{3;1} = \eta_{i_1 i_2}^{3;12} = 0 \qquad \text{vs.} \qquad H_1 : \eta_{i_1}^{3;1} \neq 0 \text{ und bzw. oder } \eta_{i_1 i_2}^{3;12} \neq 0.$$

$C^* = (X_1, X_3, X_4)$ ist die eindeutige Clique von Graph (a) aus Abbildung 6.1, die die zu testende Kante enthält. Nach Frydenberg und Lauritzen (1989) werden nun die durch C^* induzierten graphischen Modelle (a′) und (b′) aus Abbildung 6.2 verglichen. Das obige Testproblem kann somit umformuliert werden, so dass eine eindimensionale Komponente des Parametervektors zugrunde liegt:

$$H_0' : \eta_{i_1}^{3;1} = 0 \quad \text{vs.} \quad H_1' : \eta_{i_1}^{3;1} \neq 0.$$

Gemäß den obigen Ausführungen setzt sich die suffiziente Statistik für die Parameter in Modell (b′) zusammen aus $N(i_1)$, $S(i_1)_4$, $SP(i_1)_{44}$. Damit bleibt auf eine eindimensionale Komponente des interessierenden Parametervektors zu testen.

Weitere nötige Schritte für die Herleitung bedingter Tests in einer k-parametrigen Exponentialfamilie sind in Ferguson (1967), Kapitel 5.4 beschrieben. So ist beispielsweise die Verteilung auf $\mathscr{X}_t$, der Menge aller möglichen Datensätze mit beobachteter suffizienter Statistik t, zu bestimmen. Die Hauptschwierigkeit dieses Ansatzes ist die

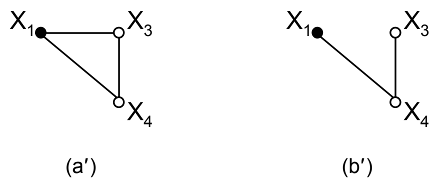

Abbildung 6.2: Durch das Testproblem (a′) und (b′) aus Abbildung 6.2 induzierte graphische Modelle nach Frydenberg und Lauritzen (1989).

Generierung weiterer Datensätze mit vorgegebener realisierter suffizienter Statistik. Denkbar wäre, den zugrunde liegenden Datensatz in eine diskrete und eine stetige Komponente aufzuteilen. Wie in Kapitel 3.3 beschrieben, wird mit dem Diaconis-Sturmfels-Algorithmus ein neuer Zustand der diskreten Beobachtungen vorgeschlagen. Gesucht wird nun eine Vorschlagsdichte für die stetigen Variablen, die ausschließlich Datensätze generiert, die denselben Wert der zugrunde liegenden suffizienten Statistik aufweisen wie der beobachtete Datensatz. Lindquist und Taraldsen (2005 und 2007) greifen eine Idee von Trotter und Tukey (1956) auf, um hierzu Algorithmen vorzuschlagen. In den so genannten „conditional Monte Carlo“-Verfahren werden bedingte Stichproben durch adäquat gewichtete unbedingte Stichproben generiert. Diese Methode produziert nicht direkt Datensätze mit der gegebenen suffizienten Statistik; eine Implementierung als Vorschlagsdichte im Metropolis-Hastings-Algorithmus scheint daher zu rechenintensiv zu sein.

Ein vielversprechender Ansatz für die Verwendung des Diaconis-Sturmfels-Algorithmus bei der Analyse gemischt stetiger-diskreter Variablen ergibt sich durch eine von Edwards (2000), Kapitel 5.4, beschriebene Darstellung des gegebenen Datensatzes. Es sei angenommen, dass die zugrunde liegenden Daten unabhängig und identisch verteilte Beobachtungen einer diskreten Variablen X_1 mit I_1 möglichen Ausprägungen und einer stetigen Variablen X_2 mit Realisationen aus $\mathbb{R}$ umfassen. Getestet wird die Unabhängigkeit dieser beiden Merkmale. Edwards (2000) verwendet die geordneten Beobachtungen des stetigen Merkmals und stellt somit den Datensatz als Kontingenztafel dar. Treten bei den Beobachtungen der stetigen Variable Bindungen auf, so ist der vorliegende Datensatz darstellbar als

$$
\begin{array}{c|ccc}
 & x_{2\,(1)} & \cdots & x_{2\,(O)} \\
\hline
x_1 = 1 & n_{11} & \cdots & n_{1O} \\
\vdots & & & \\
x_1 = I_1 & n_{I_1 1} & \cdots & n_{I_1 O}
\end{array}
,
$$

wobei $x_{2\,(1)}, \ldots, x_{2\,(O)}$ die verschiedenen Werte der geordneten Statistik $X_{2\,(1)}, \ldots, X_{2\,(O)}$ sind. $n_{i_1 i_2}$ bezeichne die Anzahl der Beobachtungen mit $x_1 = i$, $i = 1, \ldots, I_1$, und $x_2 = x_{2\,(i_2)}$, $i_2 = 1, \ldots, O$. Der Unabhängigkeitstest eines stetigen und eines diskreten Merkmals ist somit zurückgeführt auf einen entsprechenden Test zweier diskreter Variablen. Damit kann der Algorithmus von Diaconis und Sturmfels für diese gemischt stetigen-diskreten Daten verwendet werden. Die suffiziente Statistik besteht aus den Zeilen- und Spaltensummen der zugrunde liegenden Kontingenztafel. Die Verteilung auf der Menge aller möglichen Tafeln mit identischer beobachteter suffizienter Statistik ist hypergeometrisch. Dieses Vorgehen kann für einen Test auf bedingte Unabhängigkeit, $X_1 \perp\!\!\!\perp X_2 | X_3$, erweitert werden, wobei X_1 und X_3 diskret und X_2 stetig oder diskret sein kann. Das in Kapitel 5 beschriebene Vorgehen kann unmittelbar auf diese Situation übertragen werden und wird daher an dieser Stelle nicht vorgeführt; das zugrunde liegende Modell entspricht damit Modell 3 aus Abbildung 5.1. Je nach Anzahl der verschiedenen Ausprägungen der stetigen Variable besteht die zugehörige Gröbner-Basis aus sehr vielen Polynomen. Beispielsweise seien X_1 und X_3 binär und X_2 besitze nur zehn unterschiedliche Realisationen. Dann setzt sich die Gröbner-Basis für das bedingte Unabhängigkeitsmodell bereits aus 90 Basispolynomen zusammen. Zudem besitzt die interessierende Tafel aufgrund des Konstruktionsprinzips überwiegend Einträge von Null und Eins. Damit ist zu erwarten, dass viele vorgeschlagene neue Zustände negative Einträge besitzen. Gemäß den Ausführungen in Kapitel 3 verweilt die Kette dann in ihrem vorherigen Zustand. Für eine adäquate Annäherung der Verteilung auf $\mathscr{Z}_t$, der Menge aller möglichen Datensätze mit demselben realisierten Wert der suffizienten Statistik t, wird daher vermutlich eine sehr lange Kette benötigt.

Die Verwendung des Diaconis-Sturmfels-Algorithmus ist prinzipiell auch bei gemischt stetigen-diskreten Variablen möglich. Der damit verbundene Simulationsaufwand scheint jedoch sehr groß zu sein, so dass alternative Verfahren für die praktische Anwendung geeigneter erscheinen.

Kapitel 7

Zusammenfassung

Die Analyse struktureller Abhängigkeiten der Komponenten eines multivariaten Zufallsvektors ist eine wichtige Aufgabenstellung in der Statistik. Zur Bewertung solcher Abhängigkeitsstrukturen wurden in der vorliegenden Arbeit Anpassungstests bzw. Konfidenzintervalle als Kriterium herangezogen. Diesen liegen oftmals eine Approximation der Verteilung der jeweiligen Teststatistik oder exakte Berechnungen zugrunde. Unter Verwendung von MCMC-Methoden ermöglicht der Algorithmus von Diaconis und Sturmfels (1998) die Simulation aus der bedingten Verteilung einer diskreten Exponentialfamilie mit beobachteter suffizienter Statistik. Die so konstruierten Tests weisen als „exakte" Tests in einer k-parametrigen Exponentialfamilie bestimmte Optimalitätseigenschaften auf, siehe z. B. Witting (1985). Speziell beschreiben Diaconis und Sturmfels die Konstruktion der Vorschlagdichte im Metropolis-Hastings-Algorithmus. Darauf basierend wurden in dieser Arbeit alternative algebraische Tests und Konfidenzintervalle entwickelt.

In den Kapiteln 2 und 3 wurden die theoretischen Grundlagen dieser Arbeit erläutert. Für die Beschreibung struktureller Zusammenhänge multivariater Zufallsvektoren wurden log-lineare und graphische Modelle herangezogen. Liegen der Untersuchung ausschließlich kategoriale Merkmale zugrunde, so können mit log-linearen Modellen verschiedene Abhängigkeitsstrukturen in einem Datensatz wie z. B. Symmetrie- oder Unabhängigkeitsbeziehungen dargestellt werden. Graphische Modelle repräsentieren die Unabhängigkeitsbeziehungen zwischen den Komponenten eines multivariaten Zufallsvektors durch einen Graphen. Einfache graphische Modelle mit multinomialverteiltem Zufallsvektor sind in der Klasse log-linearer Modelle enthalten. Graphische

Modelle können zudem auch für Datensätze mit stetigen oder gemischt stetigen-diskreten Merkmalen angewendet werden. Nach Lauritzen und Wermuth (1989) wurde für den interessierenden Fall gemischt stetiger-diskreter Variablen die Familie der CG-Verteilungen als gemeinsame Verteilung eingeführt.

Es folgte die Darstellung der verwendeten Begriffe und Verfahren der computergestützten Algebra bzw. der algebraischen Statistik. Es wurden Ideale und Varietäten definiert, ihr Zusammenhang erklärt sowie ihre Struktur anhand einer Gröbner-Basis charakterisiert. Darauf aufbauend konnte der Diaconis-Sturmfels-Algorithmus eingeführt werden, der anschließend für die Entwicklung neuer algebraischer Tests und Konfidenzintervalle verwendet wurde. Dabei wurden keine Datensätze mit strukturellen Nullzellen betrachtet. Eine resultierende Gröbner-Basis müsste entsprechend der Position der strukturellen Nullzelle modifiziert werden und ist somit immer ein Sonderfall.

Liegen der Untersuchung so genannte gepaarte Beobachtungen zugrunde, so ist häufig von Interesse, ob die Zellwahrscheinlichkeiten des interessierenden Datensatzes eine Symmetriestruktur aufweisen; ein wichtiges Beispiel hierfür sind Gutachterverlässlichkeitsstudien. Zunächst wurde das perfekte Symmetriemodell betrachtet. Dieses ist jedoch sehr restriktiv, so dass zusätzlich einige alternative Modelle, speziell das bedingte, diagonale sowie das ordinale Quasi-Symmetriemodell, vorgestellt wurden. Es wurde die polynomiale Parametrisierung dieser Modelle bestimmt und die jeweiligen Gröbner-Basen berechnet. Diese wurden anschließend im Metropolis-Hastings-Algorithmus implementiert und so die Verteilungen der jeweiligen Teststatistiken simuliert.

Exemplarisch wurden für den Test auf perfekte Symmetrie der Einfluss der Parameter der Markov-Kette auf das Testergebnis analysiert. Da Voruntersuchungen bereits gezeigt hatten, dass keine Wechselwirkungen zwischen der Ketten- und Schrittlänge sowie der Einschwingphase zu erwarten sind, wurde jeweils ein Parameter variiert, während die übrigen auf festgelegten Standardwerten verblieben. Dabei zeigte sich, dass insbesondere die Kettenlänge adäquat gewählt werden muss.

Lag der Untersuchung ein Datensatz mit geringem Stichprobenumfang zugrunde, so wurden die neuen algebraischen Tests (auf perfekte, bedingte, diagonale und ordinale quasi-Symmetrie) mit den entsprechenden exakten Verfahren verglichen.

Hier zeigte sich eine gute Übereinstimmung der Testergebnisse. Bei einem größeren Stichprobenumfang erfolgte ein Vergleich der approximativen und algebraischen Tests. Die p-Wert-Paare stimmten zum großen Teil überein, allerdings konnten einige Abweichungen entdeckt werden. Ursächlich hierfür scheinen die betrachteten Simulationsmodelle (vgl. Tabelle 4.1) zu sein, die einer möglichst realistischen (Daten-) Situation entsprechen sollen. Selbst bei einem Stichprobenumfang von $n = 100$ werden nur wenige Realisationen außerhalb der Hauptdiagonalen erwartet, so dass die Approximation noch nicht gerechtfertigt scheint.

Häufig ist nicht nur von Bedeutung, einen Zusammenhang zwischen zwei kategorialen Merkmalen zu erkennen, sondern zusätzlich die Stärke der Abhängigkeit zu quantifizieren. Das Odds Ratio sowie das zugehörige Konfidenzintervall geben hierzu Aufschluss. Das Konstruktionsprinzip eines algebraischen Konfidenzintervalls stützt sich sowohl auf das exakte wie auch auf das approximative Verfahren. So wurde ausgenutzt, dass ein Wert des Odds Ratios von Eins äquivalent ist zur Unabhängigkeit der beiden erhobenen Merkmale. Mit der polynomialen Parametrisierung des Unabhängigkeitsmodells wurde die zugehörige Gröbner-Basis bestimmt. Damit konnten anschließend die Normalverteilungsquantile des approximativen Verfahrens durch die algebraischen Quantile ersetzt werden.

Es zeigte sich, dass das algebraische Konfidenzintervall bei einem Datensatz mit kleinem Stichprobenumfang häufig Unendlich als obere Grenze angibt. Dies ist der Fall, wenn „zu viele“ Datensätze mit Einträgen n_{12} bzw. n_{21} gleich Null generiert werden. Abhilfe kann der ebenfalls eingeführte modifizierte Schätzer für das Odds Ratio $\widehat{OR}_{\text{mod}}$ schaffen.

Interessieren verschiedene mögliche Modelle zur Beschreibung der Abhängigkeitsstrukturen in einem Datensatz, so sind algebraische Anpassungstests nach Diaconis-Sturmfels möglich; der nötige Simulationsaufwand ist jedoch groß und in der Praxis oft nicht zu erbringen. Am Beispiel der Graphischen Modelle für kategoriale Daten sowie der betrachteten Symmetriemodelle wurde gezeigt, wie der erforderliche Simulationsaufwand z. T. erheblich verringert werden kann. Häufig weist die interessierende Modellklasse eine hierarchische Struktur auf; das bedeutet, es gibt in dieser Modellklasse ein Modell, das ein Spezialfall (Untermodell) aller weiteren betrachteten Modelle ist. Diese Anordnung findet sich ebenfalls in der suffizienten Statistik und damit auch in der Menge aller möglichen Datensätze mit beobachteter suffizienter Statistik wieder und konnte für die Entwicklung einer algebraischen Modellselektion ausgenutzt werden. Speziell konnte gezeigt werden, dass das Ideal eines Modells die Ideale aller darauf aufbauenden Modelle enthält. Entsprechend gilt, dass die Gröbner-Basis für das Untermodell entweder die Basispolynome aller weiteren Modelle enthält oder diese durch Linearkombination darstellen kann. Die vorgeschlagene algebraische Modellselektion beruht daher auf der Simulation gemäß der Gröbner-Basis des Untermodells. Anschließend werden die verschiedenen Markov-Ketten anhand der beobachteten suffizienten Statistik extrahiert und so die jeweilige Verteilung der Teststatistik simuliert. Die resultierenden algebraischen p-Werte können dann für ein gewähltes Modellselektionsverfahren, wie z. B. die Edwards-Havránek-Prozedur, verwendet werden.

Diese neue Strategie erwies sich insbesondere bei kleinen Datensätzen als sinnvolle Alternative zu traditionellen asymptotischen und exakten Verfahren oder zur herkömmlichen Diaconis-Sturmfels Methode. Die hypergeometrische Verteilung auf $\mathscr{Z}_t$, der Menge aller möglichen Datensätze mit realisierter suffizienter Statistik t, konnte adäquat angenähert werden. Der zusätzliche Selektionsschritt in der entwickelten Methode bedeutete oftmals einen geringeren Aufwand als einzelne Simulationen für jedes betrachtete Modell. Anhand der Datenbeispiele wurden aber auch die Grenzen der vorgestellten Strategie deutlich: Insbesondere die Mächtigkeit von $\mathscr{Z}_t$ beeinflusste den Nutzen des neuen Verfahrens. So musste der nötige Simulationsaufwand beispielsweise bei großem Stichprobenumfang entsprechend adjustiert werden. Für Datensatz 4.2 konnten trotz der großen Anzahl generierter Zustände nur sehr wenige Kontingenztafeln extrahiert werden, die dieselben beobachteten suffizienten Statistiken für die Modelle CS, DS und QOS aufweisen wie der zugrunde liegende Datensatz. Die Annäherung der hypergeometrischen Verteilung auf $\mathscr{Z}_t$ und damit die Aussagekraft des neuen Verfahrens war daher für diesen Datensatz fraglich. Ein weiterer wichtiger Aspekt ist die Dimension der zugrunde liegenden Kontingenztafel, denn die Anzahl der Basispolynome erhöht sich entsprechend. Ein zusätzliches Problem für eine algebraische Analyse höherdimensionaler Kontingenztafeln ist die hohe Berechnungszeit der entsprechenden Gröbner-Basen. Dobra (2003) sowie Dobra und Sullivant (2004) haben in diesem Kontext effiziente Verfahren zur Bestimmung von Gröbner-Basen entwickelt. Takemura und Aoki (2004) leiten charakteristische Eigenschaften einer Markov-Basis her, die eine irreduzible Markov-Kette generiert. Hoşten und Sullivant (2007) zeigen, dass sich unter bestimmten Voraussetzungen die Markov-Basis einer multidimensionalen Tafel aus Markov-Basen kleinerer Tafeln zusammensetzt. Es erscheint sinnvoll, diese Methoden für weitergehende Untersuchungen zu implementieren.

Die Erschließung weiterer Anwendungsgebiete der Algebraischen Statistik ist Gegenstand aktueller Forschung. In dieser Arbeit wurde der Nutzen des Diaconis-Sturmfels-Algorithmus bei gemischt stetigen-diskreten Variablen diskutiert. Liegen der Untersuchung eine stetige und eine bzw. zwei diskrete Variablen zugrunde, so beschreibt Edwards (2000) eine diskretisierte Darstellungsmöglichkeit des zugrunde liegenden Datensatzes. Damit ist der Diaconis-Sturmfels-Algorithmus prinzipiell auch für einige Testsituationen bei gemischt stetigen-diskreten Datensätzen anwendbar. Allerdings scheint der nötige Simulationsaufwand so groß, dass ein Gebrauch in der Praxis noch nicht zu erwarten ist.

Weitere interessante Forschungsfragen ergeben sich aufgrund des verwendeten MCMC-Verfahrens. Für eine Analyse des Konvergenzverhaltens der generierten Markov Ketten sollte speziell die Wahl der Ketten- und Schrittlänge sowie der Einschwingphase Gegenstand einer umfassenden Simulationsstudie sein. Mögliche Verfahren zur Bewertung des Konvergenzverhaltens können beispielsweise El Adlouni et al. (2006) sowie Robert und Casella (2004), Kapitel 12, entnommen werden. Zudem wäre ein weiterer Vergleich der algebraischen, asymptotischen und exakten Verfahren mit weiteren Methoden wie z. B. anderen simulationsbasierten Verfahren interessant. Insbesondere für die Modellselektion bei log-linearen Modellen bietet sich das von Booth und Butler (1999) und von Caffo und Booth (2001) weiterentwickelte Verfahren an.

ANHANG A

Anhang

A.1 Buchberger Algorithmus

Nachfolgend wird der Algorithmus von Buchberger zur Bestimmung einer Gröbner Basis beschrieben. Dazu bezeichne $\overline{p}^{\,P}$ den Rest aus der Polynomdivision eines Polynoms p durch das $s-$Tupel $P = \{p_1, \ldots, p_s\}$. Weiter seien p und $g \in K[x_1, \ldots, x_n]$ zwei von Null verschiedene Polynome mit $multideg(p) = (\alpha_1, \ldots, \alpha_n) =: \alpha$ und $multideg(g) = (\beta_1, \ldots, \beta_n) =: \beta, \quad \alpha, \beta \in \mathbb{Z}^n_{\geq 0}$. Dann heißt $x^\gamma = (x_1^{\gamma_1}, \ldots, x_n^{\gamma_n})$ das kleinste gemeinsame Vielfache von p und g, wobei gilt $\gamma = (\gamma_1, \ldots, \gamma_n) \in \mathbb{Z}^n_{\geq 0}$ und $\gamma_i := \max(\alpha_i, \beta_i)$, $i = 1, \ldots, n$. Grundlegend für den Buchberger-Algorithmus zur Bestimmung von Gröbner Basen ist das so genannte S-Polynom.

Definition A.1.1 (S-Polynom)

Es seien $p, g \in K[x_1, \ldots, x_n]$ zwei von Null verschiedene Polynome mit $multideg(p) = \alpha = (\alpha_1, \ldots, \alpha_n)$ und $multideg(g) = \beta = (\beta_1, \ldots, \beta_n)$. Dann heißt

$$S(p, g) = \frac{x^\gamma}{LT(p)} \cdot p - \frac{x^\gamma}{LT(g)} \cdot g$$

das S-Polynom von p und g.

Sind also die Polynome p und g Elemente eines Ideals $\mathscr{I}$, dann ist auch $S(p, g)$ aus $\mathscr{I}$. Aufgrund der Konstruktion des S-Polynoms $S(p, g)$ heben sich die Leitterme von p und g auf. Weitergehend kann gezeigt werden, dass jede Aufhebung der Leitterme von Polynomen mit gleichem Multigrad auf das Prinzip des S-Polynoms zurückgeführt werden kann, vgl. Cox et al. (1997), Kapitel 2.6. Unter Ausnutzung der Definition sowie der Eigenschaften von S-Polynomen kann das so genannte Buchberger-Kriterium bewiesen werden.

Theorem A.1.2 (Buchberger-Kriterium)

Sei $\mathscr{I} \subset K[x_1, \ldots, x_n]$ *ein Polynomideal und* $\mathscr{G} = \{g_1, \ldots, g_t\}$ *eine Basis von* $\mathscr{I}$. *Dann ist* $\mathscr{G}$ *genau dann eine Gröbner Basis von* $\mathscr{I}$, *wenn bei der Division von* $S(g_i, g_j)$ *durch* $\mathscr{G}$ *der Rest Null ist für alle* $i \neq j$.

Beweis: Cox et al. (1997), S. 82f.

Der Algorithmus von Buchberger zur Bestimmung von Gröbner Basen beruht auf diesem Kriterium. Aufgrund seiner großen Bedeutung wird der Pseudocode angegeben.

Theorem A.1.3 (Buchberger-Algorithmus)

Es sei $\mathscr{I} =< p_1, \ldots, p_s >\neq \{0\} \subset K[x_1, \ldots, x_n]$ *ein Polynomideal und* $\succ$ *eine fest gewählte Monomordnung. Die Bildung einer Gröbner Basis für* $\mathscr{I}$ *erfolgt in endlich vielen Schritten nach dem Pseudocode:*

Eingabe:	$P = (p_1, \ldots, p_s)$
Ausgabe:	*Gröbner Basis* $\mathscr{G} = (g_1, \ldots, g_t)$ *für* $\mathscr{I}$ *mit* $P \subset \mathscr{G}$
Setze	$\mathscr{G} := P$
Wiederhole	
	$\mathscr{G}' := \mathscr{G}$
	Berechne für jedes Paar $\{g, q\}$, $g \neq q$ *von*
	Polynomen aus $\mathscr{G}'$ $\quad S := \overline{S(g,q)}^{\mathscr{G}'}$
	Ist $S \neq 0$, *dann setze* $\mathscr{G} := \mathscr{G}' \cup \{S\}$
Stoppe, wenn gilt: $\mathscr{G} = \mathscr{G}'$	

Beweis: Cox et al (1997), S. 87f.

A.2 Anhang zu Algebraische Testprozeduren und Konfidenzintervalle

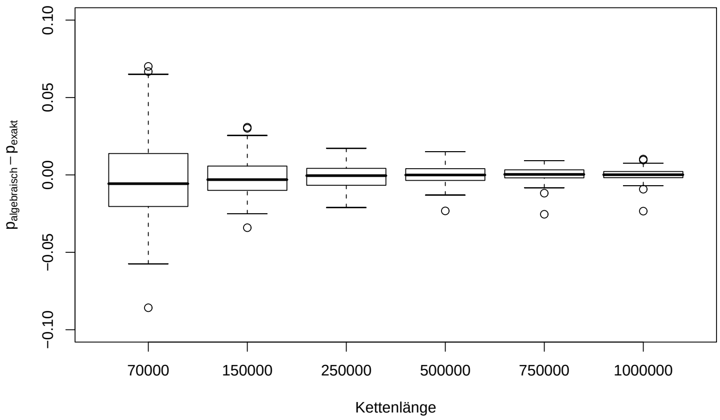

Abbildung A.1: Boxplots der Differenzen der exakten und algebraischen p-Werte des stetigkeitskorrigieren χ^2-Tests für verschiedene Kettenlängen. Die Datengenerierung erfolgt gemäß dem S Modell, siehe Tabelle 4.1 (i) mit $n = 25$.

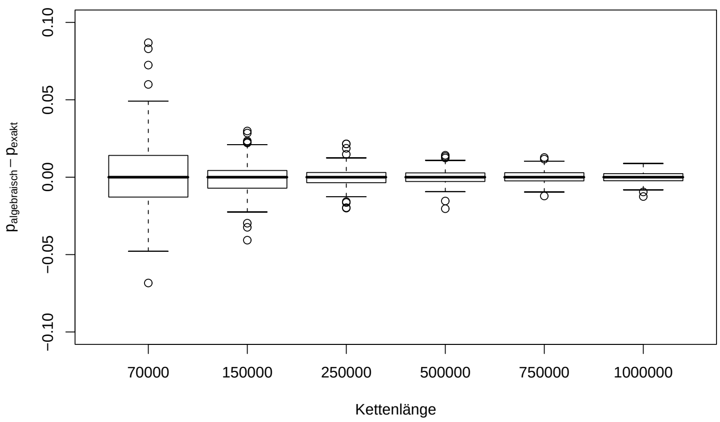

Abbildung A.2: Boxplots der Differenzen der exakten und algebraischen p-Werte des Tests von May und Johnson für verschiedene Kettenlängen. Die Datengenerierung erfolgt gemäß dem S Modell, siehe Tabelle 4.1 (i) mit $n = 25$.

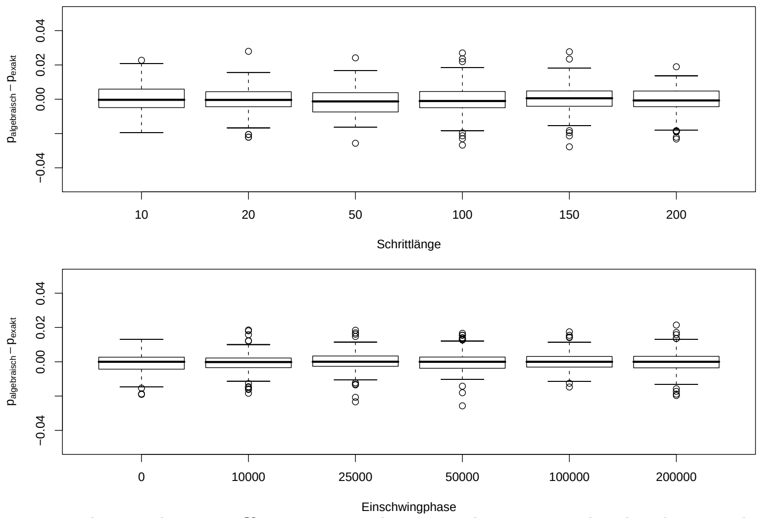

Abbildung A.3: Boxplots der Differenzen der exakten und algebraischen p-Werte des stetigkeitskorrigieren χ^2-Tests für verschiedene Schrittlängen (oben) bzw. Einschwingphasen (unten). Die Datengenerierung erfolgt gemäß dem S Modell, siehe Tabelle 4.1 (i) mit $n = 25$.

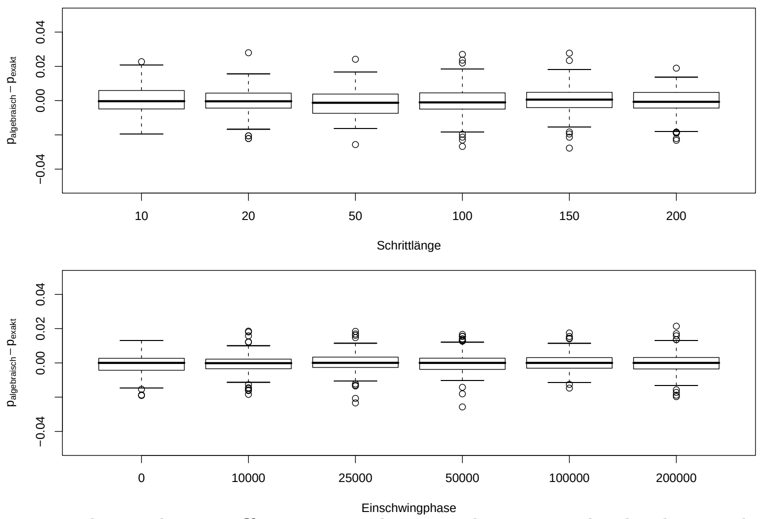

Abbildung A.4: Boxplots der Differenzen der exakten und algebraischen p-Werte des Tests von May und Johnson für verschiedene Schrittlängen (oben) bzw. Einschwingphasen (unten). Die Datengenerierung erfolgt gemäß dem S Modell, siehe Tabelle 4.1 (i) mit $n = 25$.

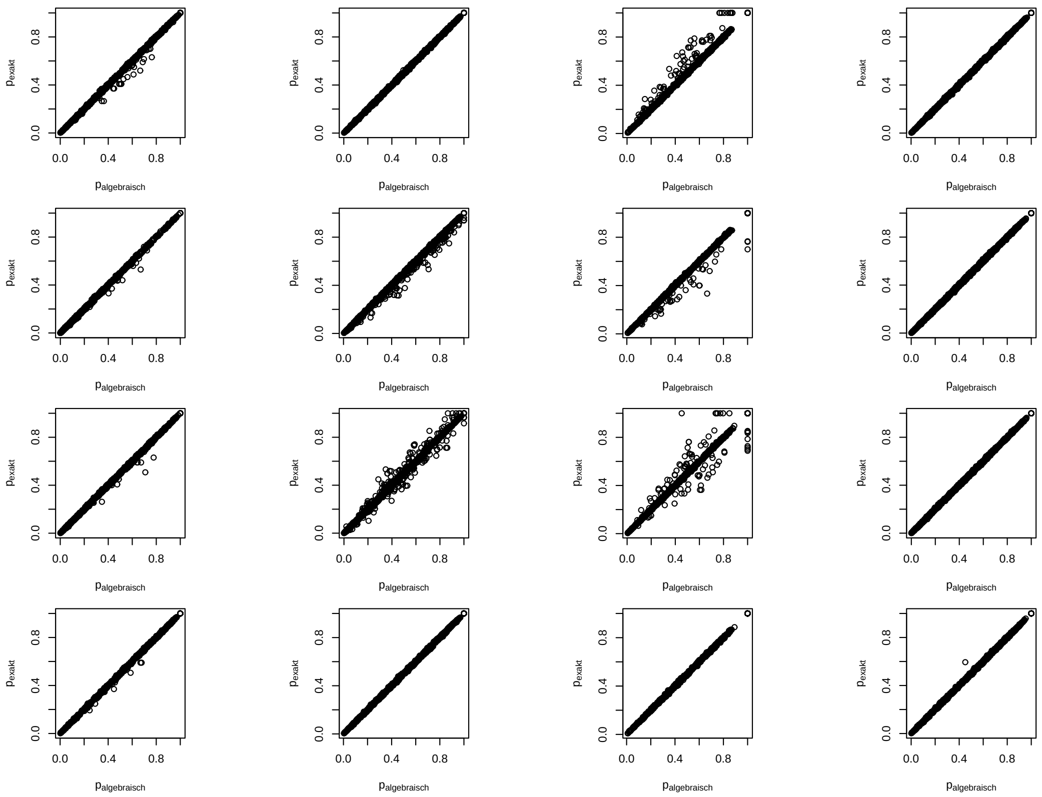

Abbildung A.5: Vergleich der exakten und algebraischen p-Werte der Likelihood-Quotienten-Tests. Das zu testende Modell wird repräsentiert durch die Spalten (S-, CS-, DS-, OQS-Modell, von links nach rechts). Die Datengenerierung wird durch die Zeilen wiedergegeben (S-, CS-, DS-, OQS-Modell, von oben nach unten, siehe Tabelle 4.1 (i)-(iv) mit $n = 25$).

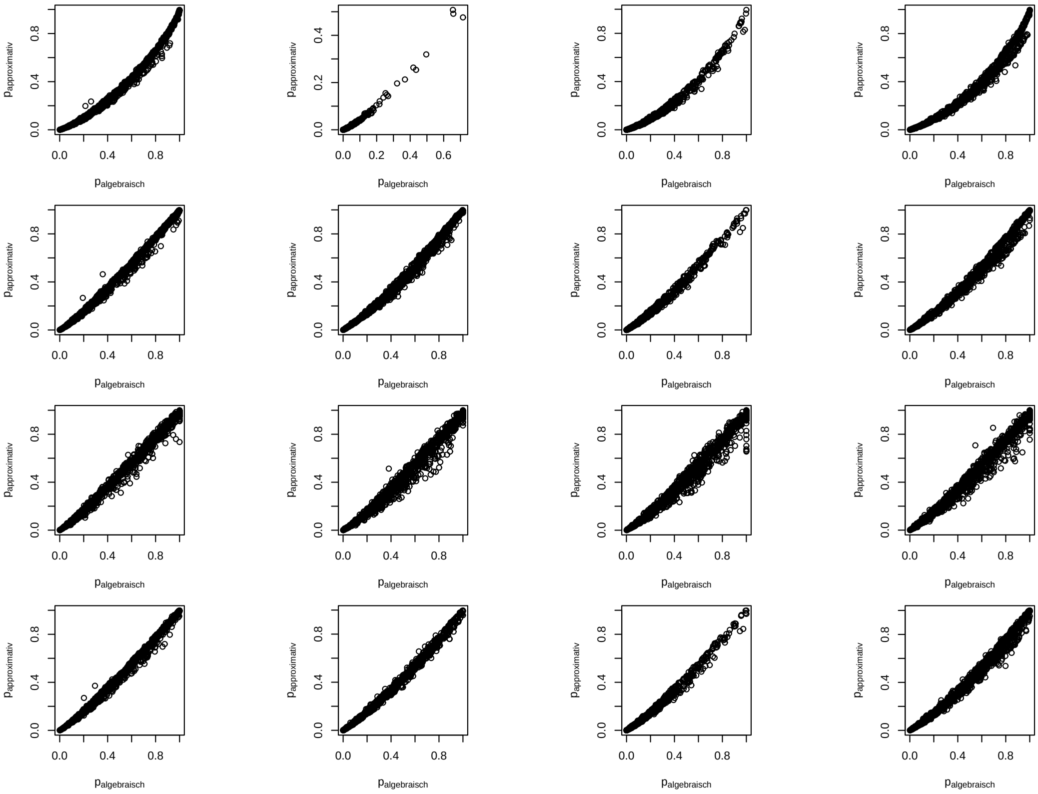

Abbildung A.6: Vergleich der approximativen und algebraischen p-Werte der Likelihood-Quotienten-Tests. Das zu testende Modell wird repräsentiert durch die Spalten (S-, CS-, DS-, OQS-Modell, von links nach rechts). Die Datengenerierung wird durch die Zeilen wiedergegeben (S-, CS-, DS-, OQS-Modell, von oben nach unten, siehe Tabelle 4.1 (i)-(iv) mit $n = 100$).

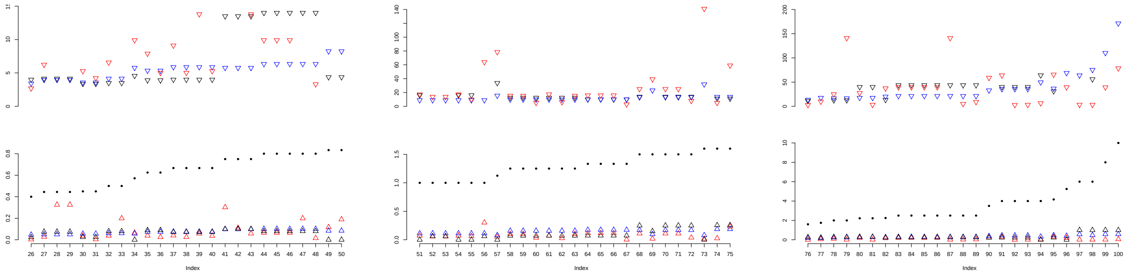

Abbildung A.7: Vergleich der approximativen, exakten und algebraischen 95% Konfidenzgrenzen (obere Grenze: Bild oben; untere Grenze: Bild unten) für die 26-50 (links), 51-75 (Mitte) bzw. 76-100 (rechts) kleinsten ML-Schätzwerte für das Odds Ratio aus Modell (i), aus Tabelle 4.6 mit $n = 15$. Der Wert des geschätzten Odds Ratios wird durch Punkte repräsentiert. Die Konfidenzintervalle werden farblich unterschieden (approximativ: blau, exakt: rot, algebraisch: schwarz).

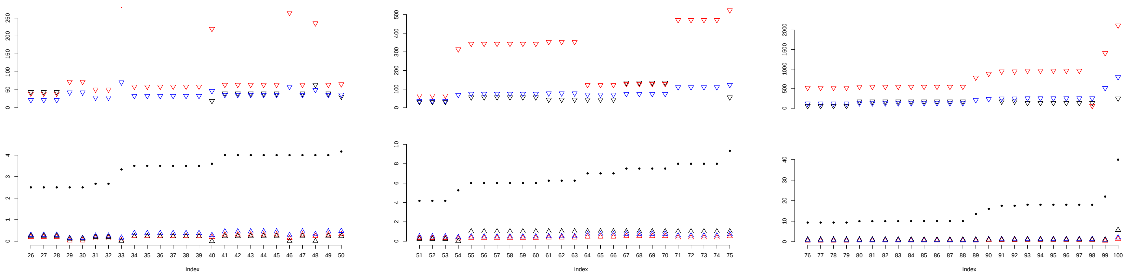

Abbildung A.8: Vergleich der approximativen, exakten und algebraischen 95% Konfidenzgrenzen (obere Grenze: Bild oben; untere Grenze: Bild unten) für die 26-50 (links), 51-75 (Mitte) bzw. 76-100 (rechts) kleinsten ML-Schätzwerte für das Odds Ratio aus Modell (ii), aus Tabelle 4.6 mit $n = 15$. Der Wert des geschätzten Odds Ratios wird durch Punkte repräsentiert. Die Konfidenzintervalle werden farblich unterschieden (approximativ: blau, exakt: rot, algebraisch: schwarz).

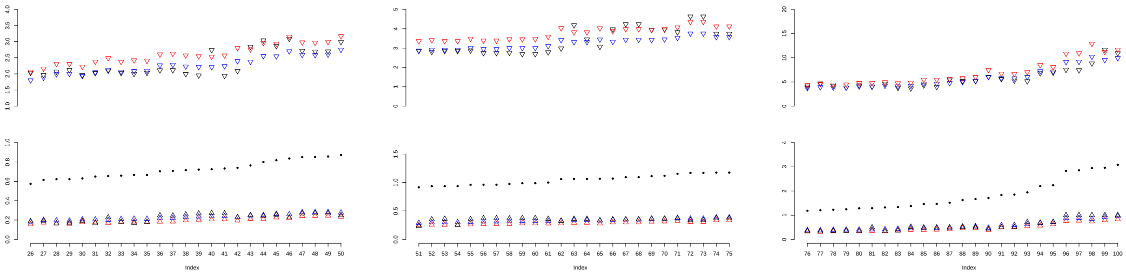

Abbildung A.9: Vergleich der approximativen, exakten und algebraischen 95% Konfidenzgrenzen (obere Grenze: Bild oben; untere Grenze: Bild unten) für die 26-50 (links), 51-75 (Mitte) bzw. 76-100 (rechts) kleinsten ML-Schätzwerte für das Odds Ratio aus Modell (i), aus Tabelle 4.6 mit $n = 50$. Der Wert des geschätzten Odds Ratios wird durch Punkte repräsentiert. Die Konfidenzintervalle werden farblich unterschieden (approximativ: blau, exakt: rot, algebraisch: schwarz).

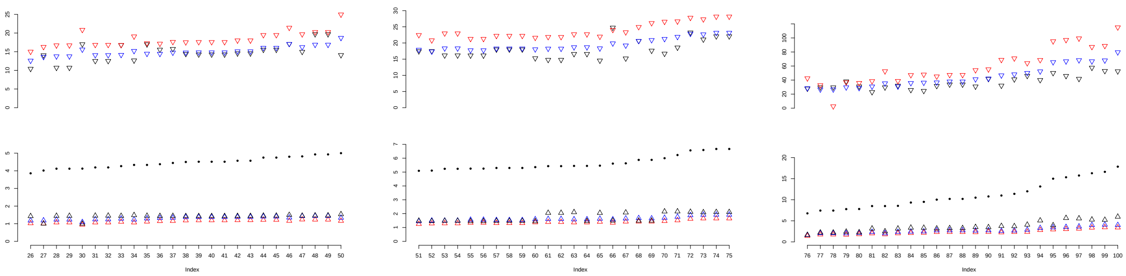

Abbildung A.10: Vergleich der approximativen, exakten und algebraischen 95% Konfidenzgrenzen (obere Grenze: Bild oben; untere Grenze: Bild unten) für die 26-50 (links), 51-75 (Mitte) bzw. 76-100 (rechts) kleinsten ML-Schätzwerte für das Odds Ratio aus Modell (ii), aus Tabelle 4.6 mit $n = 50$. Der Wert des geschätzten Odds Ratios wird durch Punkte repräsentiert. Die Konfidenzintervalle werden farblich unterschieden (approximativ: blau, exakt: rot, algebraisch: schwarz).

A.3 Anhang zur Modellselektion

Beispiel A.3.1

Gegeben sei eine $2 \times 2 \times 2$-Tafel und es interessieren die in Tabelle 5.1 angegebenen log-linearen Modelle. Dann sind die entsprechenden $T^{*(i)}$, $i = 1, \ldots, 7$ gegeben durch

$T^{*(1)}((1,1,1)) = (1,0,0,0,1,0,0,0)'$, $\quad T^{*(1)}((1,1,2)) = (1,0,0,0,0,1,0,0)'$,

$T^{*(1)}((1,2,1)) = (0,1,0,0,0,0,1,0)'$, $\ldots$, $T^{*(1)}((2,2,2)) = (0,0,0,1,0,0,0,1)'$;

$T^{*(2)}((1,1,1)) = (1,0,0,0,1,0,0,0)'$, $\quad T^{*(2)}((1,1,2)) = (1,0,0,0,0,1,0,0)'$,

$T^{*(2)}((1,2,1)) = (0,1,0,0,1,0,0,0)'$, $\ldots$, $T^{*(2)}((2,2,2)) = (0,0,0,1,0,0,0,1)'$;

$T^{*(3)}((1,1,1)) = (1,0,0,0,1,0,0,0)'$, $\quad T^{*(3)}((1,1,2)) = (0,1,0,0,0,1,0,0)'$,

$T^{*(3)}((1,2,1)) = (1,0,0,0,0,0,1,0)'$, $\ldots$, $T^{*(3)}((2,2,2)) = (0,0,0,1,0,0,0,1)'$;

$T^{*(4)}((1,1,1)) = (1,0,0,0,1,0)'$, $\quad T^{*(4)}((1,1,2)) = (1,0,0,0,0,1)'$,

$T^{*(4)}((1,2,1)) = (0,1,0,0,1,0)'$, $\ldots$, $T^{*(4)}((2,2,2)) = (0,0,0,1,0,1)'$;

$T^{*(5)}((1,1,1)) = (1,0,1,0,0,0)'$, $\quad T^{*(5)}((1,1,2)) = (1,0,0,1,0,0)'$,

$T^{*(5)}((1,2,1)) = (1,0,0,0,1,0)'$, $\ldots$, $T^{*(5)}((2,2,2)) = (0,1,0,0,0,1)'$;

$T^{*(6)}((1,1,1)) = (1,0,1,0,0,0)'$, $\quad T^{*(6)}((1,1,2)) = (1,0,0,1,0,0)'$,

$T^{*(6)}((1,2,1)) = (0,1,1,0,0,0)'$, $\ldots$, $T^{*(6)}((2,2,2)) = (0,1,0,0,0,1)'$;

$T^{*(7)}((1,1,1)) = (1,0,1,0,1,0)'$, $\quad T^{*(7)}((1,1,2)) = (1,0,1,0,0,1)'$,

$T^{*(7)}((1,2,1)) = (1,0,0,1,1,0)'$, $\ldots$, $T^{*(7)}((2,2,2)) = (0,1,0,1,0,1)'$.

	Wahrscheinlichkeit			Wahrscheinlichkeit	
exakt	algebraisch	neues Verfahren	exakt	algebraisch	neues Verfahren
0,29	0,28	0,27	0,21	0,22	0,21
0,21	0,22	0,21	0,21	0,20	0,21
0,19	0,18	0,21	0,21	0,22	0,22
0,14	0,15	0,14	0,04	0,04	0,04
0,05	0,05	0,05	0,04	0,03	0,04
0,07	0,07	0,07	0,21	0,21	0,22
0,03	0,03	0,03	0,04	0,03	0,04
0,02	0,02	0,02	0,04	0,04	0,04

Tabelle A.1: Exakte hypergeometrische und algebraische Wahrscheinlichkeiten für alle Elemente von $\mathscr{Z}_{t^{(1)}}$ (links) und $\mathscr{Z}_{t^{(3)}}$ (rechts) des Datensatzes 5.1.

	Wahrscheinlichkeit			Wahrscheinlichkeit	
exakt	algebraisch	neues Verfahren	exakt	algebraisch	neues Verfahren
0,01	0,02	0,01	0,01	0,01	0,01
0,09	0,09	0,09	0,02	0,02	0,02
0,04	0,04	0,04	0,01	0,01	0,01
0,06	0,06	0,06	0,04	0,04	0,04
0,09	0,09	0,09	0,01	0,01	0,01
0,04	0,05	0,04	0,01	0,01	0,01
0,04	0,05	0,04	0,01	0,02	0,01
0,01	0,01	0,01	0,01	0,01	0,01
0,01	0,01	0,02	0,01	0,01	0,01
0,01	0,01	0,01	0,01	0,00	0,00
0,02	0,02	0,02	0,00	0,00	0,00
0,04	0,04	0,04	0,00	0,00	0,00
0,03	0,03	0,03	0,01	0,01	0,01
0,04	0,04	0,04	0,00	0,00	0,00
0,06	0,06	0,06	0,00	0,00	0,00
0,03	0,03	0,03	0,00	0,00	0,00
0,02	0,02	0,02	0,01	0,01	0,01
0,12	0,12	0,12	0,00	0,00	0,00
0,01	0,01	0,01	0,00	0,00	0,00
0,02	0,01	0,02	0,00	0,00	0,00
0,01	0,01	0,01	0,00	0,00	0,00
0,01	0,01	0,01	0,00	0,00	0,00

Tabelle A.2: Exakte hypergeometrische und algebraische Wahrscheinlichkeiten für alle Elemente von $\mathscr{Z}_{t^{(4)}}$ des Datensatzes 5.1.

Wahrscheinlichkeit			Wahrscheinlichkeit		
exakt	algebraisch	neues Verfahren	exakt	algebraisch	neues Verfahren
0,02	0,02	0,02	0,00	0,00	0,00
0,00	0,00	0,01	0,05	0,05	0,05
0,08	0,08	0,08	0,01	0,01	0,01
0,08	0,08	0,08	0,08	0,07	0,08
0,05	0,05	0,06	0,01	0,01	0,01
0,08	0,07	0,08	0,03	0,03	0,03
0,06	0,07	0,06	0,03	0,02	0,03
0,04	0,04	0,04	0,02	0,02	0,02
0,12	0,13	0,12	0,01	0,01	0,01
0,01	0,01	0,01	0,01	0,01	0,01
0,05	0,05	0,06	0,01	0,01	0,01
0,04	0,04	0,04	0,01	0,01	0,01
0,03	0,03	0,03	0,00	0,00	0,00
0,01	0,01	0,01	0,01	0,01	0,01
0,00	0,00	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,01	0,01	0,01
0,00	0,00	0,00	0,00	0,00	0,00
0,01	0,01	0,01	0,00	0,00	0,00
0,01	0,01	0,01	0,00	0,00	0,00
0,01	0,01	0,01	0,00	0,00	0,00

Tabelle A.3: Exakte hypergeometrische und algebraische Wahrscheinlichkeiten für alle Elemente von $\mathscr{Z}_{t^{(5)}}$ des Datensatzes 5.1.

Wahrscheinlichkeit			Wahrscheinlichkeit		
exakt	algebraisch	neues Verfahren	exakt	algebraisch	neues Verfahren
0,04	0,04	0,05	0,01	0,01	0,01
0,04	0,04	0,05	0,04	0,05	0,05
0,04	0,04	0,05	0,01	0,01	0,01
0,01	0,02	0,01	0,01	0,01	0,01
0,01	0,01	0,01	0,01	0,01	0,01
0,14	0,13	0,13	0,02	0,02	0,02
0,07	0,07	0,07	0,01	0,01	0,01
0,04	0,05	0,04	0,01	0,01	0,01
0,04	0,04	0,04	0,01	0,01	0,01
0,04	0,05	0,04	0,01	0,01	0,01
0,01	0,01	0,01	0,01	0,01	0,01
0,07	0,07	0,07	0,00	0,01	0,01
0,02	0,02	0,02	0,01	0,01	0,01
0,01	0,01	0,01	0,00	0,01	0,00
0,00	0,00	0,00	0,02	0,03	0,02
0,01	0,01	0,01	0,01	0,01	0,01
0,04	0,04	0,04	0,00	0,00	0,00
0,07	0,07	0,07	0,01	0,01	0,01
0,04	0,05	0,04	0,00	0,00	0,00
0,01	0,01	0,01	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	0,00

Tabelle A.4: Exakte hypergeometrische und algebraische Wahrscheinlichkeiten für alle Elemente von $\mathscr{Z}_{t(6)}$ des Datensatzes 5.1.

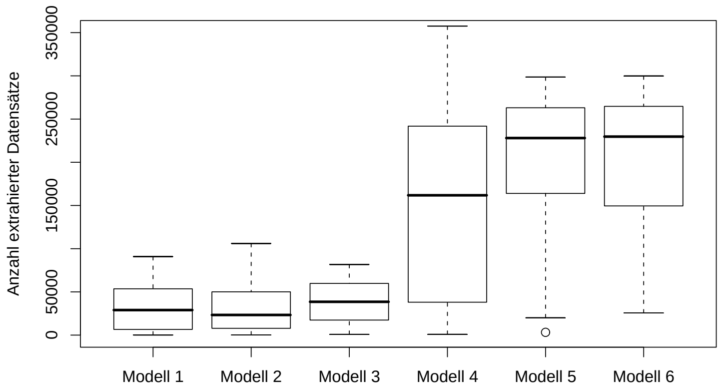

Abbildung A.11: Boxplots der Anzahl selektierter Datensätze der Modelle 1–6 bei zugrunde liegendem Simulationsmodell 4.

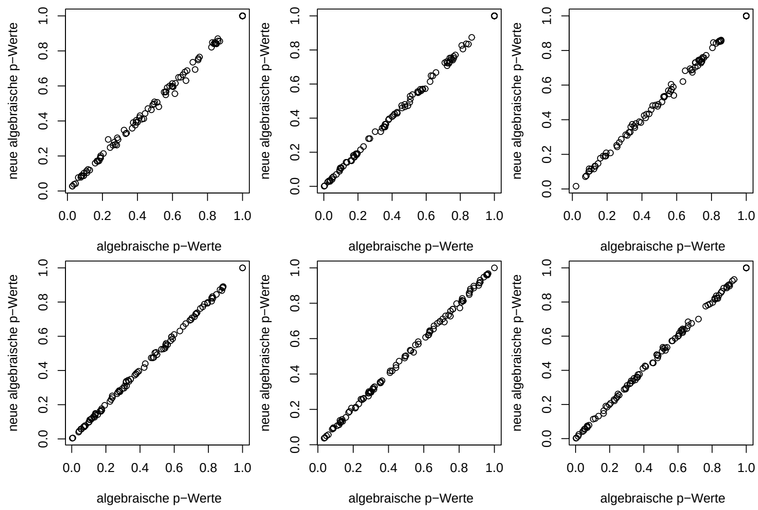

Abbildung A.12: Streudiagramme der p-Werte des χ^2-Anpassungstests des bekannten Diaconis-Sturmfels-Verfahrens und der neuen Prozedur für die Modelle 1–6 (beginnend oben von links nach rechts) bei zugrunde liegendem Simulationsmodell 4.

Symbolverzeichnis

Allgemein

$I\!N$	Menge der natürlichen Zahlen
$\mathbb{Z}$	Menge der ganzen Zahlen
$\mathbb{Z}_{\geq 0}$	$I\!N \cup \{0\} = I\!N_0$
$\mathbb{Q}$	Menge der rationalen Zahlen
$I\!R$	Menge der rationalen Zahlen
$\mathbb{C}$	Menge der komplexen Zahlen
$N \subseteq M$	N ist Teilmenge von M
$N \subset M$	N ist echte Teilmenge von M
$N \cup M$	Vereinigung der Mengen N und M
$N \cap M$	Durchschnitt der Mengen N und M
$\|N\|$	Mächtigkeit der Menge N
$N \setminus \{n\}$	Menge N ohne Element n
$\emptyset$	leere Menge

Strukturmodellierung

$\Delta = \{1, \ldots, q\}$	endliche Indexmenge für kategoriale Variablen, synonym für Menge kategorialer Variablen
$X_\Delta = (X_\delta)_{\delta \in \Delta}$	q-dimensionaler, kategorialer Zufallsvektor
i_δ	Realisation von X_δ, $\delta \in \Delta$

$\mathcal{I}_\delta = \{1, \ldots, I_\delta\}$	endliche Menge möglicher Realisationen eines kategorialen Merkmals
$i = (i_\delta)_{\delta \in \Delta}$	Zelle einer Kontingenztafel
$\mathcal{I} = \times_{\delta \in \Delta}$	Kartesisches Produkt über alle $I_\delta, \delta \in \Delta$
n	Stichprobenumfang
$\pi_i = \pi_{(i_1, \ldots, i_\Delta)'}$	Wahrscheinlichkeit für Ausprägung $X_\Delta = i_\Delta$
n_{i_a}	Zelleinträge einer Randtafel
$\lambda_{i_a}^a$	Effekte eines log-linearen Modells
m_i	erwartete Zellhäufigkeit
$d = \{d_1, \ldots, d_l\}$	Erzeugendenmenge eines hierarchischen log-linearen Modells
$\mathcal{G} = (V, E)$	Graph mit Kantenmenge E und Knotenmenge V
$bd(v)$	Menge aller Nachbarn eines Knoten $v \in V$
$cl(A)$	$bd(A) \cup A$, $A \subseteq V$ Abschluss von A
$\mathcal{G}_A$	durch A induzierter Untergraph von $\mathcal{G}$
Γ	synonym für Menge stetiger Variablen
$\mathscr{H}$	endliche, nicht-leere Menge, Stichprobenraum
$(\mathscr{H}, A, \mathcal{P})$	statistisches Modell
$(\mathscr{H}, A, P)$	Wahrscheinlichkeitsraum
$X_i \perp\!\!\!\perp X_j \mid X_{V \setminus \{i,j\}}$	X_i unabhängig von X_j, gegeben $X_{V \setminus \{i,j\}}$
$M(\mathcal{G})$	graphisches Modell

Algebraische Statistik

$p(x) = \sum_{i \in \mathbb{N}_0} a_i x^i$	Polynom in der Unbestimmten x über dem Körper K, $a_i \in K$
$deg(p)$	Grad des Polynoms p
$x^\alpha := x_1^{\alpha_1} \cdot x_2^{\alpha_2} \cdots x_n^{\alpha_n}$	Monom in $x_1, \ldots, x_n$ mit $\alpha_1, \ldots, \alpha_n \in \mathbb{N}_0$
$K[x_1, \ldots, x_n]$	kommutativer Polynomring mit Einselement in den Unbestimmten $x_1, \ldots, x_n$ über einem Körper K

$V(p_1, \dots, p_s)$	affine Varietät
$\mathscr{I}$	Polynomideal oder kurz: Ideal
$< p_1, \dots, p_s >$	ein durch $p_1, \dots, p_s$ erzeugtes Ideal
$\succ$	Monomordnung
	lex: lexikographische Ordnung
	grlex: gradlexikographische Ordnung
$LC(p)$	Leitkoeffizient von p
$LM(p)$	Leitmonom von p
$LT(p)$	Leitterm von p
$\mathscr{I} =< x^\alpha : \alpha \in \mathbb{Z}^n_{\geq 0} >$	Monomideal
$LT(\mathscr{I})$	Menge der Leitterme von Elementen aus dem Ideal $\mathscr{I}$
$\mathscr{I} =< LT(\mathscr{I}) >$	Leittermideal
$V(\mathscr{I})$	Varietät des Ideals $\mathscr{I}$
$\mathscr{G} = \{g_1, \dots, g_t\}$	Gröbner Basis
$\mathscr{X} = \{X_s, s \in S\}$	Stochastischer Prozess mit Parameterraum S und Zustandsraum E
$\mathtt{P}$	Übergangsmatrix
π^∞	Grenzverteilung der Markov Kette
π	stationäre Verteilung der Markov Kette
$m_1, \dots, m_L$	Markov Basis
$\mathscr{I}_T$	Torisches Ideal
$\mathscr{Z}_t$	Menge aller Datensätze mit realisierter suffizienter Statistik t
$H(z)$	Hypergeometrische Wahrscheinlichkeitsfunktion
$\mathcal{T}$	Menge der einzelnen Einträge der suffizienten Statistik

Algebraische Testprozeduren und Konfidenzintervalle

χ^2_{Bowker}	χ^2-Teststatistik des Bowker-Tests
χ^2_{Korr}	χ^2-Teststatistik des stetigkeitskorrigierten Bowker-Tests

$\chi^2_{\text{mod. Wald}}$	χ^2-Teststatistik des Tests von May und Johnson
S	perfektes Symmetriemodell
CS	bedingtes Symmetriemodell
DS	diagonales Symmetriemodell
OQS	ordinales Quasi-Symmetriemodell
OR	Odds Ratio
$\widehat{OR}$	ML-Schätzer für das Odds Ratio
KI	Konfidenzintervall

Algebraische Modellselektion

$\mathcal{M}$	Menge betrachteter Modelle
$\mathcal{A}$	Menge akzeptabler Modelle
$\mathcal{R}$	Menge abgelehnter Modelle
$N \subseteq M$	N ist Teilmenge von M
$M1 \subseteq M2$	$M1$ ist ein Untermodell von $M2$
I_a	Menge der schwach akzeptablen Modelle
I_r	Menge der schwach abgelehnten Modelle
D_r	r-Dual
D_a	a-Dual

Literaturverzeichnis

Agresti, A., 1983. A Simple Diagonals-Parameter Symmetry and Quasi-Symmetry Model. *Statistics & Probability Letters*, 1, S. 313–316.

Agresti, A., 1992. Modelling Patterns of Agreement and Disagreement. *Statistical Methods in Medical Research*, 1, S. 201–218.

Agresti, A., 1996. An Introduction to Categorical Data Analysis. Wiley, New York.

Agresti, A., 2002. Categorical Data Analysis, 2nd edition. Wiley, New York.

Allman, E.S., und Rhodes, J.A., 2007. Molecular Phylogenetics From an Algebraic Viewpoint. *Statistica Sinica*, 17, S. 1299–1316.

Aoki, S., und Takemura, A., 2005. Markov Chain Monte Carlo Exact Tests for Incomplete Two-Way Contingency Tables. *Journal of Statistical Computation and Simulation*, 75, S. 787–812.

Artin, M., 1993. Algebra. Birkhäuser, Basel.

Barbone, F., Austin, H., und Partridge, E.E., 1993. Diet and Endometrial Cancer: A Case-Control Study. *American Journal of Epidemiology*, 137, S. 393–403.

Besag, S., 1974. Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society*, 36, S. 192–236.

Bigatti, A., Scala, R., und Robbiano, L., 1999. Computing Toric Ideals. *Journal of Symbolic Computation*, 27, S. 351–365.

Bigatti, A., und Robbiano, L., 2001. Toric Ideals. *Matemática Contemporânea*, 21, S. 1–25.

Bishop, Y.M.M., Fienberg, S.E., und Holland, P.W., 1995. Discrete Multivariate Analysis. The MIT Press, Cambridge, Massachusetts.

Bodendiek, R., und Lang, R., 1995. Lehrbuch der Graphentheorie. Band 1 und 2. Spektrum Akademischer Verlag, Heidelberg.

Booth, J., und Butler, R., 1999. An Importance Sampling Algorithm for Exact Conditional Test in Log-Linear Models. *Biometrika*, 86, S. 321–332.

Bowker, A.H., 1948. A Test for Symmetry in Contingency Tables. *Journal of the American Statistical Association*, 43, S. 572–574.

Breslow, N., 1982. Covariance Adjustment of Relative-Risk Estimates in Matched Studies. *Biometrics*, 38, S. 661–672.

Buchberger, B., und Winkler, F. (Hrsg.), 1998. Gröbner Bases and Applications. London Math. Soc. Lecture Notes 251, Cambridge Univ. Press.

Buchberger, B., 2006. Bruno Buchberger's PhD Thesis 1965: An Algorithm for Finding the Basis Elements of the Residue Class Ring of an Zero Dimensional Polynomial Ideal. *Journal of Symbolic Computation*, 41, S. 475–511.

Burnham, K.P., und Anderson, D.A., 2002. Model Selction and Multimodel Inference. A Practical Information-Theoretic Approach. 2nd edition, Springer, New York.

Caffo, B.S., und Booth, J.G., 2001. A Markov Chain Monte Carlo Algorithm for Approximating Exact Conditional Probabilities. *Journal of Computational and Graphical Statistics*, 10, S. 730–745.

Caputo, A., 1998. Eine alternative Familie von Modellverteilungen für Kovarianz- und Konzentrationsgraphen. Forschungsbericht Nr. 48 aus dem Institut für Statistik der Universität München.

Carbonell, J.G., und Siekmann, J. (Hrsg.), 2004. Artificial Intelligence and Symbolic Computation. Proceedings of the 7th International Conference, AISC 2004, Linz, Austria, September 22-24, 2004. Lecture Notes in Artificial Intelligence 3249. Springer, Berlin, Heidelberg.

Caussinus, H., 1965. Contribution à l' analyse statistique de tableaux de corrélation. *Ann. Fac. Sci. Univ. Toulouse*, 29, S. 77–182.

Chatfield, C., 1995. Model Uncertainty, Data Mining and Statistical Inference. *Journal of the Royal Statistical Society A*, 158, S. 419–466.

Christensen, R., 1997. Log-Linear Models, 2nd edition. Springer, New York.

Chib, S., und Greenberg, E., 1995. Understanding the Metropolis-Hastings-Algorithm. *The American Statistician*, 49, S. 327–335.

Cicchetti, D.V., und Feinstein, A.R., 1990. High Agreement but Low Kappa: II. Resolving the Paradoxes. *Journal of Clinical Epidemiology*, 43, S. 551–558.

Čižmár, J., 2002. Gröbnersche Basen in speziellen Ringen (Aus der Geschichte der Gröbnerschen Basen). *Periodica Polytechnica Ser. Mech. Eng.*, 46, S. 45–57.

Cochran, W.G., 1954. Some Methods for Strengthening the Common χ^2 Tests. *Biometrics*, 10, S. 417–451.

CoCoATeam; CoCoA: A System for Doing Computations in Commutative Algebra. Erhältlich im Internet unter http://cocoa.dima.unige.it

Conover W.J., 1971. Practical Nonparametric Statistics. Wiley, New York.

Conover, W.J., 1974. Some Reasons for not Using the Yates Continuity Correction on 2 x 2 Contingency Tables. *Journal of the American Statistical Association*, 69, S. 374–382.

Cox, D., Little, J., und O'Shea, D., 1997. Ideals, Varieties, and Algorithms, 2nd edition. Springer, New York.

Dahlhaus, R., 2000. Graphical Interaction Models for Multivariate Time Series. *Metrika*, 51, S. 157–172.

Darroch, J.N., Lauritzen, S.L., und Speed, T.P., 1980. Markov Fields and Log-Linear Interaction Models for Contingency Tables. *The Annals of Statistics*, 8, S. 522–539.

Darroch, J.N., und Speed, T.P., 1983. Additive and Multiplicative Models and Interactions. *The Annals of Statistics*, 11, S. 724–738.

Diaconis, P., und Saloff-Coste, L., 1995. What Do We Know About the Metropolis Algorithm. Proceedings of the twenty-seventh annual ACM symposium on Theory of Computing, Dept. Mathematics, Harvard University.

Diaconis, P., und Sturmfels, B., 1998. Algebraic Algorithms for Sampling From Conditional Distributions. *The Annals of Statistics*, 26, S. 363–397.

Dinwoodie, I.H., 2002. Algebraic Methods for Polynomial Statistical Models. *Statistics and Computing*, 12, S. 307–314.

Dobra, A., 2003. Markov Bases for Decomposable Graphical Models. *Bernoulli*, 9, S. 1093–1108.

Dobra, A. and Sullivant, S., 2004. A Divide-and-Conquer Algorithm for Generating Markov Bases of Multi-Way Tables. *Computational Statistics*, 19, S. 347–366.

Drton, M., 2006. Algebraic Techniques for Gaussian Models. Erhältlich im Internet unter http://arxiv.org/abs/math/0610679v1.

Drton, M., Sturmfels B., und Sullivant S., 2007. Algebraic Factor Analysis: Tetrads, Pentads and Beyond. *Probability Theory and Related Fields*, 138, S. 463–493.

Edwards, A.L., 1948. Note on the "Correction for Continuity" in Testing the Significance of the Difference Between Correlated Proportions. *Psychometrika*, 13, S. 185–187.

Edwards, D., und Havránek, T., 1985. A Fast Procedure for Model Search in Multidimensional Contingency Tables. *Biometrika*, 72, S. 339–351.

Edwards, D., und Havránek, T., 1987. A Fast Model Selection Procedure for Large Families of Models. *Journal of the American Statistical Association*, 82, S. 205–213.

Edwards, D., 2000. Introduction to Graphical Modelling. Springer, New York.

El Adlouni, S., Favre, A.-C., und Bobée, B., 2006. Comparison of Methodologies to Assess the Convergence of Markov Chain Monte Carlo Methods. *Computational Statistics and Data Analysis*, 50, S. 2685–2701.

Everitt, B.S., 1977. The Analysis of Contingency Tables. Chapman & Hall/CRC, London.

Fahrmeir, L., Kaufmann, H.L., und Ost, F., 1981. Stochastische Prozesse. Carl Hanser Verlag, München.

Feinstein, A.R., und Cicchetti, D.V., 1990. High Agreement But Low Kappa: I. The Problems of Two Paradoxes. *Journal of Clinical Epidemiology*, 43, S. 543–549.

Ferguson, T.S., 1967. Mathematical Statistics. Academic Press, New York.

Fischer, G., und Sacher, R., 1978. Einführung in die Algebra. 2. überarbeitete Auflage, B.G. Teubner, Stuttgart.

Fleiss, J. L., Levin, B., und Paik, M.C., 2003. *Statistical Methods for Rates and Proportions*. 3rd edition, Wiley, Hoboken.

Frydenberg, M., und Lauritzen, S.L., 1989. Decomposition of Maximum Likelihood in Mixed Graphical Interaction Models. *Biometrika*, 76, S. 539–555.

Gabriel, K.R., 1969. Simultaneous Test Procedures - Some Theory of Multiple Comparisons. *The Annals of Mathematical Statistics*, 40, S. 224–250.

Geiger, D., Meek, C., und Sturmfels, B., 2006. On the Toric Algebra of Graphical Models. *The Annals of Statistics*, 34, S. 1463–1492.

Giglio, B., und Wynn, H.P., 2004. Monomial Ideals and the Scarf Complex for Coherent Systems in Reliability Theory. *The Annals of Statistics*, 32, S. 1289–1331.

Good, I.J., 1973. What are Degrees of Freedom? *The American Statistician*, 27, S. 227–228.

Goodman, L.A., 1979. Multiplicative Models for Square Contingency Tables with Ordered Categories. *Biometrika*, 66, S. 413–418.

Grizzle, J.E., 1967. Continuity Correction in the χ^2-Test for 2 x 2 Tables. *The American Statistician*, 21, S. 28–32.

Haldane, J.B.S., 1955. The Estimation and Significance of the Logarithm of a Ratio of Frequencies. *Annals of Human Genetics*, 20, S. 309–311.

Hastings, W.K., 1970. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, 57, S. 97–109.

Heinhold, J., und Riedmüller, B., 1975. Lineare Algebra und Analytische Geometrie, Teil 1. Carl Hanser Verlag, München.

Heldt, D., Kreuzer, M., Pokutta, S., und Poulisse, H., 2006. Approximate Computation of Zero-Dimensional Polynomial Ideals. Erhältlich im Internet unter http://staff.fim.uni-passau.de/kreuzer/publications.html.

Hoşten, S., und Sullivant, S., 2007. A Finiteness Therorem for Markov Bases of Hierarchical Models. *Journal of Combinatorial Theory, Series A*, 114, S. 311–321.

Kateri, M., und Agresti, A., 2007. A Class of Ordinal Quasi-Symmetry Models for Square Contingency Tables. *Statistics & Probablity Letters*, 77, S. 598–603.

Krampe, A., und Kuhnt, S., 2007. Bowker´s Test for Symmetry and Modifications Within the Algebraic Framework. *Computational Statistics and Data Analysis*, 51, S. 4124–4142.

Krampe, A., und Kuhnt, S., 2008. Model Selection for Contingency Tables with Algebraic Statistics. Eingeladener Beitrag. Erscheint in: Gibilisco, P., Riccomagno, E., Rogatin, M.-P., und Wynn, H. (Hrsg.): Algebraic and Geometric Methods in Statistics. Cambridge University Press, Cambridge.

Kreienbrock, L., und Schach, S., 1995. Epidemiologische Methoden, Gustav Fischer, Stuttgart, Jena.

Kreuzer, M., und Robbiano, L., 2000. Computational Commutative Algebra 1. Springer, Berlin.

Lauritzen, S.L., und Wermuth, N., 1989. Graphical Models for Associations Between Variables, Some of Which are Qualitative and Some Quantitative. *The Annals of Statistics*, 17, S. 31–57.

Lauritzen, S.L., 1998. Graphical Models. Oxford University Press, New York.

Landis, J.R., und Koch, G.G., 1975. A Review of Statistical Methods in the Analysis of Data Arising From Observer Reliability Studies (Part I and II), *Statistica Neerlandica*, 29, S. 101–123 und S. 151–161.

Lindquist, H., und Taraldsen, G., 2005. Monte Carlo Conditioning on a Sufficient Statistic. *Biometrika*, 92, S. 451–464.

Lindquist, H., und Taraldsen, G., 2007. Conditional Monte Carlo Based on Sufficient Statistics with Applications. In: Nair, V. (Hrsg.): Advances in Statistical Modeling and Inference. Essays in Honor of Kjell A. Doksum, S. 545–562.

Lněicka, R., und Matúš, F., 2007. On Gaussian Conditional Independence Structures, *Kybernetika*, 43, S. 327–342.

Mantel, N. and Greenhouse, S.W., 1968. What is the Continuity Correction? *The American Statistician*, 22, S. 27–30.

Mantel, N., und Haenszel, W., 1959. Statistical Aspects of the Analysis of Data From Retrospective Studies of Disease. *Journal of the National Cancer Institute*, 4, S. 719–748.

Matúš, F., 2005. Conditional Independence in Gaussian Vectors and Rings of Polynomials. In: Kern-Isberner, G., Rödder, W., Kulmann, F. (Hrsg.): Proceedings of „Conditionals, Information, and Inference“ - WCII2002, Lecture Notes in Computer Science 3301.

May, W.L., und Johnson W.D., 2001. Symmetry in Square Contingency Tables: Tests of Hypotheses and Confidence Interval Construction. *Journal of Biopharmaceutical Statistics*, 11, S. 23–33.

McCullagh, P., 1978. A Class of Parametric Models for the Analysis of Square Contingency Tables with Ordered Categories. *Biometrika*, 65, S. 413–418.

McCullagh, P., und Nelder, J.A., 1999. Generalized Linear Models, 2nd edition, Chapman & Hall, Boca Raton, Florida.

McNemar, Q., 1947. Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages. *Psychometrika*, 12, S. 153–157.

Mehta, C.R., Patel, N.R., und Gray, R., 1985. Computing an Exact Confidence Interval for the Common Odds Ratio in Several 2 x 2 Contingency Tables. *Journal of the American Statistical Association*, 80, S. 969–973.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., und Teller, E., 1953. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21, S. 1087–1092.

Mood, A.M., Graybill, F.A., und Boes, D.C., 1974. Introduction to the Theory of Statistics, 3rd edition. McGraw-Hill, Singapur.

Nelsen, R.B., 2006. An Introduction to Copulas, 2nd edition, Springer, New York.

Pachter, L., und Sturmfels, B. (Hrsg.), 2005. Algebraic Statistics for Computational Biology. Cambridge University Press, Cambridge, UK.

Pistone, G., und Wynn, H. P., 1996. Generalised Confounding with Gröbner Bases. *Biometrika*, 83, S. 653–666.

Pistone, G., Riccomagno, E., und Wynn, H.P., 2000. Algebraic Statistics. Chapman & Hall, Boca Raton, Florida.

Plackett, R.L., 1964. The Continuity Correction in 2 x 2 tables. *Biometrika*, 51, S. 327–337.

R. The R Project for Statistical Computing. Erhältlich im Internet unter http://www.r-project.org/

Rapallo, F., 2003. Algebraic Markov Bases and MCMC for Two-Way Contingency Tables. *Journal of the American Statistical Association*, 30, S. 385–397.

Rapallo, F., 2005. Algebraic Exact Inference for Rater Agreement Models. *Statistical Methods and Applications*, 14, S. 45–66.

Rapallo, F., 2006. Markov Bases and Structural Zeros. *Journal of Symbolic Computation*, 41, S. 164–172.

Robert, C.P., und Casella, G., 2004. Monte Carlo Statistical Methods, 2nd edition. Springer, New York.

Rubinstein, R.Y., und Kroese, D.P., 2008. Simulation and the Monte Carlo Method, 2nd edition. John Wiley & Sons, Inc., Hoboken, New Jersey.

Sørensen, D., und Gianola, D., 2002. Likelihood, Bayesian, and MCMC Methods in Qualitative Genetics. Springer, New York.

Sturmfels, B., 2002. Solving Systems of Polynomial Equations. Amer. Math. Soc., CBMS Regional Conferences Series, No 97, Providence, Rhode Island.

Sturmfels, B., und Sullivant, S., 2005. Toric Ideals of Phylogenetic Invariants. *Journal of Computational Biology*, 12, S. 204–228.

Takemura, A., und Aoki, S., 2004. Some Characteristics of Minimal Markov Basis for Sampling From Discrete Conditional Distributions. *Annals of the Institute of Statistical Mathematics*, 56, S. 1–17.

Trotter, H.F., und Tukey, J.W., 1956. Conditional Monte Carlo for Normal Samples. In: Meyer, H.A. (Hrsg.): Symposium on Monte Carlo Methods, S. 64–79, Wiley, New York.

Viana, M.A.G., und Richards, D.S.P. (Hrsg.), 2002. Algebraic Methods in Statistics and Probability, Contemporary Mathematics, American Mathematical Society, Providence, Rhode Island.

Volkmann, L., 1996. Fundamente der Graphentheorie, Springer, Wien.

Weihs, C., und Jessenberger, J., 1999. Statistische Methoden zur Qualitätssicherung und -optimierung, Wiley-VCH, Weinheim.

Wermuth, N., 1976. Analogies Between Multiplicative Models in Contingency Tables and Covariance Selection. *Biometrics*, 32, S. 95–108.

Whittaker, J., 1990. Graphical Models in Applied Mathematical Multivariate Statistics, Wiley, New York.

Witting, H., 1985. Mathematische Statistik, B.G. Teubner, Stuttgart.

Zelen, M., 1971. The Analysis of Several 2 x 2 Contingency Tables. *Biometrika*, 58, S. 129–137.