
Studienabschlussarbeiten
Fakultät für Mathematik, Informatik
und Statistik

Nießl, Christina:

# CLG Bayes-Netze versus Pfadmodelle - Ein Vergleich zweier graphischer Modelle

**Bachelorarbeit, Sommersemester 2017**

Fakultät für Mathematik, Informatik und Statistik

Ludwig-Maximilians-Universität München

https://doi.org/10.5282/ubm/epub.40351

Ludwig-Maximilians-Universität

Institut für Statistik


Bachelorarbeit

---

# CLG Bayes-Netze versus Pfadmodelle - Ein Vergleich zweier graphischer Modelle

---

**Autorin:** Christina Nießl

**Betreuer:** Prof. Dr. Thomas Augustin

Eva Endres

**Datum:** 04. April 2017

# Abstract

Sowohl in CLG Bayes-Netzen als auch in Pfadmodellen werden Zufallsvariablen und die Abhängigkeitsstrukturen zwischen diesen graphisch dargestellt. Die vorliegende Arbeit beschreibt zunächst Aufbau und Eigenschaften der beiden Modelle und vergleicht diese anschließend hinsichtlich verschiedener Kriterien. Neben der graphischen Darstellung von Abhängigkeitsstrukturen werden dabei noch weitere Gemeinsamkeiten aber auch einige Unterschiede festgestellt: Beide Ansätze sind zur Modellierung von kausalen Beziehungen geeignet, in CLG Bayes-Netzen stellt dies jedoch keine Notwendigkeit dar. Bezüglich der Variablen im Modell und deren Abhängigkeiten sind Pfadmodelle außerdem oftmals eingeschränkter als CLG Bayes-Netze. Beim Vergleich der Methoden zur Bestimmung weiterer Abhängigkeiten ergibt sich, dass die Stärke des Einflusses in CLG Bayes-Netzen nur für direkte Abhängigkeiten und in Pfadmodellen nur für kausale Effekte bestimmt wird. Besonders in diesem Punkt wird dabei der wichtigste Unterschied deutlich, der in den jeweiligen Anwendungszielen der Modelle liegt: Während in CLG Bayes-Netzen die kompakte Repräsentation der Verteilung genutzt wird, um Schlussfolgerungen unter Unsicherheit zu ziehen, werden in Pfadmodellen kausale Zusammenhänge analysiert. Bei der Entscheidung, welches Modell verwendet wird, sollte daher vor allem dieser Punkt berücksichtigt werden.

# Notation

Nachfolgend werden einige der in dieser Arbeit verwendeten Notationen aufgeführt. Sofern nicht anders definiert, stellen große Buchstaben Zufallsvariablen und kleine Buchstaben deren Realisierungen dar.

| | |
|---|---|
| $Val(X)$ | Wertebereich der Zufallsvariable $X$ |
| $p(x)$ | Dichte der Zufallsvariable $X$ |
| $P(X = x)$ | Wahrscheinlichkeit, mit der $X$ den Wert $x \in Val(X)$ annimmt |
| $P$ | Wahrscheinlichkeitsverteilung |
| $\perp\!\!\!\perp_P$ | stochastisch unabhängig bezüglich $P$ |
| $\mathcal{I}(P)$ | Menge der stochastischen Unabhängigkeiten, die in $P$ gelten |
| $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ | Graph mit Knotenmenge $\mathcal{V}$ und Kantenmenge $\mathcal{E}$ |
| $Pa_{\nu_i}$ | Eltern von Knoten $\nu_i$ |
| $\perp_{\mathcal{G}}$ | d-separiert in $\mathcal{G}$ |
| $\mathcal{I}(\mathcal{G})$ | globale Markov Unabhängigkeiten in $\mathcal{G}$ |
| | **CLG Bayes-Netze** |
| $\mathcal{B} = (\mathcal{G}, P)$ | Bayes-Netz mit Graph $\mathcal{G}$ und Verteilung $P$ |
| $\boldsymbol{X^D}$ | Menge der diskreten Variablen in einem CLG Bayes-Netz |
| $\boldsymbol{X^C}$ | Menge der stetigen Variablen in einem CLG Bayes-Netz |
| $\boldsymbol{D}$ | diskrete Eltern einer stetigen Variable |
| $\boldsymbol{C}$ | stetige Eltern einer stetigen Variable |
| $\delta_{\boldsymbol{d},0}, \boldsymbol{\delta_d}$ | Regressionskoeffizienten einer CLG Verteilung |
| $\epsilon$ | Menge der beobachteten Variablen |
| | **Pfadmodelle** |
| $\boldsymbol{Z}$ | Vektor der exogenen Variablen |
| $\boldsymbol{Y}$ | Vektor der endogenen Variablen |
| $\boldsymbol{\zeta}$ | Vektor der Fehlerterme |
| $\boldsymbol{B}, \boldsymbol{\Gamma}$ | Matrix der Pfadkoeffizienten für endogene / exogene Einflussgrößen |
| $\boldsymbol{\Psi}, \boldsymbol{\Phi}$ | Kovarianzmatrix der Fehlerterme / exogenen Variablen |
| $\boldsymbol{S}$ | empirische Kovarianzmatrix der beobachtbaren Variablen |
| $\boldsymbol{\Sigma(\theta)}$ | vom Modell implizierte Kovarianzmatrix der beobachtbaren Variablen |

# Inhaltsverzeichnis


**1 Einleitung** **6**

**2 Grundlagen** **8**

2.1 Kausalität . . . . . 8

2.2 Unabhängigkeit von Zufallsvariablen . . . . . 8

2.3 Graphentheorie . . . . . 9

**3 CLG Bayes-Netze** **12**

3.1 Repräsentation . . . . . 12

3.2 Inferenz . . . . . 15

3.3 Lernen . . . . . 16

**4 Pfadmodelle** **18**

4.1 Aufbau . . . . . 18

4.1.1 Pfaddiagramme . . . . . 18

4.1.2 Strukturgleichungen . . . . . 20

4.1.3 Annahmen . . . . . 21

4.2 Schätzung . . . . . 22

4.3 Effektzerlegung . . . . . 23

4.4 Anpassungsgüte . . . . . 24

**5 Vergleich der Modelle** **26**

5.1 Graphische Darstellung . . . . . 26

5.2 Kausalität . . . . . 27

5.3 Variablen im Modell . . . . . 28

5.4 Modellierung der Abhängigkeiten . . . . . 29

5.5 Inferenz versus Effektzerlegung . . . . . 32

**6 Fazit** **34**

**Literaturverzeichnis** **36**

**Abbildungsverzeichnis** **38**

**Elektronischer Anhang** **39**

# 1 Einleitung

Um komplexe Zusammenhänge zwischen Variablen übersichtlich und intuitiv darzustellen, werden häufig graphische Modelle verwendet. Die Idee, Abhängigkeitsstrukturen graphisch zu veranschaulichen, findet ihren Ursprung in verschiedenen Bereichen der Wissenschaft und existiert seit vielen Jahren. Bereits Anfang der 1920er Jahre verwendete der Genetiker Sewall Wright gerichtete Graphen, um die Eigenschaften von Verteilungen zu analysieren. Unter der Annahme von bestimmten kausalen Zusammenhängen berechnete Wright die Korrelationen zwischen Variablen und verglich das Ergebnis mit den beobachteten Korrelationen. Die in dem sogenannten Pfadmodell vermuteten Zusammenhänge stellte er graphisch in einem Pfaddiagramm dar (Wright, 1921). Die Methode, die dementsprechend Pfadanalyse genannt wurde, wird bis heute insbesondere in den Verhaltenswissenschaften häufig angewendet (Mueller, 1996, S. 22).
Ebenfalls eine graphische Darstellung von Zufallsvariablen und deren Abhängigkeitsstrukturen liegt den sogenannten Bayes-Netzen zugrunde. Hier wird die kompakte Darstellung der gemeinsamen Verteilung verwendet, um effizient Schlussfolgerungen unter Unsicherheit durchzuführen, was auch als Inferenz in Bayes-Netzen bezeichnet wird. Ursprünglich im Bereich der künstlichen Intelligenz unter anderem von Pearl (1982) entwickelt, werden Bayes-Netze mittlerweile auch in zahlreichen weiteren Wissenschaftsbereichen wie beispielsweise Risikomanagement, Epidemiologie oder Forensik angewendet. Der Name *Bayes*-Netz leitet sich dabei aus dem Satz von Bayes ab, der bei der Durchführung der Inferenz in Bayes-Netzen eine zentrale Rolle spielt (Koski and Noble, 2012, S. 53-57). Abhängig davon, welche Variablen das Modell beinhaltet und welche Annahmen getroffen werden, ergeben sich verschiedene Bayes-Netze. In dieser Arbeit werden CLG Bayes-Netze behandelt, die sowohl diskrete als auch stetige Variablen beinhalten.

Da sowohl CLG Bayes-Netze als auch Pfadmodelle Abhängigkeitsstrukturen graphisch darstellen, stellt sich die Frage, welche Unterschiede beziehungsweise Gemeinsamkeiten die beiden Modelle aufweisen und in welchen Situationen welches Modell geeignet ist. Ziel dieser Arbeit ist es daher, einen Überblick über CLG Bayes-Netze und Pfadmodelle zu geben und diese hinsichtlich verschiedener Kriterien zu vergleichen, wodurch die Wahl eines passenden Modells erleichtert werden soll.
Zu diesem Zweck werden in Kapitel 2 zunächst einige für beide Modelle grundlegende Begriffe erläutert. Anschließend werden in den Kapiteln 3 und 4 der Aufbau und die Eigenschaften von CLG Bayes-Netzen und Pfadmodellen beschrieben. Kapitel 5 beinhal-

tet den Vergleich der beiden Modelle. Darin werden CLG Bayes-Netze und Pfadmodelle nicht nur hinsichtlich ihrer graphischen Darstellung, sondern auch bezüglich Kausalität, Variablen, Modellierung der Abhängigkeiten und Bestimmung weiterer Zusammenhänge untersucht. In Kapitel 6 werden abschließend die wichtigsten Erkenntnisse in einem Fazit zusammengefasst.

# 2 Grundlagen

Das folgende Kapitel beschreibt einige für CLG Bayes-Netze und Pfadmodelle wesentliche Grundlagen. Zuerst wird kurz das Konzept der Kausalität vorgestellt (Abschnitt 2.1) und anschließend die Unabhängigkeit von Zufallsvariablen erläutert (Abschnitt 2.2). Der letzte Abschnitt beinhaltet einige grundlegende Begriffe aus der Graphentheorie (Abschnitt 2.3).

## 2.1 Kausalität

Kausalität ist ein höchst umstrittener Begriff, der Wissenschaftler und Philosophen seit vielen Jahren beschäftigt (Mueller, 1996, S. xii). Im Folgenden werden fünf allgemeine Bedingungen aufgeführt, die nach Kline (2011, S. 98 f.) mindestens erfüllt sein müssen, bevor auf eine Ursache-Wirkungs-Beziehung geschlossen werden kann.
Zunächst wird eine zeitliche Rangfolge gefordert, das heißt die Ursache muss vor der Wirkung eintreten (*temporal precedence*). Eine zusätzliche Bedingung ist dabei, dass die Richtung der kausalen Beziehung korrekt spezifiziert wird (*correct effect priority*), da zum Beispiel in Längsschnittstudien die Ursache möglicherweise nicht sofort beobachtbar ist und der Effekt dadurch unter Umständen vor der Ursache gemessen wird.
Zu den weiteren Voraussetzungen zählt, dass zwischen Ursache und Wirkung ein Zusammenhang gemessen wird (*association*) und dieser auch noch vorhanden ist, wenn zusätzlich auf gemeinsame Ursachen kontrolliert wird oder weitere Ursachen des Effekts berücksichtigt werden (*isolation*).
Sowohl in CLG Bayes-Netzen als auch in Pfadmodellen wird keine deterministische, sondern probabilistische Kausalität modelliert, das heißt die Ursache hat Einfluss auf die Wahrscheinlichkeitsverteilung des Effekts. In diesem Fall müssen die Form der Verteilung sowie die entsprechenden Parameter bestimmt werden (*known distributional form*).

## 2.2 Unabhängigkeit von Zufallsvariablen

Der Begriff der bedingten Unabhängigkeit spielt sowohl in Pfadmodellen als auch in CLG Bayes-Netzen eine zentrale Rolle. In diesem Abschnitt wird die bedingte stochastische Unabhängigkeit von Zufallsvariablen erläutert. Wie sich zeigen wird, gibt es im Zusammenhang mit CLG Bayes-Netzen noch eine weitere Form von Unabhängigkeit, die in Abschnitt 2.3 beschrieben wird.
Im Folgenden bezeichnet $p(\boldsymbol{x})$ die gemeinsame Dichte einer Menge von Zufallsvariablen

$\boldsymbol{X} = \{X_1, \ldots, X_n\}$. Besteht $\boldsymbol{X}$ nur aus diskreten Zufallsvariablen, so gilt:

$$p(\boldsymbol{x}) = \begin{cases} P(\boldsymbol{X} = \boldsymbol{x}) & \boldsymbol{x} \in Val(\boldsymbol{X}) \\ 0 & \text{sonst,} \end{cases} \tag{2.1}$$

wobei $P(\boldsymbol{X} = \boldsymbol{x})$ der Wahrscheinlichkeit entspricht, dass $\boldsymbol{X}$ den Wert $\boldsymbol{x} \in Val(\boldsymbol{X})$ annimmt (Fahrmeir et al., 2009, S. 462).
Seien $\boldsymbol{X}, \boldsymbol{Y}$ und $\boldsymbol{Z}$ Mengen von Zufallsvariablen. Dann ist $\boldsymbol{X}$ bedingt unabhängig von $\boldsymbol{Y}$ gegeben $\boldsymbol{Z}$ bezüglich einer Verteilung $P$, wenn gilt, dass

$$p(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{z}) = p(\boldsymbol{x} \mid \boldsymbol{z}), \;\; \forall \boldsymbol{x} \in Val(\boldsymbol{X}), \; \forall \boldsymbol{y} \in Val(\boldsymbol{Y}), \; \forall \boldsymbol{z} \in Val(\boldsymbol{Z}). \tag{2.2}$$

Die Relation ist symmetrisch und wird als $\boldsymbol{X} \perp\!\!\!\perp_P \boldsymbol{Y} \mid \boldsymbol{Z}$ beziehungsweise $\boldsymbol{Y} \perp\!\!\!\perp_P \boldsymbol{X} \mid \boldsymbol{Z}$ notiert. Ist die Menge $\boldsymbol{Z}$ leer, so kann statt $\boldsymbol{X} \perp\!\!\!\perp_P \boldsymbol{Y} \mid \varnothing$ auch $\boldsymbol{X} \perp\!\!\!\perp_P \boldsymbol{Y}$ geschrieben werden. In diesem Fall werden $\boldsymbol{X}$ und $\boldsymbol{Y}$ als marginal unabhängig bezeichnet (Koller and Friedman, 2009, S. 24, 31).
Die Gesamtheit aller in einer Verteilung $P$ geltenden Unabhängigkeiten wird definiert als $\mathcal{I}(P)$ (Koller and Friedman, 2009, S. 60).

## 2.3 Graphentheorie

Ein Graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ besteht aus einer endlichen Menge von Knoten $\mathcal{V} = \{\nu_1, \ldots, \nu_n\}$ und einer Kantenmenge $\mathcal{E} \subseteq (\mathcal{V} \times \mathcal{V})$. Enthält $\mathcal{E}$ sowohl $(\nu_i, \nu_j)$ als auch $(\nu_j, \nu_i)$, so sind die Knoten $\nu_i$ und $\nu_j$ durch eine ungerichtete Kante $\nu_i$ — $\nu_j$ verbunden. Gilt hingegen $(\nu_i, \nu_j) \in \mathcal{E}$ und gleichzeitig $(\nu_j, \nu_i) \notin \mathcal{E}$, so spricht man von einer gerichteten Kante $\nu_i \rightarrow \nu_j$ mit Elternknoten $\nu_i$ und Kindknoten $\nu_j$. Die Menge aller Eltern- und Kindknoten von $\nu_i$ wird als $Pa_{\nu_i}$ beziehungsweise $Ch_{\nu_i}$ notiert. Enthält ein Graph ausschließlich gerichtete Kanten, so wird dieser als gerichteter Graph bezeichnet (Kjaerulff and Madsen, 2006, S. 4 f.).
Neben einer direkten Kante können zwei Knoten auch durch einen Pfad verbunden sein. Ein Pfad $\rho$ von $\nu_i$ nach $\nu_j$ ist eine Abfolge paarweise verschiedener Knoten, die jeweils durch Kanten verbunden sind. Wird ein gerichteter Graph betrachtet, so sind an jedem Knoten $\nu_k, k \neq i, j$, im Pfad drei verschiedene Kantenrichtungen möglich:

- eine serielle Kantenrichtung $\nu_{k-1} \rightarrow \nu_k \rightarrow \nu_{k+1}$ oder $\nu_{k-1} \leftarrow \nu_k \leftarrow \nu_{k+1}$,
- eine divergierende Kantenrichtung $\nu_{k-1} \leftarrow \nu_k \rightarrow \nu_{k+1}$,
- oder eine konvergierende Kantenrichtung $\nu_{k-1} \rightarrow \nu_k \leftarrow \nu_{k+1}$.

Ein Pfad, der ausschließlich serielle Kantenrichtungen enthält und in dem somit alle Kanten in dieselbe Richtung weisen, ist ein gerichteter Pfad und wird notiert als $\nu_i \overset{\rho}{\rightsquigarrow} \nu_j$.

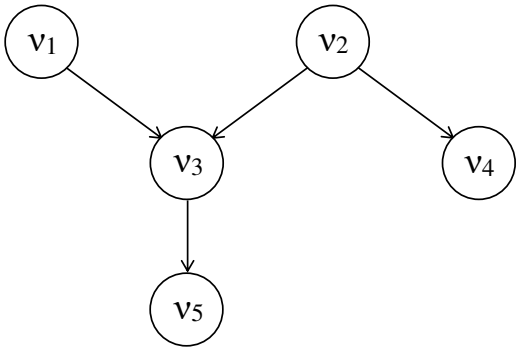



Abbildung 2.1: Beispiel für einen gerichteten, azyklischen Graphen $\mathcal{G}$ mit $\mathcal{V} = \{\nu_1, \nu_2, \nu_3, \nu_4, \nu_5\}$

Die Menge der Knoten, zu denen ein von $\nu_i$ ausgehender, gerichteter Pfad führt, stellt die Nachfahren (engl. *descendants*) von $\nu_i$ dar:

$$Descs_{\nu_i} = \{\nu_j \in \mathcal{V} : \exists \rho : \nu_i \overset{\rho}{\rightsquigarrow} \nu_j\}. \tag{2.3}$$

Umgekehrt beschreibt die Menge der Knoten, die weder Nachfahren von $\nu_i$ noch $\nu_i$ selbst sind, die Nicht-Nachfahren (engl. *non-descendants*) von $\nu_i$:

$$Non\text{-}Descs_{\nu_i} = \mathcal{V} \backslash \{\nu_i\} \backslash Descs_{\nu_i}. \tag{2.4}$$

Besteht ein gerichteter Pfad von $\nu_i$ nach $\nu_j$ und überdies ein gerichteter Pfad von $\nu_j$ nach $\nu_i$, so ergibt sich ein Zyklus. Enthält ein gerichteter Graph keine Zyklen, so wird dieser als azyklisch bezeichnet (Kruse et al., 2011, S. 365-367, 381).
Abbildung 2.1 zeigt ein Beispiel für einen gerichteten, azyklischen Graphen (engl. *directed acyclic graph*, DAG) mit Knotenmenge $\mathcal{V} = \{\nu_1, \nu_2, \nu_3, \nu_4, \nu_5\}$. In diesem Graphen hat beispielsweise Knoten $\nu_3$ einen Kindknoten ($\nu_5$), der zugleich der Menge der Nachfahren entspricht, zwei Eltern ($\nu_1$ und $\nu_2$) sowie einen Nicht-Nachfahren ($\nu_4$).
Abhängig von den in einem Pfad vorliegenden Kantenrichtungen wird dieser als aktiv oder blockiert bezüglich einer Knotenmenge $\omega \subseteq \mathcal{V}$ bezeichnet. Ein Pfad von $\nu_i$ nach $\nu_j$ ist dabei aktiv, wenn er folgende Kriterien erfüllt:

1. Für jeden Knoten $\nu_k$ entlang des Pfades mit konvergierenden Kanten ($\nu_{k-1} \rightarrow \nu_k \leftarrow \nu_{k+1}$) gilt, dass entweder $\nu_k$ oder ein Nachfahre von $\nu_k$ in $\omega$ ist.
2. Kein anderer Knoten entlang des Pfades ist in $\omega$.

Dies impliziert auch, dass weder $\nu_i$ noch $\nu_j$ in $\omega$ sein dürfen (Koller and Friedman, 2009, S. 71). Gelten für einen Pfad die obigen Kriterien nicht, so ist er durch $\omega$ blockiert (Kruse et al., 2011, S. 381).
Betrachtet man als Beispiel den Pfad von $\nu_1$ nach $\nu_4$ in Abbildung 2.1, so ist dieser

durch $\omega = \varnothing$ blockiert, da $\omega$ weder $\nu_3$ noch einen Nachfahren von $\nu_3$ enthält. Gilt jedoch beispielsweise $\omega = \{\nu_5\}$, so ist der Pfad aktiv bezüglich $\omega$, da $\omega$ einen Nachfahren von $\nu_3$ und keinen Knoten ohne konvergierende Kanten ($\nu_1, \nu_2$, oder $\nu_4$) beinhaltet.
Sind alle Pfade von $\nu_i$ nach $\nu_j$ blockiert, so werden $\nu_i$ und $\nu_j$ als d-separiert in $\mathcal{G}$ durch $\omega$ bezeichnet, dies wird notiert als

$$\nu_i \perp_{\mathcal{G}} \nu_j \mid \omega. \tag{2.5}$$

Die Notation der d-Separation lässt bereits erkennen, dass es sich hier um eine weitere Form von Unabhängigkeit handelt. Diese ist ebenfalls symmetrisch, bezieht sich jedoch im Gegensatz zur stochastischen Unabhängigkeit auf die Knoten in Graphen und nicht auf Zufallsvariablen und deren Wahrscheinlichkeiten (Kjaerulff and Madsen, 2006, S. 18 f.).
Die Aussagen, die sich durch d-Separation aus $\mathcal{G}$ ableiten lassen, werden auch als globale Markov Unabhängigkeiten bezeichnet und notiert als $\mathcal{I}(\mathcal{G})$ (Koller and Friedman, 2009, S. 72). Im Allgemeinen sind die meisten aus $\mathcal{G}$ ableitbaren Unabhängigkeitsaussagen nicht direkt ablesbar. Eine Ausnahme bilden die in $\mathcal{I}(\mathcal{G})$ enthaltenen lokalen Markov Unabhängigkeiten $\mathcal{I}_\ell(\mathcal{G})$, die besagen, dass jeder Knoten in $\mathcal{G}$ durch seine Eltern von seinen Nicht-Nachfahren d-separiert ist:

$$\mathcal{I}_\ell(\mathcal{G}) = \{(\nu_i \perp_{\mathcal{G}} \mathit{Non\text{-}Descs}_{\nu_i} \mid Pa_{\nu_i}) : i = 1, \ldots, n\} \tag{2.6}$$

(Koller and Friedman, 2009, S. 57).
Dementsprechend können die lokalen Markov Unabhängigkeiten auch aus dem DAG in Abbildung 2.1 direkt abgelesen werden:

$$\begin{aligned}\mathcal{I}_\ell(\mathcal{G}) = \{&(\nu_1 \perp_{\mathcal{G}} \nu_2, \nu_4 \mid \varnothing), (\nu_2 \perp_{\mathcal{G}} \nu_1 \mid \varnothing), (\nu_3 \perp_{\mathcal{G}} \nu_4 \mid \nu_1, \nu_2), (\nu_4 \perp_{\mathcal{G}} \nu_1, \nu_3, \nu_5 \mid \nu_2),\\ &(\nu_5 \perp_{\mathcal{G}} \nu_1, \nu_2, \nu_4 \mid \nu_3)\}.\end{aligned}$$

Da weder $\nu_1$ noch $\nu_2$ Eltern haben, werden die Unabhängigkeitsaussagen $\nu_1 \perp_{\mathcal{G}} \nu_2, \nu_4 \mid \varnothing$ und $\nu_2 \perp_{\mathcal{G}} \nu_1 \mid \varnothing$ bezüglich der leeren Menge formuliert.

# 3 CLG Bayes-Netze

CLG Bayes-Netze sind eine spezielle Form von Probabilistischen Graphischen Modellen (PGMs). PGMs nutzen die Struktur komplexer Verteilungen aus, um diese mithilfe von Graphen kompakt darzustellen (Koller and Friedman, 2009, S. 3). Wie Verteilungen in CLG Bayes-Netzen repräsentiert werden, wird in Abschnitt 3.1 beschrieben. Die graphische Darstellung kann darüber hinaus auch genutzt werden, um effizient Schlussfolgerungen unter Unsicherheit durchzuführen, was in Abschnitt 3.2 veranschaulicht wird. Der letzte Abschnitt (3.3) befasst sich mit der Erstellung von CLG Bayes-Netzen, wobei hier der Fokus auf dem Lernen von Parametern liegt.

## 3.1 Repräsentation

Da CLG Bayes-Netze Spezialfälle von Bayes-Netzen sind, werden diese im Folgenden kurz vorgestellt.
Grundlage für ein (diskretes) Bayes-Netz sind ein DAG $\mathcal{G}$ und eine Verteilung $P$, wobei die Knoten in $\mathcal{G}$ den diskreten Zufallsvariablen $\boldsymbol{X} = \{X_1, \ldots, X_n\}$ in $P$ entsprechen. Die entscheidende Eigenschaft von Bayes-Netzen ist dabei, dass alle durch d-Separation aus $\mathcal{G}$ ableitbaren Unabhängigkeiten als stochastische Unabhängigkeiten in $P$ enthalten sind:

$$\mathcal{I}(\mathcal{G})) \subseteq \mathcal{I}(P) \tag{3.1}$$

(Koller and Friedman, 2009, S. 62, 72). Diese Eigenschaft macht $\mathcal{G}$ zu einer Unabhängigkeitskarte (engl. *independency map*, I-Map) von $P$. Sind mehrere DAGs verfügbar, die die Eigenschaft in (3.1) erfüllen, so wird der DAG bevorzugt, der die größtmögliche Anzahl von in $P$ gültigen Unabhängigkeiten kodiert (engl. *minimal I-Map*) (Korb and Nicholson, 2010, S. 33).
Da $\mathcal{I}(\mathcal{G})$ und somit auch $\mathcal{I}_\ell(\mathcal{G})$ in $P$ gültig sind, kann die Verteilung als Produkt einzelner Terme dargestellt werden, was in diesem Zusammenhang als Faktorisierung von $P$ bezüglich $\mathcal{G}$ bezeichnet wird:

$$P(\boldsymbol{X}) = P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid Pa_{X_i}). \tag{3.2}$$

Gleichung (3.2) wird auch Kettenregel für Bayes-Netze genannt. Sie ermöglicht eine effiziente Darstellung der Verteilung $P$, da nur die durch $\mathcal{G}$ festgelegten bedingten Verteilungen (engl. *conditional probability distributions*, CPDs) $P(X_i \mid Pa_{X_i})$ für jede Zufallsvariable

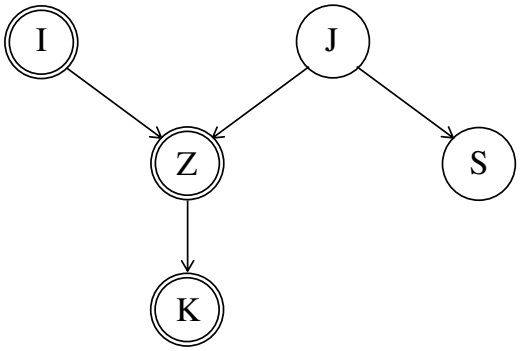



Abbildung 3.1: Beispiel für ein CLG Bayes-Netz mit zwei binären Variablen *J= Job*, *S= Stress* und drei stetigen Variablen *I= Interesse*, *Z= Zeit*, *K= Klausur*

$X_i$ bestimmt werden müssen. Besitzt eine Zufallsvariable in $\mathcal{G}$ keine Eltern, so ist die Menge der bedingenden Variablen leer und die CPD entspricht einer marginalen Verteilung (Koller and Friedman, 2009, S. 53, 62).
Ein Bayes-Netz kann somit formal definiert werden als ein Paar $\mathcal{B} = (\mathcal{G}, P)$, bestehend aus einem DAG $\mathcal{G}$, dessen Knoten die Zufallsvariablen $\boldsymbol{X} = \{X_1, \ldots, X_n\}$ darstellen und der eine Teilmenge der in einer durch CPDs bestimmten Verteilung $P$ gültigen stochastischen Unabhängigkeiten repräsentiert (Koller and Friedman, 2009, S. 62).
Je nachdem welche Variablen das Bayes-Netz beinhaltet, welche Einschränkungen $\mathcal{G}$ zugrunde liegen und welche Verteilung repräsentiert wird, ergeben sich verschiedene Bayes-Netze. Im Folgenden werden die Annahmen und Voraussetzungen von CLG Bayes-Netzen erläutert.

CLG Bayes-Netze kennzeichnen sich dadurch, dass sie sowohl diskrete Variablen $\boldsymbol{X^D}$ als auch stetige Variablen $\boldsymbol{X^C}$ enthalten, das heißt es gilt $\boldsymbol{X} = \boldsymbol{X^D} \cup \boldsymbol{X^C}$ (Madsen, 2008, S. 504). Zur Veranschaulichung wird im Folgenden ein aus Adachi (2016, S. 127 ff.) abgeleitetes Beispiel für ein CLG Bayes-Netz betrachtet, dessen DAG in Abbildung 3.1 dargestellt ist. Das CLG Bayes-Netz beinhaltet zwei binäre und drei stetige Variablen, wobei letztere graphisch durch eine doppelte Umrandung gekennzeichnet sind. In dem Beispiel wird angenommen, dass die Anzahl der Punkte, die ein Student in einer Klausur erreicht ($K$), von der in Stunden gemessenen Zeit ($Z$) abhängt, in der er sich während des Semesters mit dem Vorlesungsstoff beschäftigt hat. Diese wiederum wird dadurch beeinflusst, wie sehr sich der Student für das Thema der Vorlesung interessiert ($I$) und davon, ob er einen Job ($J$) ausübt. Ob der Student zusätzlich zum Studium arbeitet, wirkt sich zudem darauf aus, wie gestresst ($S$) er ist.
Der DAG in Abbildung 3.1 erfüllt die wichtige Voraussetzung in CLG Bayes-Netzen, dass diskrete Variablen keine stetigen Eltern haben dürfen. Für jedes $X^D \in \boldsymbol{X^D}$ gilt also

$$Pa_{X^D} \subseteq \boldsymbol{X^D} \tag{3.3}$$

(Kjaerulff and Madsen, 2006, S. 54). Daraus folgt, dass die bedingte Verteilung der einzelnen diskreten Variablen $X^D \in \boldsymbol{X^D}$ einer Multinomialverteilung entspricht, die für jede Zustandskombination der diskreten Eltern bestimmt wird (Koski and Noble, 2012, S. 56). Die bedingten Verteilungen werden meist in Tabellenform (*conditional probability tables*, CPTs) dargestellt (Koller and Friedman, 2009, S. 157).
Im Beispiel aus Abbildung 3.1 müssen folglich eine Binomialverteilung für die Variable $J$ und zwei Binomialverteilungen für die Variable $S$ bestimmt werden, da ihre Elternvariable $J$ zwei Ausprägungen hat.
Eine weitere Annahme betrifft die bedingte Verteilung der stetigen Variablen. Da im stetigen Fall die Wahrscheinlichkeit einzelner Ausprägungen gegen 0 geht, erscheint die Darstellung in Form von CPTs nicht besonders sinnvoll. Stattdessen wird im Fall von CLG Bayes-Netzen jeder stetigen Variable $X^C \in \boldsymbol{X^C}$ für jede Zustandskombination ihrer diskreten Eltern eine univariate Normalverteilung der Form

$$X^C \mid (\boldsymbol{D} = \boldsymbol{d}, \boldsymbol{C} = \boldsymbol{c}) \sim \mathcal{N}(\delta_{\boldsymbol{d},0} + \boldsymbol{\delta_d}^T \boldsymbol{c},\ \sigma^2_{\boldsymbol{d}}) \tag{3.4}$$

zugeordnet, wobei $\boldsymbol{d}$ und $\boldsymbol{c}$ die Realisierungen der diskreten und stetigen Eltern $\boldsymbol{D} \subseteq \boldsymbol{X^D}$ beziehungsweise $\boldsymbol{C} \subseteq \boldsymbol{X^C}$ darstellen. $\boldsymbol{\delta_d}$ ist ein von der Zustandskombination der diskreten Eltern abhängiger Vektor, der einen Regressionskoeffizienten für jede stetige Elternvariable beinhaltet. Der Mittelwert der Normalverteilung ist somit linear abhängig von $\boldsymbol{C}$. Die Varianz $\sigma^2_{\boldsymbol{d}}$ hingegen wird nur von der Ausprägung der diskreten Eltern $\boldsymbol{D}$ beeinflusst. Aufgrund dieser Eigenschaften wird die bedingte Verteilung in (3.4) auch als *conditional linear Gaussian distribution* (CLG Verteilung) bezeichnet (Kjaerulff and Madsen, 2006, S. 55). Alternativ ist in der Literatur auch der Begriff *CG regression* zu finden (Lauritzen and Jensen, 2001, S. 192).
Bezogen auf das Beispiel in Abbildung 3.1 müssen für die stetige Variable $Z$ aufgrund der binären Elternvariable $J$ zwei Normalverteilungen bestimmt werden, deren Mittelwerte linear abhängig von $I$ sind, das heißt:

$$\begin{aligned} Z \mid (J = 0, I = i) &\sim \mathcal{N}(\delta_{J=0,0} + (\delta_{J=0,1} \cdot i), \sigma^2_{J=0}), \\ Z \mid (J = 1, I = i) &\sim \mathcal{N}(\delta_{J=1,0} + (\delta_{J=1,1} \cdot i), \sigma^2_{J=1}). \end{aligned}$$

Für die ebenfalls stetige Variable $K$ ist dagegen nur eine einzelne Normalverteilung erforderlich, da sie nur eine stetige Elternvariable ($Z$) hat:

$$K \mid (Z = z) \sim \mathcal{N}(\delta_0 + (\delta_1 \cdot z), \sigma^2).$$

Für die stetige Variable $I$ muss ebenfalls nur eine einzelne Normalverteilung bestimmt werden. Diese ist jedoch nicht auf andere Variablen bedingt, da die Menge der Eltern von $I$ leer ist.

## 3.2 Inferenz

Die kompakte Darstellung der Verteilung $P$ im CLG Bayes-Netz kann genutzt werden, um Schussfolgerungen unter Unsicherheit bezüglich einer Menge von beobachteten Variablen $\epsilon$ durchzuführen (Kruse et al., 2011, S. 403). Dies soll im Folgenden veranschaulicht werden. Entspricht die Menge der beobachteten Variablen $\epsilon$ im CLG Bayes-Netz der leeren Menge, so werden die marginalen Verteilungen der Variablen als A-priori-Verteilungen bezeichnet. Ist jedoch der Zustand einer oder mehrerer Variablen bekannt ($\epsilon \neq \varnothing$), so kann diese Information die A-posteriori-Verteilungen der übrigen, nicht beobachteten Variablen beeinflussen. In diesem Fall gilt für $X \in \boldsymbol{X}$:

$$p(x \mid \epsilon = \varnothing) \neq p(x \mid \epsilon \neq \varnothing) \tag{3.5}$$

(Korb and Nicholson, 2010, S. 36). Die Berechnung der entsprechenden A-posteriori-Verteilungen unter Berücksichtigung der beobachteten Variablen wird als Inferenz oder *belief update* in CLG Bayes-Netzen bezeichnet (Madsen, 2008, S. 504).
Dabei kann die A-posteriori-Verteilung einer Variable prinzipiell von der Beobachtung jeder Variable im CLG Bayes-Netz beeinflusst werden (Korb and Nicholson, 2010, S. 33), solange zwischen den beiden Variablen ein Pfad existiert, der nicht bezüglich der übrigen Variablen in $\epsilon$ blockiert ist. Dies soll anhand des CLG Bayes-Netzes aus Abbildung 3.1 basierend auf Koller and Friedman (2009, S. 54 f.) verdeutlicht werden. Wird hier beispielsweise als einzige Variable das Interesse ($I$) des Studenten an der Vorlesung beobachtet, also $\epsilon = \{i\}$, so wirkt sich dies auf das erwartete Klausurergebnis ($K$) aus, das heißt:

$$p(k \mid \epsilon = \varnothing) \neq p(k \mid \epsilon = \{i\}).$$

Diese Form der Schlussfolgerung entlang der Kantenrichtungen wird *causal reasoning* genannt. Umgekehrt kann auch die Information über das Klausurergebnis die A-posteriori-Verteilung von $I$ beeinflussen, was als *evidential reasoning* bezeichnet wird. Wird jedoch gleichzeitig beobachtet, wie viele Stunden ($Z$) sich der Student mit dem Stoff der Vorlesung beschäftigt hat, so ist die zusätzliche Information über das Interesse unwesentlich für das Klausurergebnis und umgekehrt, da der Pfad von $I$ nach $K$ durch $Z$ blockiert ist. Anders verhält es sich mit der Abhängigkeit von Interesse und der Variable Job ($J$), die über eine konvergierende Verbindung miteinander verbunden sind ($I \rightarrow Z \leftarrow J$). Ist nur bekannt, ob der Student sich für die Vorlesung interessiert, so entspricht die A-posteriori-

Verteilung von $J$ der A-priori-Verteilung, das heißt:

$$p(j \mid \epsilon = \varnothing) = p(j \mid \epsilon = \{i\}).$$

Wird dagegen $Z$ beobachtet, so ergibt sich ein aktiver Pfad, wodurch beispielsweise eine trotz großem Interesse für die Vorlesung geringe Anzahl an Lernstunden durch einen vorhandenen Nebenjob erklärt wird. Diese Form der Schlussfolgerung wird auch als *intercausal reasoning* bezeichnet.
Es gilt zu beachten, dass die drei Grundformen der Schlussfolgerung (*causal, evidential* und *intercausal reasoning*) auch kombiniert werden können (Korb and Nicholson, 2010, S. 35).
Zur Berechnung der A-posteriori-Verteilungen sind verschiedene Algorithmen verfügbar, die alle die kompakte Darstellung der gemeinsamen Verteilung ausnutzen. Genau Ausführungen zu exakten Inferenzalgorithmen sind beispielsweise in Madsen (2008) und Lauritzen and Jensen (2001) zu finden.

## 3.3 Lernen

In den bisherigen Abschnitten wurden sowohl die Struktur von $\mathcal{G}$ als auch die Parameter der CPDs als gegeben betrachtet. Dies ist jedoch allgemein nicht der Fall, weshalb im Folgenden verschiedene Ansätze zur Konstruktion von CLG Bayes-Netzen beschrieben werden. Eine Möglichkeit stellt die manuelle Bestimmung von Graphenstruktur und Parametern der CPDs basierend auf theoretischen Überlegungen und Expertenwissen dar. Dies ist in vielen Fällen aber sehr zeitaufwendig und oftmals gar nicht möglich, beispielsweise wenn kein Experte auf dem entsprechenden Gebiet verfügbar ist (Koller and Friedman, 2009, S. 697).
Ein alternativer Ansatz ist das Lernen von CLG Bayes-Netzen aus einem Trainingsdatensatz $\mathcal{D}$. Dabei wird angenommen, dass alle Beobachtungseinheiten im Datensatz, $\xi[j], j = 1, \ldots, m$, unabhängig voneinander sind und von der wahren Verteilung $P_0$ generiert wurden, die vom CLG Bayes-Netz modelliert werden soll. Da die Parameter der CPDs von der Struktur des DAG abhängig sind, wird diese zuerst bestimmt (Kjaerulff and Madsen, 2006, S. 117, 188). Algorithmen zum Strukturlernen in allgemeinen Bayes-Netzen sind beispielsweise in Koller and Friedman (2009) zu finden. Beim Strukturlernen in CLG Bayes-Netzen muss dabei die Einschränkung berücksichtigt werden, dass diskrete Variablen keine stetigen Eltern haben dürfen.
Nach Festlegung von $\mathcal{G}$ können die Parameter $\boldsymbol{\theta}$ geschätzt werden. Ist der Trainingsdatensatz vollständig, kann dazu die Maximum-Likelihood-Methode verwendet werden. Da für jede Variable $X_i, i = 1, \ldots, n$, eine nur von den Eltern $Pa_{X_i}$ abhängige CPD bestimmt wird, kann die Likelihood-Funkion $L$ als Produkt von $n$ lokalen Likelihood-Funktionen

$L_{X_i}$ dargestellt werden, die getrennt maximiert werden:

$$\begin{aligned} L(\boldsymbol{\theta} : \mathcal{D}) &= \prod_{i=1}^{n} L_{X_i}(\boldsymbol{\theta}_{X_i|Pa_{X_i}} : \mathcal{D}) \\ &= \prod_{i=1}^{n} \prod_{j=1}^{m} p(x_i[j] \mid pa_{X_i}[j] : \boldsymbol{\theta}_{X_i|Pa_{X_i}}). \end{aligned} \tag{3.6}$$

Dabei repräsentieren $x_i[j]$ und $pa_{X_i}[j]$ die Werte von Beobachtung $j$ für $X_i$ beziehungsweise $Pa_{X_i}$ und $\boldsymbol{\theta}_{X_i|Pa_{X_i}}$ die Teilmenge der Parameter, die die CPD von $X_i$ bestimmen (Koller and Friedman, 2009, S. 724 f.).

Bei der Schätzung der Parameter kann zusätzlich ausgenutzt werden, dass allen diskreten und stetigen Variablen $\boldsymbol{X^D}$ beziehungsweise $\boldsymbol{X^C}$ für jede Zustandskombination der diskreten Eltern jeweils eine Multinomialverteilung beziehungsweise Normalverteilung zugeordnet wird (Koller and Friedman, 2009, S. 726). Für jede stetige Variable $X^C \in \boldsymbol{X^C}$ können somit die Parameter $\boldsymbol{\theta}_{X^C|\boldsymbol{d},\boldsymbol{C}} = (\delta_{\boldsymbol{d},0}, \boldsymbol{\delta}_{\boldsymbol{d}}^T, \sigma_{\boldsymbol{d}}^2)$ aus den Daten geschätzt werden, indem für die entsprechende Zustandskombination $\boldsymbol{d} \in Val(\boldsymbol{D})$ der diskreten Eltern eine lineare Regression mit den stetigen Eltern $\boldsymbol{C}$ als Prädiktoren durchgeführt wird (vgl. Koller and Friedman (2009, S. 728 f.) für *Gaussian Bayesian networks*).

Zur Schätzung der Parameter für diskrete Variablen siehe Koller and Friedman (2009, S. 725 f.).

# 4 Pfadmodelle

Pfadmodelle sind Strukturgleichungsmodelle, die ausschließlich beobachtbare Variablen enthalten. Sie können allerdings auch als erweiterte Regressionsmodelle betrachtet werden, in denen Variablen Einfluss- und Zielgröße zugleich sein können (Weiber and Mühlhaus, 2014, S. 26, 36).
Pfadmodelle werden in der Pfadanalyse verwendet, die ein multivariates statistisches Verfahren zur Untersuchung der Beziehungen zwischen Variablen darstellt. Ausgangspunkt ist die Formulierung von Zusammenhängen, die auf theoretischen Überlegungen basieren und kausal sind oder zumindest als kausal angenommen werden. Aus diesem Grund wurde die Pfadanalyse auch ursprünglich als *causal modeling* bezeichnet (Kline, 2011, S. 16). Das resultierende Modell wird durch ein Pfaddiagramm und eine Menge von Strukturgleichungen dargestellt (Abschnitt 4.1). Aus der Kovarianzmatrix der empirischen Daten kann anschließend die Stärke der Zusammenhänge geschätzt werden (Abschnitt 4.2). Die Aufteilung in verschiedene Effekte ermöglicht dabei eine genaue Analyse der oft komplexen Beziehungen zwischen den Variablen (Abschnitt 4.3). Wie gut ein Pfadmodell die Daten repräsentiert, kann anhand verschiedener Anpassungskriterien beurteilt werden (Abschnitt 4.4). Grundsätzlich kann jedoch nicht geprüft werden, ob die im Modell vermuteten Abhängigkeiten tatsächlich kausal sind (Kline, 2011, S. 8).

## 4.1 Aufbau

Pfadmodelle können sowohl durch Pfaddiagramme (Abschnitt 4.1.1) als auch durch Strukturgleichungen (Abschnitt 4.1.2) dargestellt werden, die im Folgenden jeweils näher beschrieben werden. In Abschnitt 4.1.3 werden anschließend die in Pfadmodellen getroffenen Annahmen erläutert.

### 4.1.1 Pfaddiagramme

Die graphische Repräsentation der zwischen den Variablen vermuteten Abhängigkeitsstrukturen erfolgt durch ein Pfaddiagramm. Genau wie in CLG Bayes-Netzen können zwei Variablen auch in Pfadmodellen durch einen Pfad verbunden sein. Anders als im DAG eines CLG Bayes-Netzes entspricht jedoch ein Pfad in einem Pfaddiagramm einem geraden, einseitig gerichteten Pfeil ($\rightarrow$). Existiert zwischen zwei Variablen ein solcher Pfad, so wird dies im Folgenden als kausaler Zusammenhang bezeichnet, wobei der Pfad von Ursache zu Wirkung führt (Adachi, 2016, S. 127). Es gilt allerdings zu beachten, dass die Kausalität

in den meisten Fällen nur angenommen werden kann (Kline, 2011, S. 16). Wird zwischen zwei Variablen hingegen kein kausaler Zusammenhang, sondern lediglich eine Korrelation angenommen, so wird dies durch einen gebogenen, zweiseitig gerichteten Pfeil gekennzeichnet ($\leftrightsquigarrow$) (Adachi, 2016, S. 127).

Anhand der im Pfaddiagramm vorliegenden Verbindungen kann in Pfadmodellen zwischen drei Arten von Variablen unterschieden werden. Zu den exogenen Variablen zählen alle Variablen, die von keinen Faktoren innerhalb des Modells kausal beeinflusst werden (Mueller, 1996, S. 23). Diese werden als $Z_j, j = 1, \ldots, m$, notiert und entsprechen im Pfaddiagramm allen Variablen, zu denen kein Pfad führt (Adachi, 2016, S. 131). Von den endogenen Variablen $Y_i, i = 1, \ldots, n$, wird hingegen angenommen, dass sie von mindestens einer Variable im Modell kausal beeinflusst werden, weshalb im Pfaddiagramm mindestens ein Pfad zu jeder endogenen Variable führt. Im Gegensatz zu klassischen Regressionsmodellen können endogene Variablen in Pfadmodellen jedoch zugleich auch auf andere (endogene) Variablen wirken (Mueller, 1996, S. 23). Alle endogenen und exogenen Variablen werden in Pfadmodellen als beobachtbar vorausgesetzt, was einen entscheidenden Unterschied zu allgemeinen Strukturgleichungsmodellen darstellt. Die einzige Ausnahme bilden die Fehlerterme $\zeta_i, i = 1, \ldots, n$, die für jedes $Y_i$ modelliert werden und unerklärte, durch unberücksichtigte Ursachen entstehende Varianz repräsentieren. Da Art und Anzahl dieser Ursachen unbekannt sind, werden die Fehlerterme als latente (exogene) Variablen betrachtet. Um dies graphisch zu verdeutlichen, werden die latenten Fehlerterme im Pfaddiagramm als Kreise gekennzeichnet, während die übrigen beobachtbaren Variablen als Rechtecke dargestellt werden (Kline, 2011, S. 103 f.).

Unter Verwendung der eingeführten Begriffe kann nun zwischen rekursiven und nicht-rekursiven Pfadmodellen unterschieden werden. Im Gegensatz zu den nicht-rekursiven Modellen ergeben sich in rekursiven Modellen durch die Pfade weder direkte Wechselwirkungen ($Y_1 \rightleftarrows Y_2$, *direct feedback loops*) noch indirekte Schleifen (engl. *indirect feedback loops*) wie beispielsweise $Y_1 \to Y_2 \to Y_3 \to Y_1$. Zusätzlich dürfen die Fehlerterme zweier endogener Variablen in rekursiven Modellen nicht korreliert sein (siehe Abschnitt 4.1.3). Nicht-rekursive Modelle sind somit flexibler, aber auch komplexer bezüglich Interpretation und Schätzung der Parameter, weshalb in dieser Arbeit nur rekursive Pfadmodelle behandelt werden (Kline, 2011, S. 106 ff.).

Als Beispiel wird in den folgenden Abschnitten ein rekursives Pfadmodell betrachtet, das auf Adachi (2016, S. 127 ff.) basiert und dessen Pfaddiagramm in Abbildung 4.1 dargestellt ist. Analog zu dem CLG Bayes-Netz aus Abbildung 3.1 enthält es die stetigen Variablen *Interesse* ($Z_1$), *Zeit* ($Y_1$) und *Klausur* ($Y_3$), die das Interesse an dem Vorlesungsstoff, die Lernzeit und die in der Klausur erreichte Punktzahl beschreiben. Zusätzlich werden die Variablen *Wissen* ($Z_2$) und *Abwesenheit* ($Y_2$) modelliert, die bereits vorhandenes Wissen über den Vorlesungsstoff sowie die Anzahl der Abwesenheitsstunden repräsentieren und ebenfalls stetig sind. Das Pfadmodell beinhaltet somit zwei exogene Va-

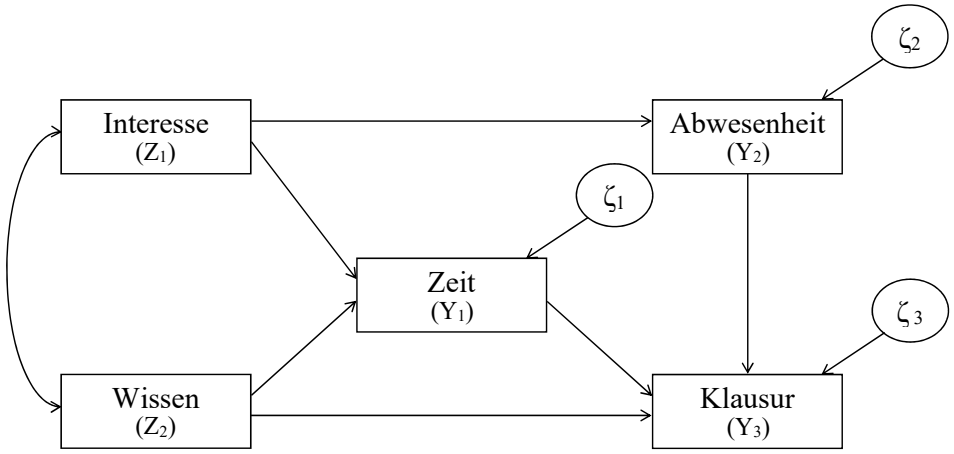



Abbildung 4.1: Beispiel für ein Pfaddiagramm mit zwei exogenen Variablen $\boldsymbol{Z} = (Z_1, Z_2)^T$ und drei endogenen Variablen $\boldsymbol{Y} = (Y_1, Y_2, Y_3)^T$ mit den Fehlertermen $\boldsymbol{\zeta} = (\zeta_1, \zeta_2, \zeta_3)^T$ nach Adachi (2016)

riablen $\boldsymbol{Z} = (Z_1, Z_2)^T$ und drei endogene Variablen $\boldsymbol{Y} = (Y_1, Y_2, Y_3)^T$, für die jeweils ein Fehlerterm modelliert wird ($\boldsymbol{\zeta} = (\zeta_1, \zeta_2, \zeta_3)^T$).

### 4.1.2 Strukturgleichungen

In Pfadmodellen wird anhand des Pfaddiagramms für jede endogene Variable eine univariate Regressionsgleichung bestimmt, die alle direkten kausalen Einflüsse enthält und in diesem Zusammenhang auch Strukturgleichung genannt wird. Neben einem konstanten Term $\alpha$ und einem Fehlerterm $\zeta$ beinhaltet jede Strukturgleichung Pfadkoeffizienten, die abhängig von der jeweiligen Einflussgröße als $\beta$ oder $\gamma$ notiert werden. Konkret entspricht in der Regressionsgleichung mit $Y_i$ als Zielvariable $\beta_{ik}$ dem Pfadkoeffizienten der endogenen Einflussvariable $Y_k$ und $\gamma_{ij}$ dem Pfadkoeffizienten der exogenen Einflussvariable $Z_j$. Da die Daten vor Schätzung der Parameter oftmals zentriert werden, wird dies im Folgenden ebenfalls angenommen. Somit können alle Gleichungen um den konstanten Term $\alpha$ reduziert werden und die Darstellung der Strukturgleichungen in Matrixform erfolgt durch

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} \beta_{11} & \dots & \beta_{1n} \\ \vdots & & \vdots \\ \beta_{n1} & \dots & \beta_{nn} \end{pmatrix} \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} + \begin{pmatrix} \gamma_{11} & \dots & \gamma_{1m} \\ \vdots & & \vdots \\ \gamma_{n1} & \dots & \gamma_{nm} \end{pmatrix} \begin{pmatrix} Z_1 \\ \vdots \\ Z_m \end{pmatrix} + \begin{pmatrix} \zeta_1 \\ \vdots \\ \zeta_n \end{pmatrix} \tag{4.1}$$

oder kurz

$$\boldsymbol{Y} = \boldsymbol{B}\boldsymbol{Y} + \boldsymbol{\Gamma}\boldsymbol{Z} + \boldsymbol{\zeta}. \tag{4.2}$$

Um ein Pfadmodell vollständig zu definieren, müssen neben den Pfadkoeffizienten ($\boldsymbol{B}$ und $\boldsymbol{\Gamma}$) zusätzlich die Kovarianzmatrizen der Fehlerterme ($\boldsymbol{\Psi}$) und der exogenen Variablen ($\boldsymbol{\Phi}$)

bestimmt werden. Diese haben die Form

$$\boldsymbol{\Psi} = \begin{pmatrix} \sigma^2_{\zeta_1} & & 0 \\ & \ddots & \\ 0 & & \sigma^2_{\zeta_n} \end{pmatrix} \tag{4.3}$$

und

$$\boldsymbol{\Phi} = \begin{pmatrix} \sigma^2_{Z_1} & \sigma_{Z_{12}} & \dots & \sigma_{Z_{1m}} \\ \sigma_{Z_{21}} & \sigma_{Z_{22}} & \dots & \sigma_{Z_{2m}} \\ \vdots & \dots & \ddots & \vdots \\ \sigma_{Z_{m1}} & \dots & \dots & \sigma^2_{Z_{mm}} \end{pmatrix} \tag{4.4}$$

(Mueller, 1996, S. 24 f.).
In dem Beispiel aus Abbildung 4.1 sind drei endogene Variablen enthalten (*Zeit, Abwesenheit* und *Klausur*), dementsprechend ergeben sich drei Strukturgleichungen:

$$\begin{aligned} Y_1 &= \gamma_{11} \cdot Z_1 + \gamma_{12} \cdot Z_2 + \zeta_1, \\ Y_2 &= \gamma_{21} \cdot Z_1 + \zeta_2, \\ Y_3 &= \beta_{31} \cdot Y_1 + \beta_{32} \cdot Y_2 + \gamma_{32} \cdot Z_2 + \zeta_3. \end{aligned}$$

Die vier Matrizen des Pfadmodells haben also die Form $\boldsymbol{B} = \left(\begin{smallmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ \beta_{31} & \beta_{32} & 0 \end{smallmatrix}\right)$, $\boldsymbol{\Gamma} = \left(\begin{smallmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & 0 \\ 0 & \gamma_{32} \end{smallmatrix}\right)$, $\boldsymbol{\Psi} = \begin{pmatrix} \sigma^2_{\zeta_1} & 0 & 0 \\ 0 & \sigma^2_{\zeta_2} & 0 \\ 0 & 0 & \sigma^2_{\zeta_3} \end{pmatrix}$ und $\boldsymbol{\Phi} = \begin{pmatrix} \sigma^2_{Z_1} & \sigma_{Z_{12}} \\ \sigma_{Z_{21}} & \sigma^2_{Z_2} \end{pmatrix}$.

### 4.1.3 Annahmen

Um Probleme bei der Schätzung und Interpretation der Parameter zu vermeiden, werden in Pfadmodellen einige Annahmen getroffen, die zum Teil bereits implizit in den Gleichungen (4.1) bis (4.4) enthalten sind. Gleichung (4.1) setzt beispielsweise voraus, dass die kausalen Zusammenhänge im Pfadmodell linear sind. Weiter entspricht die Kovarianzmatrix der Fehlerterme in (4.3) einer Diagonalmatrix, dementsprechend werden die Fehlerterme als untereinander unkorreliert betrachtet. Zusätzlich geht man davon aus, dass zwischen exogenen Variablen und Fehlertermen keine Korrelation vorliegt, weshalb die Kovarianzmatrizen $\boldsymbol{\Psi}$ und $\boldsymbol{\Phi}$ getrennt formuliert werden können (Mueller, 1996, S. 25 f.).
Außerdem wird angenommen, dass alle Fehlerterme über die Beobachtungen unkorreliert sind und einer Normalverteilung mit Erwartungswert 0 und Varianz $\sigma^2_{\zeta_i}$ folgen, das heißt:

$$\zeta_i \sim \mathcal{N}(0, \sigma^2_{\zeta_i}) \tag{4.5}$$

(Mueller, 1996, S. 26).
Eine weitere wesentliche Annahme in Pfadmodellen ist, dass alle beobachtbaren Variablen $\boldsymbol{Z} = (Z_1, \ldots, Z_m)^T$ und $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ mit keinem oder vernachlässigbarem Fehler gemessen werden (Mueller, 1996, S. 25).
Insgesamt gilt zu beachten, dass die in rekursiven Pfadmodellen gemachten Annahmen verhältnismäßig stark sind und in allgemeineren Strukturgleichungsmodellen abgeschwächt werden können, gleichzeitig vereinfachen sie jedoch auch die Schätzung und Interpretation der Parameter (Kline, 2011, S. 108).

## 4.2 Schätzung

Sind die im vorherigen Abschnitt genannten Annahmen erfüllt, so können die freien Parameter $\boldsymbol{\Theta}$ in $\boldsymbol{B}, \boldsymbol{\Gamma}, \boldsymbol{\Psi}$ und $\boldsymbol{\Phi}$ aus den vorliegenden Daten, das heißt aus der empirischen Kovarianzmatrix der beobachtbaren Variablen $\boldsymbol{S}$, geschätzt werden. Ziel der Schätzung ist es, $\hat{\boldsymbol{\Theta}}$ so zu wählen, dass die wahre Kovarianzmatrix der beobachtbaren Variablen $\boldsymbol{\Sigma}$ möglichst gut durch die vom Modell implizierte Kovarianzmatrix $\boldsymbol{\Sigma}(\boldsymbol{\Theta})$ approximiert wird. Um dies zu erreichen, wird eine geeignete *fit function* $F(\boldsymbol{S}, \boldsymbol{\Sigma}(\boldsymbol{\Theta}))$ minimiert, die die Abweichung von $\boldsymbol{\Sigma}(\boldsymbol{\Theta})$ zu $\boldsymbol{S}$ misst. Meist erfolgt die Schätzung durch die Maximum-Likelihood-Methode (ML-Methode). In diesem Fall entspricht der Vektor $\hat{\boldsymbol{\Theta}}$, der die *fit function* $F_{ML}$ minimiert, dem ML-Schätzer (Mueller, 1996, S. 151-155). Alle geschätzten Parameter werden in das entsprechende Pfaddiagramm eingetragen, was in Abbildung 4.2 für das Pfaddiagramm aus Abbildung 4.1 veranschaulicht ist.
Aufgrund der restriktiven Annahmen in rekursiven Pfadmodellen können die Pfadkoeffizienten und Varianzen der Fehlerterme alternativ auch geschätzt werden, indem für jede endogene Variable eine lineare Regression berechnet wird (Mueller, 1996, S. 26). Die Kovarianzmatrix der exogenen Variablen ($\boldsymbol{\Phi}$) muss in diesem Fall separat geschätzt werden und entspricht der empirischen Kovarianzmatrix (Kline, 2011, S. 162).
Oftmals werden zusätzlich die standardisierten Parameter des Pfadmodells berechnet. Die Standardisierung der Pfadkoeffizienten erfolgt analog zur Standardisierung von Regressionskoeffizienten. Die standardisierte Kovarianzmatrix $\boldsymbol{\Phi}$ entspricht einer Korrelationsmatrix, wodurch alle exogenen Variablen eine Varianz von 1.0 besitzen. Theoretisch gilt dies ebenfalls für die Fehlerterme. Hier wird jedoch stattdessen für jedes $\zeta_i$ der Anteil der unerklärten Varianz von $Y_i$ angegeben (Kline, 2011, S. 160-163). Dies ist äquivalent zu $1 - R^2_{Y_i}$, wobei $R^2_{Y_i}$ dem Bestimmtheitsmaß der Regression mit $Y_i$ als Zielvariable entspricht (Mueller, 1996, S. 32).
Die maximale Anzahl an freien Parametern, die in rekursiven Pfadmodellen geschätzt werden kann, entspricht der Anzahl der nicht redundanten Varianzen und Kovarianzen in $\boldsymbol{S}$. In diesem Fall kann die empirische Kovarianzmatrix $\boldsymbol{S}$ perfekt reproduziert werden und das Pfadmodell wird als genau identifiziert bezeichnet (Kline, 2011, S. 124 f.). In vielen Fällen

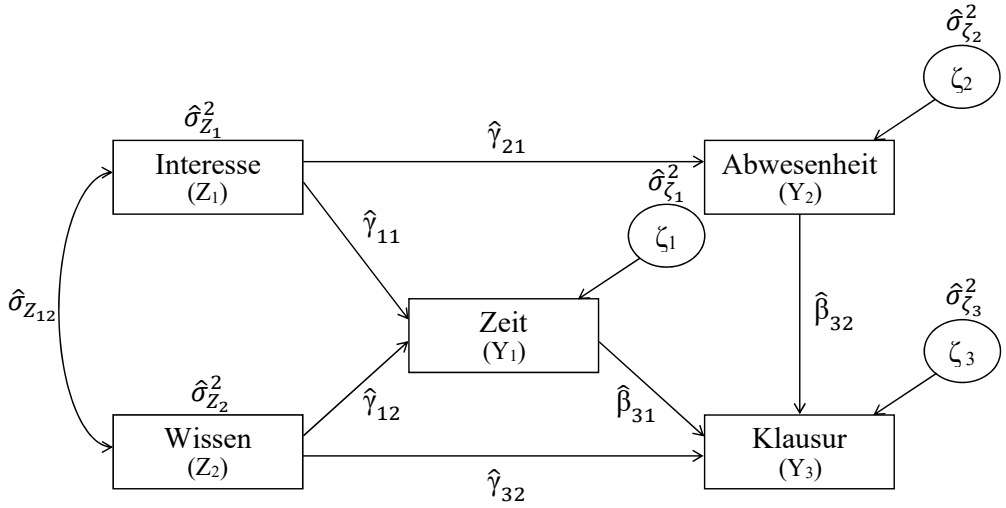



Abbildung 4.2: Pfaddiagramm aus Abbildung 4.1 mit undstandardisierten geschätzten Parametern

wird jedoch der Pfadkoeffizient mancher Pfade auf 0 festgelegt. Dies entspricht der Annahme, dass zwischen den jeweiligen Variablen kein direkter kausaler Zusammenhang besteht (Mulaik, 2009, S. 119-121). Ein solches eingeschränktes Modell gilt als überidentifiziert (Mueller, 1996, S. 47).

## 4.3 Effektzerlegung

In Pfadmodellen kann der oft komplexe Zusammenhang zwischen zwei Variablen ausführlich analysiert werden, indem die vom Modell implizierte Kovarianz in verschiedene Effekte zerlegt wird. Grundsätzlich wird dabei zwischen kausalen und nicht-kausalen Effekten unterschieden, die im Folgenden jeweils näher erläutert werden.

### Kausale Effekte

Kausale Effekte werden nochmals in direkte, indirekte und totale Effekte unterteilt und können nur entlang der Pfadrichtungen bestimmt werden.
Existiert zwischen zwei Variablen ein Pfad, so spricht man von einem direkten Effekt ($DE$), der durch den jeweiligen Pfadkoeffizienten bestimmt wird. Werden die Variablen durch mehrere, in dieselbe Richtung weisende Pfade verbunden, so ergibt sich ein indirekter Effekt. Dieser wird durch das Produkt aller in der Verbindung enthaltenen Pfadkoeffizienten berechnet. Der totale indirekte Effekt zwischen zwei Variablen ($IE$) wird bestimmt, indem alle indirekten Effekte addiert werden. Der totale kausale Effekt ($TE$) setzt sich schließlich aus dem direkten und dem totalen indirekten Effekt zusammen:

$$TE = DE + IE. \tag{4.6}$$

Ist zwischen zwei Variablen keine indirekte Verbindung vorhanden, so besteht der totale Effekt lediglich aus dem direkten Effekt. Gleiches gilt, wenn nur ein indirekter Effekt vorliegt (Mueller, 1996, S. 36).
Abhängig davon, ob sie mit standardisierten oder unstandardisierten Pfadkoeffizienten berechnet wurden, können alle kausalen Effekte wie (un-)standardisierte Regressionskoeffizienten in der klassischen linearen Regression interpretiert werden (Kline, 2011, S. 162-167). Im Beispiel aus Abbildung 4.2 setzt sich der unstandardisierte totale kausale Effekt zwischen *Wissen* ($Z_2$) und *Klausur* ($Y_3$) aus einem direkten Effekt und einem indirekten Effekt über die Variable *Zeit* ($Y_1$) zusammen. Für den totalen Effekt ergibt sich also:

$$TE = \hat{\gamma}_{32} + (\hat{\gamma}_{12} \cdot \hat{\beta}_{31}).$$

#### Nicht-kausale Effekte

Haben zwei Variablen dieselbe Ursache oder enthält die Verbindung eine Korrelation ($\backsim$), so wird dies als nicht-kausaler Effekt bezeichnet (Weiber and Mühlhaus, 2014, S. 32). Um alle nicht-kausalen Effekte zwischen zwei Variablen zu bestimmen, werden unter Verwendung der *tracing rule* zunächst alle zulässigen Verbindungen (engl. *valid tracings*) identifiziert. Eine Verbindung ist demnach zulässig, wenn sie folgende Kriterien erfüllt:

1. Keine Variable wird durch eine Pfeilspitze erreicht und anschließend durch eine Pfeilspitze wieder verlassen.

2. Keine Variable kommt zweimal vor.

Alle zulässigen Verbindungen, die weder direkte noch indirekte kausale Verbindungen darstellen, sind nicht-kausale Verbindungen. In dem Pfaddiagramm aus Abbildung 4.1 ergeben sich beispielsweise zwischen *Wissen* und *Klausur* zwei nicht-kausale Verbindungen: $Wissen \backsim Interesse \rightarrow Zeit \rightarrow Klausur$ und $Wissen \backsim Interesse \rightarrow Abwesenheit \rightarrow Klausur$. Der entsprechende nicht-kausale Effekt wird als Produkt aller relevanten Pfadkoeffizienten und Korrelationen berechnet, wobei bei der vorgestellten *tracing rule* standardisierte Parameter gefordert werden. Im Gegensatz zu kausalen Effekten können nicht-kausale Effekte jedoch nicht wie Regressionskoeffizienten interpretiert werden, da sie entgegen der Pfadrichtungen verlaufen oder korrelative Verbindungen enthalten (Kline, 2011, S. 169 f.).

## 4.4 Anpassungsgüte

Um die Anpassung eines Pfadmodells an die empirischen Daten zu prüfen, sind verschiedene Anpassungsstatistiken verfügbar, die die empirische Kovarianzmatrix $\boldsymbol{S}$ mit der vom Modell implizierten Kovarianzmatrix $\boldsymbol{\Sigma}(\hat{\boldsymbol{\Theta}})$ vergleichen (Kline, 2011, S. 191 ff.). Dabei gilt

zu beachten, dass eine Bestimmung der Anpassungsgüte für genau identifizierte Pfadmodelle nicht sinnvoll ist, da in diesem Fall $\boldsymbol{S} = \boldsymbol{\Sigma}(\hat{\boldsymbol{\Theta}})$ gilt (Mueller, 1996, S. 176).
Ein Beispiel für eine Anpassungsstatistik ist der häufig angewandte Chi-Quadrat-Test, der sich aus der Stichprobengröße und dem Wert der minimierten *fit function* zusammensetzt (Mueller, 1996, S. 82).
Allgemein impliziert jedoch ein guter Anpassungswert nicht, dass das entsprechende Modell korrekt und theoretisch bedeutsam ist (Kline, 2011, S. 189-193). Vor allem gilt zu beachten, dass für die meisten Modelle sogenannte äquivalente Modelle existieren, das heißt Modelle mit gleichen Anpassungswerten aber unterschiedlicher Anordnung der Pfade und Korrelationen (Kline, 2011, S. 225). Daraus folgt auch, dass anhand der Anpassungsstatistiken nicht das Modell bestimmt werden kann, das tatsächlich kausale Zusammenhänge beinhaltet. Insgesamt kann die Pfadanalyse somit als *disconfirmatory technique* betrachtet werden, da Pfadmodelle abgelehnt, aber nie als korrekt bestätigt werden können (Kline, 2011, S. 16). Wird ein Modell aufgrund unzureichender Anpassungsgüte verworfen, so sollte auch die Spezifikation des neuen Modells auf theoretischen und nicht (nur) auf statistischen Überlegungen basieren (Kline, 2011, S. 94).

# 5 Vergleich der Modelle

Nachdem in den beiden vorhergehenden Kapiteln die grundlegenden Eigenschaften von CLG Bayes-Netzen und Pfadmodellen erläutert wurden, werden die beiden Modelle nun hinsichtlich ihrer Gemeinsamkeiten und Unterschiede in graphischer Darstellung, Kausalität, Variablen im Modell und Modellierung der Abhängigkeiten untersucht (Abschnitte 5.1 bis 5.4). Im letzten Abschnitt (5.5) erfolgt ein Vergleich zwischen Inferenz in CLG Bayes-Netzen und Effektzerlegung in Pfadmodellen. Manche der in diesem Kapitel genannten Unterschiede lassen sich dabei auf den wichtigsten Unterschied zwischen CLG Bayes-Netzen und Pfadmodellen zurückführen, der in den jeweiligen Anwendungszielen liegt: CLG Bayes-Netze werden verwendet, um eine Verteilung kompakt darzustellen und effizient Schlussfolgerungen unter Unsicherheit zu ziehen (Kruse et al., 2011, S. 353). In Pfadmodellen hingegen werden die (kausalen) Zusammenhänge zwischen Variablen in einem theoriebasierten Modell anhand von empirischen Daten geschätzt und mithilfe der graphischen Darstellung analysiert (Mueller, 1996, S. 57 f.).

## 5.1 Graphische Darstellung

Sowohl CLG Bayes-Netze als auch Pfadmodelle verwenden Graphen, um Abhängigkeitsstrukturen zwischen Variablen intuitiv und übersichtlich darzustellen (vgl. Abschnitt 3.1 und 4.1.1). Im Folgenden werden daher Unterschiede und Gemeinsamkeiten bezüglich der graphischen Darstellung erläutert.

Beide Modelle unterscheiden graphisch zwischen verschiedenen Variablenarten. Im DAG eines CLG Bayes-Netzes kennzeichnet man durch einfache oder doppelte Umrandungen, ob es sich um diskrete oder stetige Variablen handelt. Im Pfaddiagramm eines Pfadmodells hingegen wird durch runde und eckige Formen zwischen latenten und manifesten Variablen unterschieden.

Bezüglich der Verbindungen sind in den Graphen beider Modelle gerichtete Pfeile ($\rightarrow$) vorhanden, die direkte Abhängigkeiten repräsentieren. In rekursiven Pfadmodellen dürfen durch die Pfade dabei weder direkte noch indirekte *feedback loops* entstehen, wodurch das Pfaddiagramm genau wie der DAG im CLG Bayes-Netz gerichtet und azyklisch ist. Im Pfaddiagramm kann im Gegensatz zum DAG jedoch zusätzlich eine nicht weiter definierte Korrelation zwischen zwei Variablen modelliert werden, die durch einen gebogenen, zweiseitig gerichteten Pfeil gekennzeichnet ist ($\leftrightarrows$).

Mit dem d-Separations-Kriterium beziehungsweise der *tracing rule* (vgl. Abschnitt 2.3 und 4.3) verfügen beide Modelle über Methoden, mit denen die Abhängigkeiten zwischen den

Variablen mithilfe der graphischen Darstellung analysiert werden können. Die Kriterien für eine gültige Verbindung (*valid tracing*) in einem Pfadmodell sind dabei sehr ähnlich zu denen eines aktiven Pfades in einem CLG Bayes-Netz mit $\boldsymbol{Z} = \varnothing$, da in beiden Fällen keine konvergierende Verbindung ($\rightarrow X \leftarrow$) zulässig ist. Allerdings können in Pfadmodellen keine bedingten Abhängigkeiten bestimmt werden, da die *tracing rule* in Pfadmodellen nicht vorsieht, dass andere Variablen bereits beobachtet sind ($\boldsymbol{Z} \neq \varnothing$). Zudem werden hier im Gegensatz zu den CLG Bayes-Netzen auch korrelative Verbindungen ($\backsim$) berücksichtigt, die meist zwischen exogenen Variablen modelliert werden (Mulaik, 2009, S. 120). Besteht zwischen zwei exogenen Variablen im Pfadmodell eine korrelative Verbindung, so werden diese folglich als abhängig angenommen, während Variablen ohne Eltern in CLG Bayes-Netzen im Fall von $\boldsymbol{Z} = \varnothing$ grundsätzlich unabhängig sind.
Schließlich werden in Pfadmodellen üblicherweise die Pfadkoeffizienten sowie die Varianzen und Kovarianzen direkt in das Pfaddiagramm eingetragen, da so der Zusammenhang zwischen den Variablen noch übersichtlicher dargestellt werden kann. In CLG Bayes-Netzen dagegen werden die bedingten Verteilungen allenfalls bei Beispielen mit wenigen Variablen in den DAG aufgenommen, oftmals ist dies jedoch aufgrund der großen Anzahl an bedingten Verteilungen nicht möglich (siehe Abschnitt 5.4).

## 5.2 Kausalität

Sowohl Pfadmodelle als auch CLG Bayes-Netze sind zur Modellierung von kausalen Beziehungen geeignet (vgl. Abschnitt 4.1.1 sowie Korb and Nicholson 2010, S. 29). Daraus folgt jedoch nicht, dass ein gerichteter Pfeil in der graphischen Darstellung in jedem Fall einen kausalen Zusammenhang anzeigt.
Abgesehen von der Einschränkung, dass diskrete Variablen keine stetigen Eltern haben dürfen, ist in CLG Bayes-Netzen grundsätzlich jede Anordnung der Variablen und Kanten im DAG zulässig, die die Verteilung korrekt repräsentiert und somit gültige Schlussfolgerungen unter Unsicherheit liefert. Dennoch ist Kausalität in CLG Bayes-Netzen durchaus eine wünschenswerte Eigenschaft, da kausale Netze tendenziell weniger Kanten beinhalten und leichter verständlich sind. Zudem fällt die manuelle Konstruktion der Graphenstruktur nach kausalen Prinzipien meist leichter und Parameter können von Experten besser erfragt werden, wenn das CLG Bayes-Netz eine kausale Ordnung reflektiert (Koller and Friedman, 2009, S. 65 f., 1009).
In Pfadmodellen hingegen wird vorausgesetzt, dass alle durch Pfade dargestellten Verbindungen kausal sind oder zumindest als kausal angenommen werden (Kline, 2011, S. 16). Da eine gute Anpassung des Modells an die empirischen Daten keine kausalen Zusammenhänge impliziert, müssen Pfadmodelle basierend auf theoretischen Überlegungen und nicht (nur) auf statistischen Kriterien erstellt werden (vgl. Abschnitt 4.4). Daraus folgt auch, dass Pfadmodelle in ihrer Komplexität beschränkter sind als CLG Bayes-Netze, da

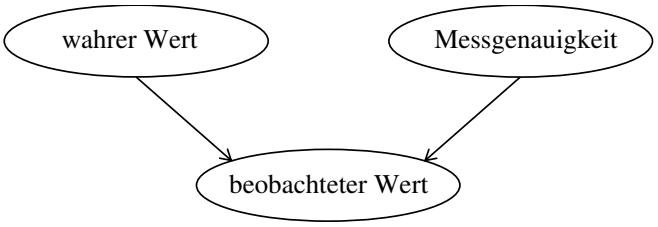



Abbildung 5.1: Explizite Modellierung von Messfehlern in CLG Bayes-Netzen

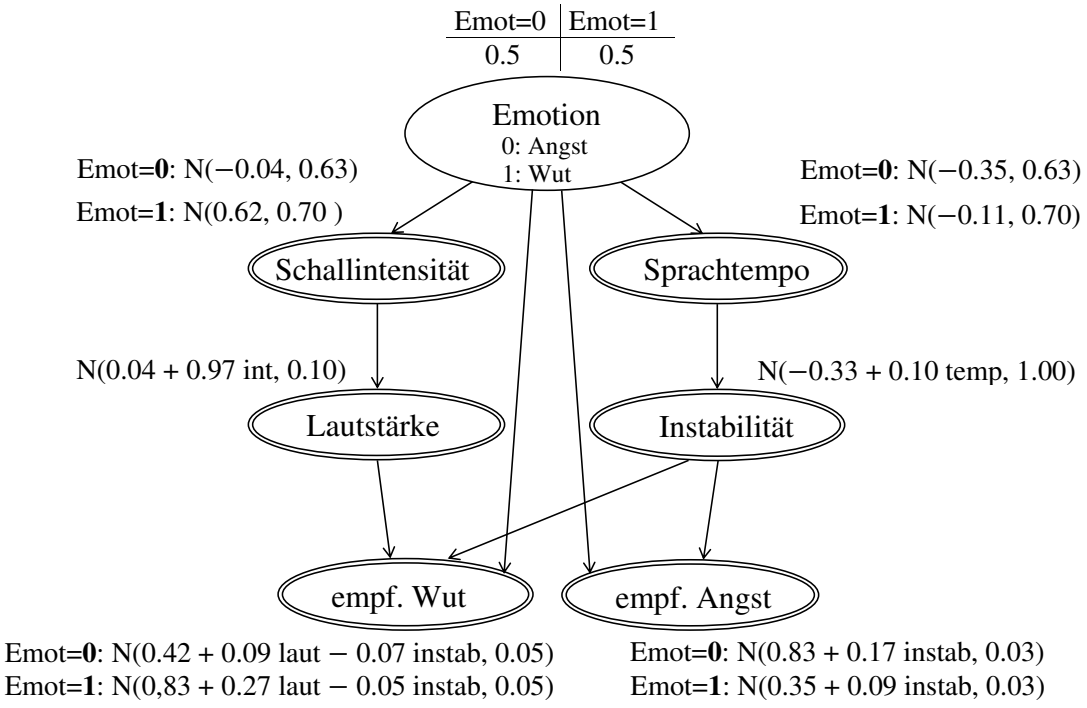
der mit der manuellen Modellkonstruktion einhergehende Zeitaufwand ab einer gewissen Anzahl von Variablen zu groß wird.

## 5.3 Variablen im Modell

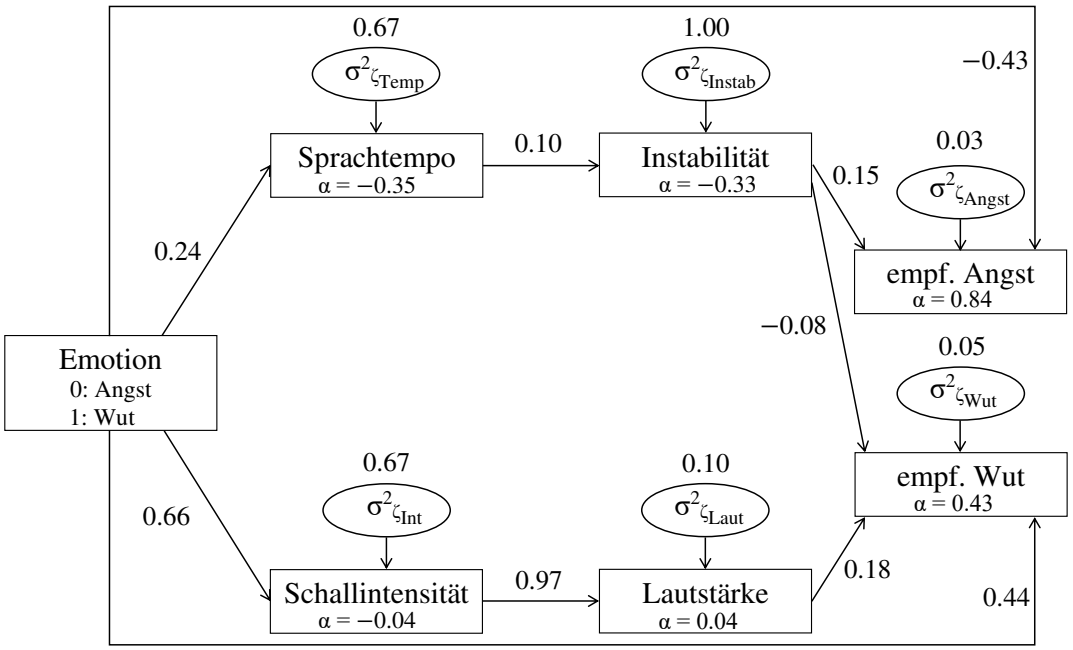
Hinsichtlich der im Modell zulässigen Variablen ergeben sich für CLG Bayes-Netze und Pfadmodelle mehrere Unterschiede, die in diesem Abschnitt beschrieben werden.
Zunächst können stetige und diskrete Variablen in CLG Bayes-Netzen beliebig angeordnet werden, solange dabei die Voraussetzung erfüllt wird, dass diskrete Variablen keine stetige Eltern haben (vgl. Abschnitt 3.1). In klassischen Pfadmodellen hingegen sind diskrete Variablen nur zulässig, wenn sie exogene Variablen darstellen (Rosseel, 2012, S. 25). Allerdings sind Ansätze verfügbar, in denen diskrete Variablen auch als endogene Variablen modelliert werden können, siehe dazu beispielsweise Kuha and Goldthorpe (2010).
Eine weitere Einschränkung von Pfadmodellen betrifft die Modellierung von latenten Variablen. Im Gegensatz zu den allgemeinen Strukturgleichungsmodellen können Pfadmodelle mit Ausnahme der Fehlerterme nur beobachtbare Variablen enthalten (Mueller, 1996, S. 129). In CLG Bayes-Netzen hingegen ist die Modellierung von latenten Variablen grundsätzlich nicht unzulässig. Auch wenn sich das Lernen von CLG Bayes-Netzen mit nicht beobachtbaren Variablen als komplex erweist, kann die Modellierung einer solchen Variable dennoch sinnvoll sein (Koller and Friedman, 2009, S. 713 f.). Zum Beispiel könnte in das CLG Bayes-Netz aus Abbildung 3.1 zusätzlich die latente Variable *Intelligenz* aufgenommen werden, da das Klausurergebnis üblicherweise nicht nur von der Lernzeit, sondern auch von der Intelligenz des jeweiligen Studenten abhängt.
Wie in Abschnitt 4.1.3 bereits erläutert wurde, werden die Variablen in Pfadmodellen nicht nur als beobachtbar, sondern auch als messfehlerfrei vorausgesetzt. Auch in dieser Hinsicht sind CLG Bayes-Netze flexibler, da hier Unsicherheit bezüglich der Messgenauigkeit explizit modelliert werden kann. Dazu werden drei Variablen erstellt, die den wahren Wert, die Messgenauigkeit und den beobachteten Wert repräsentieren (siehe Abbildung 5.1) (Kjaerulff and Madsen, 2006, S. 160 f.).

## 5.4 Modellierung der Abhängigkeiten

Ein weiterer wichtiger Aspekt im Vergleich von CLG Bayes-Netzen und Pfadmodellen ist die Modellierung der Abhängigkeiten zwischen den Variablen. In beiden Modellen wird für jede Variable eine Verteilung angenommen, die nur auf die Eltern (CLG Bayes-Netze) beziehungsweise die direkten Ursachen (Pfadmodelle) bedingt ist (vgl. Abschnitt 3.1 und 4.1.2). Da diskrete Variablen in Pfadmodellen keine abhängigen Variablen darstellen und somit auch keine bedingten Verteilungen zum Vergleich mit CLG Bayes-Netzen verfügbar sind, werden im Folgenden nur die Verteilungen von stetigen Variablen untersucht.
Hierfür wird zur Veranschaulichung ein Beispiel verwendet, das auf Bänziger et al. (2015) basiert. Dieses wurde so gewählt, dass der entsprechende DAG (Abbildung 5.2a) und das Pfaddiagramm (Abbildung 5.2b) die gleiche Struktur aufweisen und somit die Parameter direkt verglichen werden können. Das Beispiel behandelt die verbale Kommunikation von Emotionen. Zunächst wird dabei festgehalten, welche Emotion vom Sprecher ausgedrückt wird. Anschließend wird die Übertragung analysiert, wobei zwischen objektiven akustischen Messungen (engl. *distal cues*) und subjektiven Empfindungen des Zuhörers (engl. *proximal percepts*) unterschieden wird. Schließlich wird ermittelt, in wie vielen Fällen die kommunizierte Emotion vom Zuhörer richtig zugeordnet wird. Es gilt zu beachten, dass das vorliegende Beispiel nur eine Auswahl der von Bänziger et al. analysierten Variablen beinhaltet. Es werden die ausgesendeten Emotionen Angst und Wut untersucht, wodurch die Variable *Emotion* eine binäre Variable darstellt. Von dieser wird angenommen, dass sie sich direkt auf die objektiven stetigen Messungen *Schallintensität* (kurz *int*) und *Sprachtempo* (kurz *temp*) sowie die empfangenen Emotionen *(empf.) Angst* und *(empf.) Wut* auswirkt, die ebenfalls stetig sind. *Schallintensität* und *Sprachtempo* wiederum beeinflussen die subjektiven stetigen Empfindungen *Lautstärke* (kurz *laut*) und *Instabilität* (kurz *instab*). Schließlich wird für die *proximal percepts* ein Einfluss auf die empfangene Wut und Angst modelliert, wobei sich *Lautstärke* nur auf die Variable *(empf.) Wut* auswirkt.
Die entsprechenden (unstandardisierten) Parameter sind in Abbildung 5.2a beziehungsweise 5.2b zu finden und wurden in beiden Fällen mittels linearer Regression berechnet. Der R-Code zur Berechnung der Parameter befindet sich auf der dieser Arbeit beiliegenden CD. Es gilt zu beachten, dass die Daten in diesem Beispiel nicht zentriert sind, weshalb die Strukturgleichungen des Pfadmodells im Gegensatz zu der Annahme in Kapitel 4 einen konstanten Term $\alpha$ enthalten.
Sowohl in CLG Bayes-Netzen als auch in Pfadmodellen wird bei ausschließlich stetigen Einflussgrößen eine univariate Normalverteilung modelliert, deren Mittelwert linear von den bedingenden Variablen abhängt. Ist die Varianz größer als 0, so handelt es sich in beiden Fällen um eine nicht deterministische Abhängigkeit. Werden die Parameter mittels linearer Regression berechnet (vgl. Abschnitt 3.3 und 4.2), so ergibt sich für beide Modelle exakt dieselbe Verteilung. In dem vorliegenden Beispiel sind daher sowohl im CLG Bayes-




(a) CLG Bayes-Netz




(b) Pfadmodell

Abbildung 5.2: CLG Bayes-Netz und Pfadmodell zur verbalen Kommunikation von Emotionen basierend auf (Bänziger et al., 2015)

Netz als auch im Pfadmodell die Verteilungen von *Lautstärke* und *Instabilität* gegeben durch:

$$Laut \mid (Int = int) \sim \mathcal{N}(0.04 + 0.97 \cdot int, 0.10),$$
$$Instab \mid (Temp = temp) \sim \mathcal{N}(-0.33 + 0.10 \cdot temp, 1.00).$$

Enthält die Menge der bedingenden Variablen zusätzlich diskrete Variablen, so wird in CLG Bayes-Netzen für jede Zustandskombination der diskreten Eltern eine Normalverteilung bestimmt. Betrachtet man also beispielsweise die Variable *(empf.) Angst*, so wird für jede Ausprägung der binären Einflussgröße *Emotion* (0: Angst, 1:Wut) eine Normalverteilung modelliert:

$$Angst \mid (Emot = 0, Instab = instab) \sim \mathcal{N}(0.83 + 0.17 \cdot instab, 0.03),$$
$$Angst \mid (Emot = 1, Instab = instab) \sim \mathcal{N}(0.35 + 0.09 \cdot instab, 0.03).$$

Im Gegensatz dazu werden in Pfadmodellen diskrete Einflussvariablen dummy-kodiert und wie in der klassischen linearen Regression in den Mittelwert aufgenommen (Rosseel, 2012, S. 25). Im Beispiel wurde für die diskrete Variable *Emotion* die ausgesendete Emotion Angst (*Emot=0*) als Referenzkategorie gewählt, somit ergibt sich hier für die Variable *(empf.) Angst*:

$$Angst \mid (Emot = emot, Instab = instab) \sim \mathcal{N}(0.84 - 0.43 \cdot \mathbb{1}\{emot = 1\} + 0.15 \cdot instab, 0.03).$$

Die Modellierung in CLG Bayes-Netzen hat den Vorteil, dass Interaktionseffekte zwischen den diskreten und stetigen Einflussvariablen automatisch berechnet werden. So beträgt im Beispiel für die Zielgröße *(empf.) Angst* der Regressionskoeffizient von *Instabilität* abhängig von der ausgesendeten Emotion 0.17 beziehungsweise 0.09, während er im Pfadmodell unverändert bei 0.15 liegt. Daraus folgt allerdings, dass in CLG Bayes-Netzen zwangsläufig mehr Parameter als in Pfadmodellen bestimmt werden müssen. Dies impliziert auch, dass die Parameter in Pfadmodellen immer in das Pfaddiagramm eingetragen werden können, während dies für CLG Bayes-Netze in vielen Fällen nicht möglich ist.
Während die Parameter in Pfadmodellen immer aus empirischen Daten geschätzt werden müssen, ist dies für CLG Bayes-Netze nicht zwingend notwendig. Ist für eine Fragestellung die exakte Bestimmung der bedingten Verteilungen nicht entscheidend, so gibt es verschiedene Ansätze, um die Parameter mithilfe von Expertenwissen festzulegen. Dieses Vorgehen liefert allerdings oftmals keine reliablen Ergebnisse und ist in CLG Bayes-Netzen, die keine kausale Ordnung repräsentieren, nicht zu empfehlen (Koller and Friedman, 2009, S. 66). Anstatt der ML-Schätzer können in CLG Bayes-Netzen alternativ auch Bayes-Schätzer

berechnet werden. Dies ermöglicht die Kombination von Expertenwissen mit empirischen Daten, siehe dazu Koller and Friedman (2009, S. 733 ff.).

## 5.5 Inferenz versus Effektzerlegung

Wie im vorherigen Abschnitt beschrieben, werden sowohl in CLG Bayes-Netzen als auch in Pfadmodellen die Verteilungen der Variablen in Abhängigkeit ihrer Eltern beziehungsweise ihrer direkten Ursachen angegeben. Folglich kann deren Effekt auf die abhängige Variable direkt aus den Parametern der Verteilung abgelesen werden. In beiden Modellen können jedoch durch Inferenz (vgl. Abschnitt 3.2) und Effektzerlegung (vgl. Abschnitt 4.3) auch weitere Abhängigkeiten zwischen Variablen bestimmt werden. Im Folgenden werden die beide Methoden näher untersucht.

Bei der Inferenz in CLG Bayes-Netzen wird die graphische Darstellung der Verteilung verwendet, um die A-posteriori-Verteilungen von unbeobachteten Variablen im Modell hinsichtlich beliebig vieler beobachteter Variablen effizient zu berechnen. Wie in Abschnitt 3.2 bereits erwähnt wurde, kann dabei die Beobachtung jeder Variable im Netz berücksichtigt werden, unabhängig davon durch welche Kanten sie mit der interessierenden Variable verbunden ist (*causal, evidential* und *intercausal reasoning*). Die Stärke des Einflusses wird dabei nicht explizit berechnet.

Ziel in Pfadmodellen ist es hingegen, den Zusammenhang zwischen den Variablen im Modell zu analysieren. Dementsprechend wird neben dem direkten kausalen Effekt auch der indirekte und totale kausale Effekt zwischen zwei Variablen bestimmt. Zwar werden bei der vom Modell implizierten Kovarianz auch nicht-kausale Verbindungen berücksichtigt, die Stärke des Einflusses kann jedoch nur für kausale Effekte bestimmt werden. Dies entspricht in CLG Bayes-Netzen dem *causal reasoning*.

Für die Modelle zur verbalen Kommunikation von Emotionen (Abbildung 5.2) folgt daraus, dass im CLG Bayes-Netz beispielsweise die Verteilung von *Instabilität* hinsichtlich der Beobachtung von *Emotion* aktualisiert werden kann, jedoch kein Parameter angegeben wird, der die Stärke des Einflusses von *Emotion* auf *Instabilität* beschreibt. Im Pfadmodell steht dagegen nicht die A-posteriori-Verteilung von *Instabilität* im Vordergrund, sondern der indirekte und zugleich totale kausale Effekt von *Emotion* auf *Instabilität*, der sich durch $0.24 \cdot 0.10$ ergibt. Soll im Pfadmodell bestimmt werden, wie sich die Beobachtung von *Instabilität* auf *Emotion* auswirkt, müssten die Pfade umgekehrt werden. Da dies jedoch der Voraussetzung für Pfadmodelle widerspricht, dass die durch Pfade gekennzeichneten Beziehungen als kausal angenommen werden müssen, kann der Einfluss von *Instabilität* auf *Emotion* nicht bestimmt werden. Im CLG Bayes-Netz hingegen ist die Berechnung der A-posteriori-Verteilung von *Emotion* bezüglich der Beobachtung von *Instabilität* zulässig (*evidential reasoning*).

Es gilt zu beachten, dass die genannten Einschränkungen von Inferenz und Effektzer-

legung auf die unterschiedlichen Anwendungsziele von CLG Bayes-Netzen und Pfadmodellen zurückzuführen sind. Die Bestimmung von indirekten und totalen Effekten wäre auch in CLG Bayes-Netzen theoretisch möglich, hier steht jedoch die Berechnung der A-posteriori-Verteilungen von interessierenden Variablen im Vordergrund und nicht die Zusammenhänge zwischen den Variablen. Umgekehrt könnten in Pfadmodellen die Stärke der Einflüsse von nicht-kausalen Effekten sowie die A-posteriori-Verteilungen bezüglich mehrerer beobachteter Variablen berechnet werden, dies entspricht jedoch nicht dem Ziel, kausale Zusammenhänge zu analysieren.

# 6 Fazit

Ziel der Arbeit war es, CLG Bayes-Netze und Pfadmodelle hinsichtlich ihrer Gemeinsamkeiten und Unterschiede zu analysieren. Ein Vergleich schien interessant, da beide Modelle Zufallsvariablen und deren Abhängigkeitsstrukturen graphisch darstellen. Nachdem in Kapitel 3 und 4 Aufbau und Eigenschaften von CLG Bayes-Netzen und Pfadmodellen erläutert wurden, erfolgte in Kapitel 5 ein Vergleich der beiden Modelle hinsichtlich graphischer Darstellung, Kausalität, Variablen und Modellierung der Abhängigkeiten sowie Methoden zur Bestimmung weiterer Zusammenhänge.
In jedem dieser Punkte konnten dabei Gemeinsamkeiten festgestellt werden. Neben der Tatsache, dass sowohl CLG Bayes-Netze als auch Pfadmodelle Graphen verwenden, um die Abhängigkeitsstrukturen zwischen Zufallsvariablen darzustellen, ist in beiden Modellen die Modellierung von kausalen Zusammenhängen zumindest von Vorteil. Bezüglich der Variablen in CLG Bayes-Netzen und Pfadmodellen sind in beiden Fällen sowohl diskrete als auch stetige Variablen zulässig. Was die Modellierung der Abhängigkeiten betrifft, so werden für stetige Variablen in beiden Modellen Normalverteilungen angenommen, deren Parameter mittels linearer Regression geschätzt werden können. Wie anhand eines Beispiels veranschaulicht wurde, ergeben sich so für ausschließlich stetige Einflussvariablen dieselben Verteilungen. Auch wenn durch die Parameter nur der Einfluss der Eltern beziehungsweise der direkten Ursachen angegeben wird, sind durch Inferenz in CLG Bayes-Netzen und Effektzerlegung in Pfadmodellen Methoden zur Bestimmung weiterer Abhängigkeiten verfügbar.
Beim Vergleich von CLG Bayes-Netzen und Pfadmodellen ergaben sich allerdings auch einige Unterschiede. Der wohl wichtigste Unterschied liegt dabei in den Anwendungszielen der Modelle. Während in CLG Bayes-Netzen Schlussfolgerungen unter Unsicherheit gezogen werden, steht in Pfadmodellen die Analyse kausaler Zusammenhänge im Fokus. Die unterschiedlichen Anwendungsziele werden dabei besonders im Vergleich von Inferenz und Effektzerlegung deutlich. Aber auch die erforderliche Annahme von Kausalität in Pfadmodellen und die damit verbundenen Einschränkungen bei der Modellkonstruktion sowie die Eintragung der Parameter in das Pfaddiagramm können darauf zurückgeführt werden, dass in Pfadmodellen im Gegensatz zu CLG Bayes-Netzen die kausalen Zusammenhänge zwischen den Variablen im Vordergrund stehen. Die Unterschiede bezüglich der Variablen im Modell hingegen sind teilweise darauf zurückzuführen, dass (klassische) Pfadmodelle die älteste Form von Strukturgleichungsmodellen darstellen und dadurch in mancher Hinsicht eingeschränkter sind (Kline, 2011, S. 103). Wie bereits erwähnt wurde, sind jedoch

für die Modellierung von diskreten und latenten Variablen in Pfadmodellen Erweiterungen verfügbar.
Insgesamt ergab der Vergleich, dass sich CLG Bayes-Netze in einigen der untersuchten Aspekte als flexibler erweisen, dafür aber auch komplexer in Aufbau und Anwendung sind. Pfadmodelle sind diesbezüglich einfacher zu verstehen und finden daher trotz der Einschränkungen immer noch häufig Anwendung. Sowohl CLG Bayes-Netze als auch Pfadmodelle sind geeignet, um Zusammenhänge zwischen Variablen zu modellieren und diese übersichtlich darzustellen. Welches Modell dabei verwendet wird, sollte vor allem in Hinblick auf das Anwendungsziel entschieden werden.

Abschließend sollte berücksichtigt werden, dass in dieser Arbeit nur Spezialfälle von Bayes-Netzen und Strukturgleichungsmodellen untersucht wurden. Daher wäre zum Beispiel auch ein Vergleich von Modellen interessant, die sich bezüglich ihrer Variablen von CLG Bayes-Netzen und Pfadmodellen unterscheiden. Anstelle von CLG Bayes-Netzen könnte man hier beispielsweise Pfadmodelle mit sogenannten *Gaussian Bayesian networks* vergleichen, die ausschließlich stetige Variablen enthalten (Koller and Friedman, 2009, S. 251 ff.). Interessant wäre auch ein Vergleich von zwei Modellen, die hinsichtlich der Modellierung von diskreten Variablen flexibler sind als die in der Arbeit behandelten Modelle. Wie bereits erwähnt wurde, können Pfadmodelle diesbezüglich erweitert werden, siehe Kuha and Goldthorpe (2010). Bei den Bayes-Netzen könnten in diesem Fall die allgemeineren *Hybrid Bayesian networks* betrachtet werden, in denen diskrete und stetige Variablen beliebig angeordnet werden können (Koller and Friedman, 2009, S. 189 f.).
Dieser Ausblick zeigt, dass der Vergleich von allgemeinen Bayes-Netzen und Strukturgleichungsmodellen mit der vorliegenden Arbeit nicht abgeschlossen ist und diese Thematik daher in zukünftigen Arbeiten weiter untersucht werden sollte.

# Literaturverzeichnis


Adachi, K. (2016). *Matrix-based introduction to multivariate data analysis*, Springer, Singapur.

Bänziger, T., Hosoya, G. and Scherer, K. R. (2015). Path models of vocal emotion communication, *PloS one* **10**(9): 1–29.

Fahrmeir, L., Kneib, T. and Lang, S. (2009). *Regression: Modelle, Methoden und Anwendungen*, 2. edn, Springer, Berlin.

Kjaerulff, U. B. and Madsen, A. L. (2006). Probabilistic networks for practitioners – A guide to construction and analysis of Bayesian networks and influence diagrams.

Kline, R. B. (2011). *Principles and practice of structural equation modeling*, 3. edn, Guilford Press, New York.

Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: Principles and techniques*, MIT Press, Cambridge.

Korb, K. B. and Nicholson, A. E. (2010). *Bayesian Artificial Intelligence, Second Edition*, 2. edn, Chapman & Hall/CRC, Hoboken.

Koski, T. J. and Noble, J. (2012). A review of Bayesian networks and structure learning, *Mathematica Applicanda* **40**(1): 53–103.

Kruse, R. J., Borgelt, C., Klawonn, F., Moewes, C., Ruß, G. and Steinbrecher, M. (2011). *Computational Intelligence: Eine methodische Einführung in Künstliche Neuronale Netze, Evolutionäre Algorithmen, Fuzzy-Systeme und Bayes-Netze*, Springer Vieweg, Wiesbaden.

Kuha, J. and Goldthorpe, J. H. (2010). Path analysis for discrete variables: The role of education in social mobility, *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **173**(2): 351–369.

Lauritzen, S. L. and Jensen, F. (2001). Stable local computation with conditional Gaussian distributions, *Statistics and Computing* **11**(2): 191–203.

Madsen, A. L. (2008). Belief update in CLG Bayesian networks with lazy propagation, *International Journal of Approximate Reasoning* **49**(2): 503–521.

Mueller, R. O. (1996). *Basic Principles of Structural Equation Modeling: An Introduction to LISREL and EQS*, Springer, New York.

Mulaik, S. A. (2009). *Linear causal modeling with structural equations*, Chapman & Hall/-CRC, Boca Raton.

Pearl, J. (1982). Reverend Bayes on inference engines: A distributed hierarchical approach, AAAI-82 Proceedings, pp. 133–136.

Rosseel, Y. (2012). lavaan: an R package for structural equation modeling and more: Version 0.5-12 (BETA).

Weiber, R. and Mühlhaus, D. (2014). *Strukturgleichungsmodellierung: Eine anwendungsorientierte Einführung in die Kausalanalyse mit Hilfe von AMOS, SmartPLS und SPSS*, 2. edn, Springer Gabler, Berlin.

Wright, S. (1921). Correlation and causation, *Journal of agricultural research* **20**(7): 557–585.

# Abbildungsverzeichnis

2.1 Beispiel für einen gerichteten, azyklischen Graphen $\mathcal{G}$ mit $\mathcal{V} = \{\nu_1, \nu_2, \nu_3, \nu_4, \nu_5\}$ 10

3.1 Beispiel für ein CLG Bayes-Netz mit zwei binären Variablen *J= Job*, *S= Stress* und drei stetigen Variablen *I= Interesse*, *Z= Zeit*, *K= Klausur* . . . 13

4.1 Beispiel für ein Pfaddiagramm mit zwei exogenen Variablen $\boldsymbol{Z} = (Z_1, Z_2)^T$ und drei endogenen Variablen $\boldsymbol{Y} = (Y_1, Y_2, Y_3)^T$ mit den Fehlertermen $\boldsymbol{\zeta} = (\zeta_1, \zeta_2, \zeta_3)^T$ nach Adachi (2016) . . . . . 20

4.2 Pfaddiagramm aus Abbildung 4.1 mit undstandardisierten geschätzten Parametern . . . . . 23

5.1 Explizite Modellierung von Messfehlern in CLG Bayes-Netzen . . . . . 28

5.2 CLG Bayes-Netz und Pfadmodell zur verbalen Kommunikation von Emotionen basierend auf (Bänziger et al., 2015) . . . . . 30

# Elektronischer Anhang

# Erklärung zur Urheberschaft

Hiermit versichere ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

München, den 04. April 2017

(Christina Nießl)