

Technische Universität München

Fakultät für Mathematik



Bachelor-Arbeit

Homogenität von Fehlerkovarianzen

Julian Rittmaier

Aufgabensteller: Mathias Drton

Betreuer: Nils Sturma

Abgabetermin: 15.09.2022

Ich erkläre hiermit, dass ich die Bachelor-Arbeit selbständig und nur mit den angegebenen Hilfsmitteln angefertigt habe.

Julian R. Hwaider

Garching, den 14.09.2022

Summary

In statistics, *causal inference* describes the process of determining causal effects within a system. In order to accurately model observed variables, statisticians often derive estimators from graphical models such as the linear structural equation model (LSEM). These models are constructed under the assumption that it is possible to model unobserved (latent) variables by accounting for their effects through the usage of correlated error terms of the observed variables.

The main objective of this thesis is to investigate, given an observation, how the uniqueness of structural parameters in such a model is affected by external equality constraints. In this case, the constraints of interest are caused by equiconfounded random variables, i.e. multiple observed variables that are influenced by a common unobserved confounder. The main tool used to quantify such homogeneities are edge-colorings that are applied to the underlying causal graph. Using this terminology, one can establish results that are similar in nature to those found when studying uncolored graphs, like an upper bound for the number of edges in an identifiable graph. However, it quickly becomes clear that the usage of edge-colorings severely influences the identifiability of certain graphs. For instance, it is possible to greatly improve the number of identifiable graphs with a fixed bidirected structure by using a certain coloring, as discussed in section 3.2. Section 3.4 explores how colorings can also cause adverse effects by limiting identifiability for some graphs, which has significant effects on the applicability of graphical criteria for identifiability to colored graphs.

All computational results in this thesis have been produced using concepts from algebraic geometry, as described in sections 3.1 and 3.2. Specifically section 3.1 contains multiple results that are similar to well known methods from the analysis of uncolored LSEMs, which have been applied to the colored case.

Finally, every section is accompanied by multiple examples and analyses of special graph layouts, the most notable being several graphs in the appendix that highlight the undesired properties of graph colorings as discussed in section 3.4.

Inhaltsverzeichnis

1	Einleitung	1
2	Grundlegende Begriffe	1
2.1	Lineare Strukturgleichungsmodelle	1
2.2	Identifizierbarkeit von Modellen	3
2.3	Gleichheit von Fehlerkovarianzen	5
2.3.1	Motivation	5
2.3.2	Kantenfärbungen	7
3	Identifizierbarkeit von gefärbten Graphen	8
3.1	Algebraische Methoden	8
3.2	Computergestützte Analysen	12
3.2.1	Experimente	13
3.3	Verwendung von <i>partial-instrumental sets</i>	14
3.4	Durch Kantenfärbungen verursachte Nichtidentifizierbarkeit	18
4	Fazit	20
5	Anhang	21
5.1	Hilfsaussagen	21
5.2	Abbildungen	22
6	Literaturverzeichnis	23

1 Einleitung

Bei der Betrachtung komplexer Systeme ist es oft von Interesse, kausale Zusammenhänge zwischen Zufallsvariablen herzuleiten. *Lineare Strukturgleichungsmodelle* ermöglichen es, konkrete Aussagen über die Abhängigkeiten zwischen beobachteten und unbeobachteten Variablen zu machen. Dabei kann es vorkommen, dass sich diese Beziehungen aufgrund mangelnder Informationen über das System und dessen Störvariablen nicht eindeutig bestimmen lassen und die entsprechenden Parameter des Modells somit nicht eindeutig identifiziert werden können. In der Vergangenheit wurden bereits umfassende Untersuchungen angestellt, lineare Strukturgleichungsmodelle nach der Identifizierbarkeit ihrer Parameter graphisch zu charakterisieren, wodurch diverse Kriterien wie das *half-trek Kriterium* (Foygel et al., 2012a), das *single-door Kriterium* (Pearl, 2009), sowie einige Kriterien basierend auf dem Konzept instrumenteller Variablen (Brito and Pearl (2002), Weihs et al. (2018)) ermittelt werden konnten. Noch weitaus weniger erforscht ist die Frage, wie sich die Berücksichtigung externer Annahmen an ein System auf die Identifizierbarkeit seiner kausalen Zusammenhänge auswirkt. Resultate in diesem Bereich beziehen sich zum Teil nur auf spezielle Modelle (Chalakov, 2013) und lassen daher wenige Rückschlüsse über den generellen Fall zu. Ein Beispiel für einen allgemeineren Ansatz, welcher auf einer graphischen Methode zur Identifizierung beruht, wurde von Zhang et al. (2021) gegeben.

Ziel dieser Arbeit ist es zu untersuchen, wie sich bestimmte Annahmen über die Gleichheit einzelner Parameter eines solchen Modells, konkret der Wirkung latenter Variablen, auf die Identifizierbarkeit der Parameterwerte auswirken. Dazu soll zunächst der Begriff der *Kantenfärbung* definiert und graphisch motiviert werden. Anhand bestimmter Modelle wird dann erarbeitet, welche Vorteile die Einführung von Gleichheiten mit sich bringen und wie sich die Effekte von Kantenfärbungen experimentell mittels algebraischer Methoden auswerten lassen. Im Anschluss wird am Beispiel der Methode von Zhang et al. (2021) beschrieben, welche Möglichkeiten existieren, mit Hilfe graphischer Kriterien die Identifizierbarkeit eines Modells zu bestimmen. Schließlich wird gezeigt, dass sich externe Gleichungen in den Parametern auch negativ auf die Eindeutigkeit von Beziehungen zwischen Variablen auswirken können und es wird eine Klasse von Graphen beschrieben, für die sich derartige Kantenfärbungen konstruieren lassen.

2 Grundlegende Begriffe

2.1 Lineare Strukturgleichungsmodelle

Um die Korrelationen der Einträge eines standardisierten reellen Zufallsvektors $X = (X_i)_{i \in V}$ mit $|V| = m$ genauer zu untersuchen, bietet sich die Verwendung des Modells der Kovarianzmatrizen an, also der Menge

$$PD_m = \{\Sigma \in \mathbb{R}^{m \times m} \mid \Sigma \text{ ist symmetrisch und positiv definit}\}.$$

Hierdurch ist es möglich, die Menge der multivariaten Normalverteilungen über m Variablen mit Erwartungswert Null, $\mathcal{N}_m(0, \Sigma), \Sigma \in PD_m$, zu charakterisieren (*Definition 2.4.1.* in Sullivant (2018)). Für die Parametrisierung des Modells wird nun die Annahme getroffen, die Zufallsvariable X_i setze sich aus einer Linearkombination der übrigen Einträge von X und einem Fehlerterm ϵ_i zusammen, wobei $\epsilon = (\epsilon_1, \dots, \epsilon_m)$ wieder m -vari-

Normalverteilt ist mit Erwartungswert Null und Kovarianzmatrix $\Omega = (\omega_{ij}) \in PD_m$, also

$$X_i = \sum_{j \in V \setminus \{i\}} \lambda_{ji} X_j + \epsilon_i, \quad \lambda_{j,i} \in \mathbb{R}. \quad (1)$$

Die Korrelationen zwischen den ϵ_i simulieren den Einfluss von Störvariablen, welche sich auf die beobachteten Zusammenhänge zwischen den Variablen auswirken. Während zyklische Abhängigkeiten zwischen den Variablen grundsätzlich nicht ausgeschlossen sind, soll im Folgenden nur der azyklische Fall betrachtet werden. Das bedeutet, es werden nur Fälle mit $\lambda_{ji} = 0$ falls $i \leq j$ betrachtet. Somit lässt sich obige Gleichung 1 auch schreiben als

$$X = \Lambda X + \epsilon, \quad (2)$$

mit einer strikten oberen Dreiecksmatrix $\Lambda = (\lambda_{ij}) \in \mathbb{R}^{m \times m}$. Da $(I - \Lambda)$ invertierbar ist (mit der Einheitsmatrix I), lässt sich Gleichung 2 umformen zu

$$X = (I - \Lambda)^{-1} \epsilon$$

und es folgt, etwa mit *Theorem 3.1.* aus Gut (2009)), dass X multivariat normalverteilt ist mit Erwartungswert Null und der Kovarianzmatrix

$$\Sigma = (I - \Lambda)^{-T} \Omega (I - \Lambda)^{-1}.$$

Ist $\mathbb{R}_m^O := \{\Lambda \in \mathbb{R}^{m \times m} \mid \Lambda_{ij} = 0, \text{ falls } i \geq j\}$ die Menge der strikten oberen Dreiecksmatrizen, dann ist die Abbildung $\phi : \mathbb{R}_m^O \times PD_m \rightarrow PD_m, (\Lambda, \Omega) \mapsto (I - \Lambda)^{-T} \Omega (I - \Lambda)^{-1}$ wohldefiniert, da Ω positiv definit ist und surjektiv, da $\phi(0, \Sigma) = \Sigma$ für alle $\Sigma \in PD_m$ gilt. Somit kann PD_m durch ϕ parametrisiert werden.

Meist ist es sinnvoll, die Menge der Parameter durch zusätzliche Informationen über das betrachtete System einzuschränken. Hierzu eignet sich die Verwendung *gemischter Graphen*.

Definition 1. Ein *gemischter Graph* $G = (V, B, D)$ ist ein gerichteter Graph $G_D = (V, D)$ bestehend aus einer endlichen Menge von Knoten V und gerichteten Kanten $D \subseteq V \times V$ zusammen mit einer Menge bigerichteter Kanten $B \subseteq \{\{u, v\} \mid u, v \in V, u \neq v\}$. Man schreibt auch $u \rightarrow v$ für (u, v) , bzw. $u \leftrightarrow v$ für $\{u, v\}$.

Das Wissen, dass zwischen einigen Variablen kein direkter Zusammenhang besteht, lässt sich nun auf ein graphisches Modell übertragen. Sei dazu $G = (V, B, D)$ und setze $m := |V|$. Dabei ist V die Menge der betrachteten Zufallsvariablen, $i \rightarrow j \notin D$ gilt genau dann wenn i keinen direkten Einfluss auf j nimmt und $i \leftrightarrow j \notin B$ genau dann wenn i und j keiner Störung durch eine unbeobachtete Variable unterliegen. Anders ausgedrückt werden die möglichen Ω auf die Menge

$$PD(B) := \{(\omega_{ij}) \in PD_m \mid \omega_{ij} = 0, \text{ falls } i \leftrightarrow j \notin B \text{ und } i \neq j\}$$

reduziert, bzw. die Λ auf die Menge

$$\mathbb{R}^D := \{(\lambda_{ij}) \in \mathbb{R}_m^O \mid \lambda_{ij} = 0, \text{ falls } i \rightarrow j \notin D\}.$$

Da in dieser Definition \mathbb{R}^D eine Teilmenge von \mathbb{R}_m^O ist, sind hier nur azyklische Graphen von Interesse, in denen kein Knoten auf sich selbst zeigt. Das bedeutet, dass kein Kreis gerichteter Kanten der Form $i \rightarrow \dots \rightarrow i$ existiert und $i \rightarrow i \notin D$ gilt für alle Knoten i .

Definition 2 (LSEM). Gegeben ein gemischter Graph $G = (V, B, D)$, ist ein **lineares Strukturgleichungsmodell** \mathcal{M}_G , kurz **LSEM**, das Bild der Parameterabbildung

$$\phi_G : \mathbb{R}^D \times PD(B) \rightarrow PD_{|V|}, (\Lambda, \Omega) \mapsto (I - \Lambda)^{-T} \Omega (I - \Lambda)^{-1},$$

also die Menge $\text{im}(\phi_G) \subseteq PD_{|V|}$.

Es handelt sich also um die Familie aller Kovarianzmatrizen, welche die durch den Graphen gegebenen Bedingungen erfüllen. Weitere Details zu LSEMs sind z.B. in Kapitel 14.2 in (Sullivant, 2018) zu finden.



Abbildung 1: Einfaches Beispiel eines gemischten Graphen.

Beispiel 1. Evans and Ringel (1999) untersuchten, welchen Einfluss das Rauchen während der Schwangerschaft auf das Gewicht eines Kindes bei der Geburt hat. Dieses Szenario soll nun modelliert werden. Sei dazu X_1 der Zigarettenverbrauch der Mutter und X_2 das Gewicht des Kindes. Es wird angenommen, dass es neben eines direkten Effekts noch andere externe Einflüsse gibt, welche den Zusammenhang stören. Dieser Sachverhalt wird von dem Modell \mathcal{M}_G beschrieben, wobei G der Graph aus Abb. 1 ist. Jedes $\Lambda \in \mathbb{R}^D$ hat somit die Form $\Lambda = \begin{pmatrix} 0 & \lambda \\ 0 & 0 \end{pmatrix}$ und als Parametrisierung erhält man

$$\phi_G(\Lambda, \Omega) = \begin{pmatrix} 1 & -\lambda \\ 0 & 1 \end{pmatrix}^{-T} \begin{pmatrix} \omega_{11} & \omega_{12} \\ \omega_{12} & \omega_{22} \end{pmatrix} \begin{pmatrix} 1 & -\lambda \\ 0 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} \omega_{11} & \omega_{12} + \lambda\omega_{11} \\ \omega_{12} + \lambda\omega_{11} & \omega_{22} + 2\lambda\omega_{12} + \lambda^2\omega_{11} \end{pmatrix}$$

für alle $\Omega \in PD(B)$ und $\Lambda \in \mathbb{R}^D$.

2.2 Identifizierbarkeit von Modellen

Gegeben ein bestimmter Graph G stellt sich nun die Frage, ob Λ und Ω eindeutig festgelegt sind, sobald die Kovarianzmatrix bekannt ist. Dies ist genau dann der Fall, wenn die Parameterabbildung ϕ_G injektiv ist. Nicht immer ist ϕ_G auf dem gesamten Parameterraum injektiv, manchmal gilt Eindeutigkeit auch nur bis auf eine bestimmte Teilmenge von Parametern. Allgemeiner spricht man in diesem Zusammenhang von dem Problem der *Identifizierbarkeit* eines parametrisierten Modells. Grundsätzlich sind für LSEMs mehrere Arten der Identifizierbarkeit zu unterscheiden, von denen im Rahmen dieser Arbeit folgende von Interesse sind:

Definition 3 (Identifizierbarkeit). Sei \mathcal{M}_G ein LSEM zu einem Graphen $G = (V, B, D)$ und ϕ_G dessen Parametrisierung. Dann heißt \mathcal{M} ...

... **global identifizierbar**, falls ϕ_G injektiv ist.

... **generisch identifizierbar**, falls ϕ_G injektiv auf einer Menge $(\mathbb{R}^D \times PD(B)) \setminus W$ ist, wobei $W \subset \mathbb{R}^D \times PD(B)$ das Maß Null besitzt.

... **rational identifizierbar**, falls \mathcal{M} generisch identifizierbar ist und die Inverse von ϕ_G auf $(\mathbb{R}^D \times PD(B)) \setminus W$ durch eine rationale Funktion gegeben ist.

Gibt es für ein λ_{ij} (bzw. ω_{ij}) eine Nullmenge $W \subset \mathbb{R}^D \times PD(B)$, sodass eine rationale Funktion $\phi_{\lambda_{ij}}^{-1}$ existiert mit $\phi_{\lambda_{ij}}^{-1}(\phi_G(\Lambda, \Omega)) = \lambda_{ij}$ für alle $(\Lambda, \Omega) \in (\mathbb{R}^D \times PD(B)) \setminus W$, dann heißt der Parameter λ_{ij} **rational identifizierbar**.

Für mehr Informationen und Hintergründe zu Identifizierbarkeit wird hier auf Kapitel 16 in (Sullivant, 2018) verwiesen. Im Falle der rationalen Identifizierbarkeit eines LSEMs ist das W aus obiger Definition genau die Nullstellenmenge der Nennerpolynome von ϕ_G und es kann gezeigt werden, dass die Existenz einer rationalen Formel ϕ_G^{-1} mit $\phi_G^{-1}(\phi_G(\Lambda, \Omega)) = (\Lambda, \Omega)$ für alle (Λ, Ω) , für die die Nenner von ϕ_G^{-1} nicht verschwinden, bereits generische Identifizierbarkeit impliziert. Dies ist der Fall, da die Ausschlussmenge in diesem Fall eine Nullmenge in $\mathbb{R}^D \times PD(B)$ bildet (Foygel et al. (2012a), *Introduction*). Eine wichtige Konsequenz hieraus ist, dass es zur generischen Identifizierbarkeit bereits ausreicht, Λ durch eine rationale Formel darstellen zu können. Für Ω erhält man dann nämlich den Ausdruck $\Omega = (I - \Lambda)^T \phi_G(\Lambda, \Omega) (I - \Lambda)$, welcher ebenfalls rational ist.

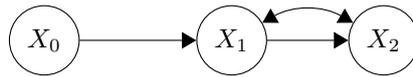


Abbildung 2: Graph unter Hinzunahme einer instrumentellen Variable.

Beispiel 2. Sei G wie in Beispiel 1. Gegeben ein beliebiges $\Sigma = (\sigma_{ij}) \in \mathcal{M}_G$, liegt (Λ, Ω) in $\phi_G^{-1}(\Sigma)$, falls $\Omega = \begin{pmatrix} \sigma_{11} & \sigma_{12} - \lambda\sigma_{11} \\ \sigma_{12} - \lambda\sigma_{11} & \sigma_{22} - 2\lambda\sigma_{12} + \lambda^2\sigma_{11} \end{pmatrix}$ mit $\Lambda \in \mathbb{R}^D$ beliebig (insbesondere folgt aus der positiven Definitheit von Σ bereits $\Omega \in PD(B)$). Somit kann \mathcal{M}_G nicht generisch identifizierbar sein und insbesondere auch nicht rational identifizierbar. Identifizierbarkeit kann erreicht werden, indem man zu einem Graphen G' mit drei Knoten übergeht. Dazu wird dem System eine sogenannte instrumentelle Variable X_0 hinzugefügt, mit sich der Effekt der Störvariablen messen lässt. Die neue Variable beeinflusst X_1 direkt, während sie mit X_2 nur über den Effekt auf X_1 in kausalem Zusammenhang steht (siehe Abb. 2). In obigem Beispiel entspricht X_0 also z.B. dem Preis der Zigaretten. Die Parameter haben somit die Form

$$\Lambda = \begin{pmatrix} 0 & \lambda_{01} & 0 \\ 0 & 0 & \lambda_{12} \\ 0 & 0 & 0 \end{pmatrix} \quad \text{und} \quad \Omega = \begin{pmatrix} \omega_{00} & 0 & 0 \\ 0 & \omega_{11} & \omega_{12} \\ 0 & \omega_{12} & \omega_{22} \end{pmatrix}.$$

D.h. man erhält für die Parameterabbildung

$$\phi_{G'}(\Lambda, \Omega) = \begin{pmatrix} \omega_{00} & \omega_{00}\lambda_{01} & \omega_{00}\lambda_{01}\lambda_{12} \\ \omega_{00}\lambda_{01} & \omega_{11} + \omega_{00}\lambda_{01}^2 & \omega_{12} + \lambda_{12}(\omega_{11} + \omega_{00}\lambda_{01}^2) \\ \omega_{00}\lambda_{01}\lambda_{12} & \omega_{12} + \lambda_{12}(\omega_{11} + \omega_{00}\lambda_{01}^2) & \omega_{22} + 2\omega_{12}\lambda_{12} + \lambda_{12}^2(\omega_{11} + \omega_{00}\lambda_{01}^2) \end{pmatrix}$$

und die Wirkung von X_1 auf X_2 gegeben $\Sigma = \phi_G(\Lambda, \Omega)$ lässt sich ermitteln über $\lambda_{12} = \frac{\sigma_{02}}{\sigma_{01}}$. Da zudem $\lambda_{01} = \frac{\sigma_{01}}{\sigma_{00}}$ gilt, besitzt Λ bereits einen rationalen Ausdruck und somit auch Ω . $\mathcal{M}_{G'}$ ist also rational identifizierbar. Für ein konkretes (Λ, Ω) ist λ_{12} offenbar genau dann

identifizierbar, wenn $\sigma_{01} \neq 0$ ist. Wie oben bereits erwähnt, ist dieser Fall in der Praxis jedoch vernachlässigbar, da es sich bei der relevanten Parametermenge um eine Menge mit Maß Null handelt.

Im Rahmen der Theorie der LSEMs ist die Kategorisierung der identifizierbaren Modelle ein offenes Problem. Zwar fanden Drton et al. (2011) für den Fall der globalen Identifizierbarkeit eine vollständige Charakterisierung, im Bereich der rationalen Identifizierbarkeit waren jedoch bisher nur Teillösungen möglich.

Häufig sind neben den Relationen innerhalb des Graphen eines Modells noch weitere, externe Informationen über die vorhandenen Parameter verfügbar. In diesen Situationen ist es erstrebenswert, die Identifizierbarkeit eines Modells unter Berücksichtigung dieser Bedingungen, etwa in Form linearer Abhängigkeiten zwischen den Komponenten von Λ und Ω , zu untersuchen. Im Folgenden soll genauer darauf eingegangen werden, wie sich Gleichheiten unter den Störeinflüssen, also den ω_{ij} , auf ein LSEM auswirken.

2.3 Gleichheit von Fehlerkovarianzen

2.3.1 Motivation

Ist bekannt, wie Störvariablen konkret auf die beobachteten Variablen einwirken, so kann man diese direkt als Variablen $L := (L_i)_{i \in \mathcal{L}}$ in die Betrachtungen miteinbeziehen. Dazu wird ein gerichteter Graph $G = (V \cup \mathcal{L}, D)$ verwendet, in dem die zusätzliche Knotenmenge \mathcal{L} die bigerichteten Kanten eines gemischten Graphen ersetzt. Eine solche *latente Variable* L_i ist nur inzident zu Kanten der Form $i \rightarrow j \in D$, wobei X_j eine beobachtete Variable ist, auf die L_i einen direkten Effekt hat. Zudem enthält D nach wie vor Kanten für alle direkten Einflüsse zwischen den beobachteten Variablen. Seien $\mu_{ij} \in \mathbb{R}$ die Parameter der Kanten $i \rightarrow j$, für die i in \mathcal{L} liegt und $\lambda_{ij} \in \mathbb{R}$ die der restlichen gerichteten Kanten. Dann haben die Variablen die Form

$$L_i = \epsilon_i, \quad i \in \mathcal{L} \quad (3)$$

im Fall der latenten Variablen und

$$X_i = \sum_{j \in V \setminus \{i\}} \lambda_{ji} X_j + \sum_{j \in \mathcal{L}} \mu_{ji} L_j + \epsilon_i, \quad i \in V \quad (4)$$

im Fall der beobachteten. Die ϵ_i sind wieder multivariat normalverteilt mit Erwartungswert Null, diesmal jedoch paarweise unkorreliert, d.h. die Kovarianzmatrix $\Omega \in PD(|V| + |\mathcal{L}|)$ hat Diagonalform.

Nun soll untersucht werden, ob die Effekte zwischen den beobachteten Variablen, also den λ_{ij} , für generische Λ eindeutig bestimmt sind, falls die Kovarianzen der X_i bekannt sind. Eine hinreichende Bedingung hierfür ist die rationale Identifizierbarkeit des LSEMs eines gemischten Graphen $G' = (V, D', B')$, welcher aus G erzeugt wurde. Dazu setzt man $D' = \{(i, j) \in D \mid i, j \in V\}$ sowie $\epsilon'_i = \sum_{j \in \mathcal{L}} \mu_{ji} \epsilon_j + \epsilon_i$ und es folgt aus den Identitäten 3 und 4, dass

$$X_i = \sum_{j \in V \setminus \{i\}} \lambda_{ji} X_j + \epsilon'_i$$

gilt für alle $i \in V$. Die ϵ'_i sind dann multivariat normalverteilt mit Erwartungswert Null und den Kovarianzen

$$\omega'_{ij} := \text{Cov}[\epsilon'_i, \epsilon'_j] = \text{Cov}\left[\sum_{i' \in \mathcal{L}} \mu_{i'i} \epsilon_{i'} + \epsilon_i, \sum_{j' \in \mathcal{L}} \mu_{j'j} \epsilon_{j'} + \epsilon_j\right] = \sum_{h \in \mathcal{L}} \mu_{hi} \mu_{hj} \omega_{hh}, \quad \forall i \neq j \quad (5)$$

wobei der letzte Schritt aus der Bilinearität der Kovarianz und der Unkorreliertheit der ϵ_i folgt. Somit ist $\omega'_{ij} = 0$ für $i \neq j$, falls keine latente Variable sowohl auf i als auch auf j einen direkten Effekt hat. B' enthält also genau diejenigen Kanten $i \leftrightarrow j$, für die i und j mindestens einen gemeinsamen Störer haben. Die möglichen Kovarianzmatrizen der X_i liegen folglich im Bild von $\phi_{G'}$. Das bedeutet, dass die rationale Identifizierbarkeit des LSEMs von G' die Existenz einer rationalen Formel für alle λ_{ij} impliziert.

Treten Homogenitäten unter den μ_{ij} auf, so spricht man von *Equiconfounding*. Chalak (2013) betrachtet dazu einige Beispiele gerichteter Graphen, in denen einige Effekte latenter Variablen gleichwertige Parameter besitzen, und prüft, ob dadurch Effekte zwischen den beobachteten Variablen identifizierbar werden. Die dabei relevanten Gleichheiten erzeugen zudem zusätzliche Relationen zwischen den ω'_{ij} in den zugehörigen gemischten Graphen, welche ebenfalls zur Identifizierung dieser Effekte verwendet werden können.



Abbildung 3: (a): Graph mit latenter Variable; die Parameter der roten gerichteten Kanten sind gleich. (b): aus linkem Graph erzeugter gemischter Graph; die Parameter der roten bigerichteten Kanten sind gleich

Beispiel 3 (Chalak (2013)). *Es sollen im Folgenden Ursachen von Veränderungen des Gehalts einer Person untersucht werden. Seien hierzu X_2 das Gehalt, X_1 eine Variable mit messbarem Einfluss auf das Gehalt (wie etwa Bildung oder Berufserfahrung) und L_0 nicht messbare Einflüsse (z.B. kognitive Fähigkeiten). Es wird angenommen, dass X_1 einen direkten Effekt auf X_2 hat und L_0 sich sowohl auf X_1 als auch auf X_2 auswirkt, also*

$$L_0 = \epsilon_0, \quad X_1 = \lambda_{01}L_0 + \epsilon_1 \quad \text{und} \quad X_2 = \lambda_{01}L_0 + \lambda_{12}X_1 + \epsilon_2.$$

Die ϵ_i sollen hierbei unabhängig und standardisiert sein. Um den Parameter λ_{12} identifizieren zu können, wird dem eine weitere Variable X_3 beobachtet, von welcher man erwartet, dass sie in gleichem Maße wie X_2 von L_0 beeinflusst wird und sonst nicht mit den X_1 und X_2 in Zusammenhang steht. Eine mögliche Messgröße hierfür ist beispielsweise der IQ. Es gilt

$$X_3 = \lambda_{03}L_0 + \epsilon_3$$

mit der Einschränkung $\lambda_{02} = \lambda_{03}$. Dieses Modell lässt sich durch den gerichteten Graphen in Abb. 3a darstellen. Der Übergang zu einem gemischten Graphen G liefert nun den

Graphen 3b, da alle beobachteten Variablen von der selben Störvariable L_0 beeinflusst werden. Zudem folgt mit Gleichung 5 und der Tatsache, dass $\omega_{22} = \omega_{33}$ gilt,

$$\omega'_{12} = \lambda_{01}\lambda_{02}\omega_{22} = \lambda_{01}\lambda_{03}\omega_{33} = \omega'_{13}.$$

Das heißt man erhält eine Homogenität in den den strukturellen Parametern von G . Diese Eigenschaft kann zur Identifizierung von λ_{12} verwendet werden. Betrachtet man die Parameterabbildung

$$\Sigma := \phi_G(\Lambda, \Omega') = \begin{pmatrix} \omega'_{11} & \omega'_{12} + \lambda_{12}\omega'_{11} & \omega'_{13} \\ \omega'_{12} + \lambda_{12}\omega'_{11} & \omega'_{22} + \lambda_{12}\omega'_{12} + \lambda_{12}(\omega'_{12} + \lambda_{12}\omega'_{11}) & \omega'_{23} + \lambda_{12}\omega'_{13} \\ \omega'_{13} & \omega'_{23} + \lambda_{12}\omega'_{13} & \omega'_{33} \end{pmatrix},$$

so erhält man die Formel $\lambda_{12} = \frac{\sigma_{12} - \omega'_{12}}{\sigma_{11}}$. Wegen $\omega'_{12} = \omega'_{13}$ ist λ_{12} also rational identifizierbar durch $\lambda_{12} = \frac{\sigma_{12} - \sigma_{13}}{\sigma_{11}}$.

2.3.2 Kantenfärbungen

Gleichheiten unter den Kantenparametern eines ungerichteten Graphen lassen sich in eine *Kantenfärbung* übersetzen. Die hier verwendete Definition ist an die in Højsgaard and Lauritzen (2008) angelehnt.

Definition 4. Gegeben ein ungerichteter/bigerichteter Graph $G = (V, E)$ und $n \leq |V|$, ist eine **Kantenfärbung** $\mathcal{C} = \{E_1, \dots, E_n\}$ eine Partition von E in disjunkte Teilmengen $E_1, \dots, E_n \subseteq E$. Eine Kante $\{i, j\} \in E$ besitzt die **Farbe** $c(i, j)$, wobei $c(i, j) = k$ ist, falls $\{i, j\}$ in E_k liegt. Gilt für eine Partitionskomponente $|E_i| = 1$, so heißt diese **atomar**.

Die Färbung \mathcal{C} des gemischten Graphen G in Beispiel 3 besitzt etwa die Komponenten $E_1 = \{1 \leftrightarrow 2, 1 \leftrightarrow 3\}$ und $E_2 = \{2 \leftrightarrow 3\}$. Besitzen die bigerichteten Kanten eines gemischten Graphen eine Färbung, dann spricht man von einem *gefärbten Graphen*. Graphen mit einer Kantenfärbung, welche ausschließlich aus atomaren Komponenten besteht, können als ungefärbte Graphen aufgefasst werden. Insbesondere ist die Menge der ungefärbten Graphen also mit einer Teilmenge der gefärbten Graphen identifizierbar. Die Färbung wirkt sich auf das zugrundeliegende statistische Modell aus, indem angenommen wird, dass zwischen den Parametern aller Kanten innerhalb einer Komponente $E_i \in \mathcal{C}$ Gleichheit gilt. Die Kovarianzmatrix Ω der Fehlerterme ϵ_i liegt dann also in der Menge

$$PD_{\mathcal{C}}(B) := \{\Omega \in PD(B) \mid \omega_{ij} = \omega_{uv}, \text{ falls } c(i, j) = c(u, v), i \neq j, u \neq v\}.$$

Das von G erzeugte Modell entspricht somit einer Einschränkung des LSEMs, welches aus G hervorgeht, wenn man die Färbung ignoriert. Kanten, welche keinen Homogenitätsbedingungen unterliegen, besitzen dabei die Farbe einer atomaren Komponente.

Definition 5. Ist $G = (V, D, B)$ ein gemischter Graph mit einer Kantenfärbung \mathcal{C} auf den bigerichteten Kanten B , dann ist das **lineare Strukturgleichungsmodell** $\mathcal{M}_G(\mathcal{C})$ **des gefärbten Graphen** gegeben durch das Bild der Parameterabbildung

$$\phi_G(\mathcal{C}) : \mathbb{R}^D \times PD_{\mathcal{C}}(B) \rightarrow PD_m, (\Lambda, \Omega) \mapsto (I - \Lambda)^{-T} \Omega (I - \Lambda)^T.$$

Der Identifizierbarkeitsbegriff kann auf solche Modelle ausgeweitet werden, indem in der ursprünglichen Definition 3 $PD_{\mathcal{C}}(B)$ anstelle von $PD(B)$ verwendet wird. In Beispiel 3 ist das Modell $\mathcal{M}_G(\mathcal{C})$ also rational identifizierbar, da λ_{12} unter der Annahme $\omega'_{12} = \omega'_{13}$ eine rationale Darstellung besitzt. Durch die Betrachtung des LSEM eines gefärbten, gemischten Graphen lassen sich somit häufig Aussagen über diejenigen Effekte zwischen beobachteten Variablen treffen, welche einem gemeinsamen Equiconfounding unterliegen.

3 Identifizierbarkeit von gefärbten Graphen

3.1 Algebraische Methoden

Eine wichtige Frage ist, in welchen Situationen es überhaupt von Nutzen ist, Gleichheiten in einem LSEM einzuführen. Meist möchte man dadurch Identifizierbarkeit ermöglichen, falls diese zuvor nicht gegeben ist. Nur in seltenen Fällen kann es vorkommen, dass ein identifizierbares Modell durch externe Einschränkungen diese Eigenschaft verliert (wie in Abschnitt 3.4 erläutert wird). Es ist daher von Interesse herauszufinden, ob ein nicht identifizierbares LSEM eines Graphen durch Hinzunahme einer Kantenfärbung rational identifizierbar wird, bzw. welche Färbungen hierfür in Frage kommen. Um dies zu entscheiden lohnt es sich, LSEMs aus einer algebraischen Perspektive zu untersuchen.

Sei $G = (V, D, B)$ wieder ein gemischter Graph mit Färbung \mathcal{C} . Dann lässt sich die Identität

$$\Sigma := \phi_G(\mathcal{C})(\Lambda, \Omega) = (I - \Lambda)^{-T} \Omega (I - \Lambda)^{-1}$$

der Parameterabbildung in ein System von Gleichungen

$$\Phi(\sigma, \lambda, \omega) := (I - \Lambda)^{-T} \Omega (I - \Lambda)^{-1} - \Sigma = 0 \quad (6)$$

in den Variablen σ, λ, ω übersetzen. Die σ, λ, ω stehen hierbei stellvertretend für alle Einträge der Matrizen Σ, Λ und Ω . Aufgrund der Symmetrie von Σ und Ω ist die Lösungsmenge dieser Gleichungen genau die Nullstellenmenge des Ideals

$$I_{\mathcal{C}}(G) := \langle \{\Phi(\sigma, \lambda, \omega)_{ij} \mid 1 \leq i \leq j \leq |V|\} \rangle \subseteq \mathbb{R}[\sigma, \lambda, \omega],$$

welches von den Polynomen auf der linken Seite in Gleichung 6 erzeugt wird.

Proposition 6. *Der Graph $G = (V, D, B)$ besitze die Kantenfärbung \mathcal{C} . Das Modell $\mathcal{M}_G(\mathcal{C})$ ist genau dann rational identifizierbar, wenn $I_{\mathcal{C}}(G) \cap \mathbb{R}[\sigma, \lambda]$ Polynome der Form $q_{ij}(\sigma)\lambda_{ij} - p_{ij}(\sigma)$, $p_{ij}, q_{ij} \in \mathbb{R}[\sigma]$ für alle Einträge $\lambda_{ij} \neq 0$ von Λ enthält, mit $q_{ij} \notin I_{\mathcal{C}}(G)$.*

Beweis. Setze $m = |V|$. Definiere für $u \in \mathbb{R}^{|\mathcal{C}|}$ und $v \in \mathbb{R}^m$ die Polynome

$$\tau(u, v)_{ij} := \begin{cases} u_{c(i,j)} & , \text{ falls } i \neq j \\ v_i & , \text{ falls } i = j \end{cases}$$

wobei $1 \leq i, j \leq m$ ist. Dadurch erhält man eine Parametrisierung aller symmetrischen $m \times m$ Matrizen, welche die durch die Kantenfärbung gegebenen Bedingungen erfüllen. Aus Lemma 14 im Anhang folgt, dass das Urbild $S := \tau^{-1}(PD_{\mathcal{C}}(B)) \subseteq \mathbb{R}^{|\mathcal{C}|+m}$ von $PD_{\mathcal{C}}(B)$

eine semialgebraische Menge ist, da es sich dabei um das Urbild einer semialgebraischen Menge unter der polynomiellen Abbildung τ handelt (Bochnak et al. (1998), Prop. 2.2.7.). Aus der Stetigkeit von τ ergibt sich außerdem, dass S offen ist, da $PD_{\mathcal{C}}(B)$ als Schnitt von $\text{im}(\tau)$ mit der offenen Menge PD_m offen in $\text{im}(\tau)$ ist. Die Behauptung folgt nun mit Proposition C.2 in Foygel et al. (2022), welche sich direkt auf $\mathcal{I}_{\mathcal{C}}(G) \cap \mathbb{R}[\sigma, \lambda]$ übertragen lässt. Die Einschränkung von τ auf S ist hierbei die verwendete Parametrisierung von $PD_{\mathcal{C}}(B)$ und $\mathcal{I}_{\mathcal{C}}(G)$ hat bereits die passende Form, da $\det(I - \Lambda) = 1$ gilt. \square

Da die Polynome $q_{ij}(\sigma)\lambda_{ij} - p_{ij}(\sigma)$ in $\mathbb{R}[\sigma, \lambda]$ liegen, lässt sich das Kriterium aus Proposition 6 noch vereinfachen. Dazu betrachtet man statt Gleichung 6 das Gleichungssystem, welches durch

$$\Phi'(\sigma, \lambda, \omega) := (I - \Lambda)^T \Sigma (I - \Lambda) - \Omega = 0 \quad (7)$$

gegeben ist. Das davon erzeugte Ideal liegt in $I_{\mathcal{C}}(G)$, da Φ' aus einer beidseitigen Multiplikation von Φ mit $(I - \Lambda)$ (bzw. $(I - \Lambda)^T$) hervorgeht. Umgekehrt erhält man Φ durch beidseitige Multiplikation von Φ' mit $(I - \Lambda)^{-1}$ (bzw. $(I - \Lambda)^{-T}$), somit gilt auch die andere Inklusion. Die beiden Ideale sind also gleich.

Proposition 7. *Sei $G = (V, D, B)$, $|V| =: m$ ein gemischter Graph mit Kantenfärbung \mathcal{C} und $\{i_k \leftrightarrow j_k \mid 1 \leq k \leq |\mathcal{C}|\}$ ein Repräsentantensystem von \mathcal{C} , d.h. $i_k \leftrightarrow j_k \in E_k$. Setze $o_{ij}(\sigma, \lambda) := [(I - \Lambda)^T \Sigma (I - \Lambda)]_{ij}$ für $i, j \in [m]$ und $o_k := o_{i_k j_k}$ für $k \in [|\mathcal{C}|]$. Definiere nun*

$$a_{ij}(\sigma, \lambda) := o_{ij}(\sigma, \lambda) - o_{c(i,j)}(\sigma, \lambda),$$

für alle $i, j \in [m]$ mit $i \leftrightarrow j \in B$. Dann gilt

$$I_{\mathcal{C}}(G) \cap \mathbb{R}[\sigma, \lambda] = \langle \{a_{ij} \mid i \leftrightarrow j \in B\} \cup \{o_{ij} \mid i \leftrightarrow j \notin B, i \neq j\} \rangle,$$

wobei die rechte Seite ein Ideal in $\mathbb{R}[\sigma, \lambda]$ ist.

Beweis. Bezeichne mit \mathcal{J} das Ideal $\langle \{a_{ij} \mid i \leftrightarrow j \in B\} \cup \{o_{ij} \mid i \leftrightarrow j \notin B, i \neq j\} \rangle$. Die Matrix Ω in Gleichung 7 erfüllt nach Annahme die Gleichheiten, welche durch \mathcal{C} festgelegt werden. Es gilt daher $\Omega_{ij} = \omega_k$ für alle Kanten $i \leftrightarrow j$ mit $c(i, j) = k$, bzw. $\Omega_{ii} = \omega_{ii}$ für die Diagonalelemente und $\Omega_{ij} = 0$ sonst. Da $I_{\mathcal{C}}(G)$ von den Φ'_{ij} aus Gleichung 7 erzeugt wird, ist

$$a_{ij} = o_{ij} - o_{c(i,j)} = \Phi'_{ij} - \Phi'_{i_{c(i,j)} j_{c(i,j)}}$$

in $I_{\mathcal{C}}(G)$ enthalten. Da zudem $o_{ij} = \Phi'_{ij} \in I_{\mathcal{C}}(G)$ gilt für alle $i \leftrightarrow j \notin B, i \neq j$, folgt somit die Inklusion $\mathcal{J} \subseteq I_{\mathcal{C}}(G) \cap \mathbb{R}[\sigma, \lambda]$.

Sei nun q ein Polynom in $I_{\mathcal{C}}(G) \cap \mathbb{R}[\sigma, \lambda]$. Dann existieren $q_{ij} \in \mathbb{R}[\sigma, \lambda, \omega]$ sodass q die Form

$$q(\sigma, \lambda) = \sum_{1 \leq i, j \leq m} q_{ij} \Phi'_{ij}$$

hat. Wendet man die Substitution $\omega_k = o_k$, bzw. $\omega_{ii} = o_{ii}$ auf die rechte Seite an, erhält man anstelle der Φ'_{ij} entweder a_{ij} , o_{ij} mit $i \neq j$ oder Null. Zudem werden die q_{ij} durch Polynome \tilde{q}_{ij} in den Variablen σ und λ ersetzt. Da q in $\mathbb{R}[\sigma, \lambda]$ liegt, ist es invariant unter der Substitution. Wendet man sie also auf beide Seiten an, erhält man

$$q = \sum_{i \leftrightarrow j \in B} \tilde{q}_{ij} a_{ij} + \sum_{\substack{i \leftrightarrow j \notin B \\ i \neq j}} \tilde{q}_{ij} o_{ij} \in \mathcal{J}.$$

Folglich ist $I_{\mathcal{C}}(G) \cap \mathbb{R}[\sigma, \lambda]$ in \mathcal{J} enthalten. \square

Mit dieser Proposition und dem Kriterium aus Proposition 6 folgt also, dass das LSEM eines gefärbten Graphen genau dann rational identifizierbar ist, wenn die Terme $q_{ij}(\sigma)\lambda_{ij} + p_{ij}(\sigma)$ in einem von Polynomen der Form $o_{ij}(\sigma, \lambda) - o_{c(i,j)}(\sigma, \lambda)$ und $o_{ij}(\sigma, \lambda)$ erzeugten Ideal liegen.

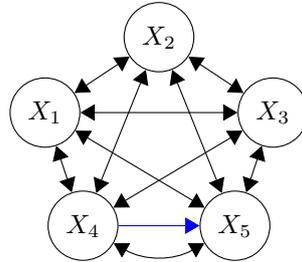


Abbildung 4: Graph mit sämtlichen bigerichteten Kanten; die gerichtete Kante ist blau markiert

Beispiel 4. Der Graph $G = (V, D, B)$ in Abb. 4 modelliert ein Szenario, in dem ein direkter Effekt zwischen den Variablen X_4 und X_5 ermittelt werden soll, obwohl keinerlei Informationen über die Existenz von Störvariablen vorliegen. Sei \mathcal{C} eine Kantenfärbung auf G . Falls keine Restriktionen bzgl. Gleichheit der Störeinträge existieren, sind alle Komponenten E_k von \mathcal{C} atomar. In diesem Fall ist das Ideal $I_{\mathcal{C}}(G)$ leer, da die a_{ij} aus Proposition 7 alle Null sind und keine i, j mit $i \leftrightarrow j \notin B$ und $i \neq j$ existieren. Wegen Proposition 6 kann das LSEM von G also nicht rational identifizierbar sein.

Im allgemeinen Fall erhält man

$$(I - \Lambda)^T \Sigma (I - \Lambda) = \begin{pmatrix} \sigma_{12} & \sigma_{13} & \sigma_{14} & \sigma_{15} - \lambda_{45}\sigma_{14} \\ \vdots & \ddots & \sigma_{23} & \sigma_{24} & \sigma_{25} - \lambda_{45}\sigma_{24} \\ \vdots & & \ddots & \sigma_{34} & \sigma_{35} - \lambda_{45}\sigma_{34} \\ \vdots & & & \ddots & \sigma_{45} - \lambda_{45}\sigma_{44} \\ \dots & \dots & \dots & & \dots \end{pmatrix}.$$

Damit kann man nun untersuchen, wann sich die Verwendung verschiedener Kantenfärbungen positiv auf die Identifizierbarkeit des Modells auswirkt:

1. Sind für alle Kanten $i \leftrightarrow 5$ die Komponenten $E_{c(i,5)}$ atomar, so verschwinden die a_{ij} mit $j = 5$ und die restlichen a_{ij} liegen in $\mathbb{R}[\sigma]$. Für jedes Polynom $q(\sigma)\lambda_{45} + p(\sigma) \in I_{\mathcal{C}}(G) \cap \mathbb{R}[\sigma, \lambda]$ muss also bereits $q \in I_{\mathcal{C}}(G) \cap \mathbb{R}[\sigma, \lambda]$ gelten, das Modell kann somit nicht rational identifizierbar sein.
2. Sind alle Komponenten bis auf die Komponente E_k atomar und enthält E_k eine Kante $l \leftrightarrow 5$, so ist rationale Identifizierbarkeit gegeben. Enthält E_k nämlich nur Kanten der Form $i \leftrightarrow 5$, so existieren Knoten u, v mit

$$a_{u5}(\sigma, \lambda) = \sigma_{u5} - \sigma_{v5} - \lambda_{45}(\sigma_{u4} - \sigma_{v4}).$$

Das Polynom $\sigma_{u4} - \sigma_{v4}$ kann nicht in $I_{\mathcal{C}}(G) \cap \mathbb{R}[\sigma, \lambda]$ liegen, da weder σ_{u4} noch σ_{v4} als Monome von Grad 1 in den erzeugenden Polynomen auftreten. Somit folgt nach

Prop. 6 die Behauptung.

Liegt ein $i \leftrightarrow j$ mit $i, j \neq 5$ in E_k , so ist

$$\begin{aligned} a_{l5}(\sigma, \lambda) - a_{ij}(\sigma, \lambda) &= (\sigma_{l5} - \lambda_{45}\sigma_{l4} - o_k(\sigma, \lambda)) - (\sigma_{ij} - o_k(\sigma, \lambda)) \\ &= \sigma_{l5} - \sigma_{ij} - \lambda_{45}\sigma_{l4} \end{aligned}$$

in $I_{\mathcal{C}}(G) \cap \mathbb{R}[\sigma, \lambda]$ enthalten, da $k = c(i, j) = c(l, 5)$ ist. Zudem kann σ_{l4} nicht in $I_{\mathcal{C}}(G) \cap \mathbb{R}[\sigma, \lambda]$ enthalten sein, da dazu a_{l4} und damit auch o_k in $I_{\mathcal{C}}(G) \cap \mathbb{R}[\sigma, \lambda]$ liegen müssten, was jedoch ausgeschlossen ist (siehe Lemma 15 im Anhang). Insgesamt erhält man also wieder eine rationale Formel für λ_{45} .

3. *Enthält die Kantenfärbung \mathcal{C} aus 2. neben E_k noch eine weitere nicht-atomare Komponente E_l , so kann es vorkommen, dass das Modell nicht mehr rational identifizierbar ist. Wähle dazu $E_k = \{1 \leftrightarrow 5, 2 \leftrightarrow 5\}$ und $E_l = \{1 \leftrightarrow 4, 2 \leftrightarrow 4\}$. Dann hat $I_{\mathcal{C}}(G) \cap \mathbb{R}[\sigma, \lambda]$ die Form*

$$\begin{aligned} I_{\mathcal{C}}(G) \cap \mathbb{R}[\sigma, \lambda] &= \langle \sigma_{15} - \sigma_{25} - \lambda_{45}(\sigma_{14} - \sigma_{24}), \sigma_{14} - \sigma_{24} \rangle \\ &= \langle \sigma_{15} - \sigma_{25}, \sigma_{14} - \sigma_{24} \rangle \subseteq \mathbb{R}[\sigma, \lambda], \end{aligned}$$

also muss für $q(\sigma)\lambda_{45} + p(\sigma) \in I_{\mathcal{C}}(G) \cap \mathbb{R}[\sigma, \lambda]$ analog zu 1. wieder $q \in I_{\mathcal{C}}(G) \cap \mathbb{R}[\sigma, \lambda]$ gelten. Es existiert also keine rationale Formel für λ_{45} .

In (Foygel et al. (2012a), Prop. 2) wird ein für die Identifizierbarkeit notwendiges Kriterium genannt, welches nur von der Anzahl der gerichteten und bigerichteten Kanten eines Graphen abhängt. Für gefärbte Graphen existiert eine ähnliche Aussage.

Proposition 8. *Ein Graph $G = (V, D, B)$ mit Kantenfärbung \mathcal{C} kann nur dann rational identifizierbar sein, wenn*

$$|D| + |\mathcal{C}| \leq \binom{|V|}{2}$$

gilt.

Beweis. Die Ungleichung lässt sich auch schreiben als $|D| + |\mathcal{C}| + |V| \leq \binom{|V|+1}{2}$. Der Beweis erfolgt durch Anwendung von Lemma 2.5 in Foygel et al. (2022). Analog zum Beweis von Korollar 2.7 in Foygel et al. (2022) kann man daraus nämlich folgern, dass das LSEM von G nur dann rational identifizierbar sein kann, wenn

$$|D| + \dim(\tau(S)) \leq \binom{|V|+1}{2}$$

gilt, wobei τ und S die selben wie im Beweis von Proposition 6 sind. Als offene semialgebraische Teilmenge von $\mathbb{R}^{|\mathcal{C}|+|V|}$ hat S die Dimension $|\mathcal{C}|+|V|$ (Bochnak et al. (1998), Prop. 2.8.4.). Da τ eine bijektive, polynomielle Abbildung ist, erhält man $\dim(\tau(S)) = \dim(S)$ (Bochnak et al. (1998), Thm. 2.8.8.) und es folgt die Ungleichung. \square

3.2 Computergestützte Analysen

Es existieren mehrere Methoden, mit denen die Identifizierbarkeit eines LSEMs auf einem ungefärbten Graphen mithilfe von *Gröbnerbasen* algebraisch überprüft werden kann (García-Puente et al. (2010), Foygel et al. (2012b)). Gröbnerbasen sind dabei ein sehr mächtiges Hilfsmittel, da mit ihnen, im Gegensatz zu graphischen Methoden, grundsätzlich beliebige LSEMs analysiert werden können. Da für die Erzeugung von Gröbnerbasen Variationen des Algorithmus von Buchberger verwendet werden (siehe hierzu Abschnitt 3.3. in Sullivant (2018)), sind rein algebraische Ansätze jedoch allgemein sehr rechenaufwändig und daher in der Praxis nur bedingt anwendbar. Ein Beispiel für die Verwendung von Gröbnerbasen ist der folgende Algorithmus, welcher von Foygel et al. (2012b) beschrieben wird.

Algorithmus 1.

Eingabe: Eine endliche Menge $S \subset \mathbb{R}[\sigma, \lambda_1, \dots, \lambda_n]$
(σ ggf. mehrere Variablen).

Ergebnis: Stellt fest, ob für jedes $i \in [n]$ ein Polynom der Form $q_i(\sigma)\lambda_i + p_i(\sigma)$ mit $q_i \notin \langle S \rangle$ in $\langle S \rangle$ enthalten ist.

1. Wähle eine Block-Monomordnung \geq auf den Variablen $\sigma, \lambda_1, \dots, \lambda_n$, sodass $\lambda_i > \sigma$ für alle $i \in [n]$ gilt.
2. Bestimme bezüglich der Monomordnung \geq eine reduzierte Gröbnerbasis \mathcal{G} des Ideals $\langle S \rangle$, welches von S erzeugt wird.
3. $\langle S \rangle$ enthält Polynome der gewünschten Form genau dann, wenn zu jedem $i \in [n]$ ein Element von \mathcal{G} existiert, dessen führendes Monom das Produkt von λ_i mit einem Monom in $\mathbb{R}[\sigma]$ ist.

(Sullivant (2018) gibt eine generelle Einführung zu Monomordnungen und Gröbnerbasen).

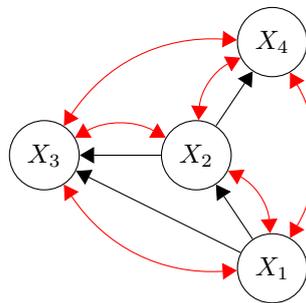


Abbildung 5: Während hier das Ideal $I_{\mathcal{C}}(G) \cap \mathbb{R}[\sigma, \lambda]$ des ungefärbten Graphen leer ist, benötigt die Berechnung der Gröbnerbasis einen Zeitaufwand von mehr als 30 Minuten (auf einem durchschnittlichen Desktop-PC), wenn alle bigerichteten Kanten dieselbe Farbe haben.

Obiger Algorithmus ermöglicht es, das Kriterium zur rationalen Identifizierbarkeit aus Proposition 6 mithilfe von Computeralgebrasystemen zu überprüfen. Als Eingabemenge

S werden dabei die a_{ij} und o_{ij} aus Prop. 7 verwendet. Die Berechnung der Gröbnerbasis ist für ungefärbte Graphen mit bis zu fünf Knoten problemlos möglich, werden jedoch Homogenitäten eingeführt, so können bereits Graphen mit nur vier Knoten einen erheblichen Rechenaufwand verursachen. Ein Beispiel hierfür ist der Graph in Abb. 5.

Um dieses Problem zu umgehen, substituieren Foygel et al. (2012b) die Einträge σ von Σ im betrachteten Ideal durch die Einträge einer zufällig gewählten Matrix $\Sigma_0 \in \mathcal{M}_G(\mathcal{C})$. Das so erzeugte Ideal liegt nun komplett in $\mathbb{R}[\lambda]$. Ergibt Algorithmus 1, dass dieses Ideal für jedes λ_{ij} ein Polynom der Form $q_{ij}\lambda_{ij} + p_{ij}$ mit $p_{ij}, q_{ij} \in \mathbb{R}$ und $q \neq 0$ enthält, dann ist das Modell mit Wahrscheinlichkeit Eins rational identifizierbar. Obwohl Σ_0 mit Wahrscheinlichkeit Null in der Ausschlussmenge der rationalen Umkehrfunktion von $\phi_G(\mathcal{C})$ liegt, kann es unter Umständen vorkommen, dass eine computergenerierte Zufallsmatrix in dieser Menge enthalten ist. In der Praxis werden daher mehrere Durchläufe mit verschiedenen Zufallsmatrizen benötigt, um eine sichere Aussage treffen zu können.

3.2.1 Experimente

Um zu verifizieren, ob eine bestimmte Kantenfärbung die rationale Identifizierbarkeit von Graphen grundsätzlich verbessert oder verschlechtert, müssen sämtliche Graphen dieser Färbung mit ihren ungefärbten Versionen verglichen werden. Graphen, welche durch Permutation der Knotenmenge eines anderen Graphen aus diesem hervorgehen, sollen hierbei miteinander identifiziert werden. Solche Graphen können als äquivalent betrachtet werden, da das Ideal $I_{\mathcal{C}}(G)$ eines Graphen G mit Kantenfärbung \mathcal{C} unter einer Permutation π der Elemente in V bis auf Umbenennung der Variablen invariant bleibt. Ist nämlich P die zu π gehörige Spalten-Permutationsmatrix und wendet man die Permutation π auf die Matrix Φ' aus Gleichung 7 an, dann erhält man

$$\begin{aligned} P^T \Phi' P &= P^T \Omega P - P^T (I - \Lambda)^T \Sigma (I - \Lambda) P \\ &= P^T \Omega P - (I - P^T \Lambda P)^T (P^T \Sigma P) (I - P^T \Lambda P). \end{aligned}$$

Ist $P^T \Lambda P$ wieder eine strikte obere Dreiecksmatrix, dann sind $\Lambda' = P^T \Lambda P$ und $\Omega' = P^T \Omega P$ die Parametermatrizen des zu G äquivalenten azyklischen Graphen G' mit Kantenfärbung \mathcal{C}' und es folgt, dass $I_{\mathcal{C}'}(G')$ und $I_{\mathcal{C}}(G)$ bis auf Variablennamen gleich sind.

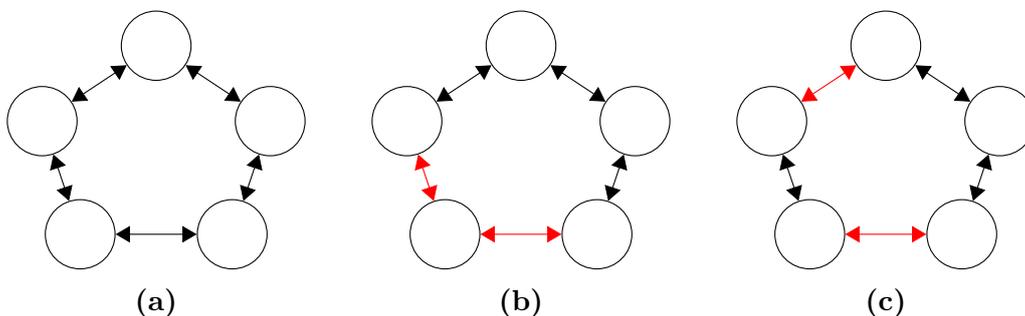


Abbildung 6: (a): unbeschrifteter, ungefärbter Graph. (b) und (c): alle Möglichkeiten, zwei Kanten gleich zu färben.

Um zu verdeutlichen, welchen Effekt die Verwendung von Kantenfärbungen auf ungefärbte Graphen haben kann, werden im Folgenden alle Graphen $G_D = (V, D, B)$ untersucht, deren bigerichteter Teil $G = (V, B)$ die Form des Graphen in Abb. 6a hat. Alle über

$ D $	Graphen insgesamt	mit Färbung (b) identifizierbar	mit Färbung (c) identifizierbar	ungefärbt identifizierbar
0	1	1	1	1
1	10	10	10	10
2	94	92	93	91
3	470	439	455	405
4	1536	1280	1378	1024
5	3090	1969	2187	1040
6	3990	1276	1452	0
7	3270	0	0	0
8	1668	0	0	0
9	480	0	0	0
10	60	0	0	0
Summe	14669	5067	5576	2571

Tabelle 1: Anzahl identifizierbarer unbeschrifteter Graphen mit Kantenfärbungen (b) und (c) aus Abb. 6

Permutationen äquivalenten Graphen werden miteinander identifiziert, die Knotenmenge V ist also unbeschriftet. Es sollen alle Kantenfärbungen \mathcal{C} betrachtet werden, in denen genau zwei Kanten gleichgesetzt werden, d.h. die einzige nicht-atomare Komponente enthält zwei Elemente. Dann hat jedes mögliche \mathcal{C} entweder die Form der Färbung in Abb. 6b oder der in Abb. 6c, es ist also ausreichend, sämtliche gerichteten Strukturen über diesen Kantenfärbungen zu berücksichtigen.

Die Werte in Tabelle 1 wurden unter Verwendung von zufälligen Matrizen für Σ ermittelt. Die Anzahl der unterschiedlichen unbeschrifteten Graphen mit gleich großer gerichteter Kantenmenge D ist für die Färbungen (b) und (c) identisch. Offenbar vergrößert sich die Menge der rational identifizierbaren Graphen durch Hinzunahme der Kantenfärbungen, falls $2 \leq |D| \leq 6$. In der Spalte „ungefärbt identifizierbar“ wird aufgeführt, wie viele Graphen G_D noch identifizierbar sind, wenn alle Komponenten der Kantenfärbung von G_D als atomar angenommen werden, der Graph also ungefärbt ist. Dabei kann es vorkommen, dass ungefärbte Graphen mehrfach betrachtet werden, da zwei unterschiedlich gefärbten Graphen derselbe ungefärbte Graph zugeordnet werden kann. Die Beobachtung, dass Graphen mit $|D| = 6$ nur unter Verwendung einer Kantenfärbung identifizierbar sein können, lässt sich mit der notwendigen Bedingung aus Proposition 8 erklären. Für diese ergibt sich nämlich für alle identifizierbaren Graphen die Ungleichung $|D| \leq \binom{5}{2} - |\mathcal{C}| = 6$ für den Fall eines gefärbten Graphen und $|D| \leq 5$ für den Fall eines ungefärbten (d.h. $|\mathcal{C}| = |V|$).

Sämtlicher *Macaulay2* Code, welcher in diesem Abschnitt verwendet wurde, ist auf GitHub zu finden unter <https://github.com/fFP2EgR/equiconfounding-m2>.

3.3 Verwendung von *partial-instrumental sets*

Bei der Identifizierung kausaler Zusammenhänge hat sich das Konzept der *instrumentellen Variable* (Bowden and Turkington, 1990) als effektives Hilfsmittel erwiesen. Dabei handelt es sich um spezielle Strukturen innerhalb eines Graphen, welche die Identifi-

zierbarkeit einzelner struktureller Parameter ermöglichen. Eine verallgemeinerte Version wird in Brito and Pearl (2002) in Form von *instrumentellen Mengen* eingeführt. Wie in Zhang et al. (2021) beschrieben, lassen sich diese Mengen auch verwenden, um externe Gleichheiten zwischen einzelnen Parametern auszunutzen. Dazu wird der Begriff des *partial-instrumental set* eingeführt.

Zur Vorbereitung werden zuerst einige graphische Konzepte definiert.

Definition 9 (Def. 1 in Brito and Pearl (2002)). Ein **Pfad** p innerhalb eines gemischten Graphen $G = (V, D, B)$ ist eine nichtleere Menge von Kanten $\{p_1, \dots, p_n\} \subseteq D \cup B$, sodass je zwei Kanten p_k und p_{k+1} zu einem gemeinsamen Knoten inzident sind. Man sagt eine Kante p_k **zeigt** auf einen Knoten i , falls entweder $p_k = j \rightarrow i$ oder $p_k = j \leftrightarrow i$ für einen Knoten j gilt. Zeigen sowohl p_k als auch p_{k+1} auf den selben Knoten i , so nennt man i einen **collider** von p .

Existiert für zwei Knoten $i, j \in V$ ein gerichteter Pfad $p \subseteq D$ ohne collider, sodass $p_1 = i \rightarrow u$ und $p_{|p|} = v \rightarrow j$ gilt für Knoten $u, v \in V$, dann heißt j ein *Nachkomme* von i . Die Menge aller Nachkommen von i wird $\text{desc}(i)$ genannt. Ist ein Knoten v inzident zu den Kanten p_k und p_{k+1} eines Pfades p mit den Endknoten i und j , dann schreibt man $p[i \sim v]$ für den Teilpfad $\{p_1, \dots, p_k\}$, bzw. $p[v \sim j]$ für $\{p_{k+1}, \dots, p_{|p|}\}$. Außerdem wird mit $\text{inc}(i)$ die Menge aller Kanten in $D \cup B$ bezeichnet, die auf den Knoten i zeigen.

Definition 10 (Def. 2 in Brito and Pearl (2002)). Sei $G = (V, D, B)$ ein Graph, $W \subseteq V$ eine Menge von Knoten und p ein Pfad in G . Dann wird p von der Menge W **blockiert**, sobald

1. W einen Knoten enthält, welcher auf p liegt und kein collider ist oder
2. p einen collider v besitzt mit $v \notin W$ und $\text{desc}(v) \cap W = \emptyset$.

Blockiert W sämtliche Pfade mit Endknoten i, j , dann werden i und j von W **d-separiert**.

In dieser Definition wird explizit nicht ausgeschlossen, dass es sich bei W um die leere Menge handelt. Partial-instrumental sets sind nun folgendermaßen definiert.

Definition 11 (Def. 4 in Zhang et al. (2021)). Gegeben ein Knoten y des Graphen $G = (V, D, B)$ und eine Menge von Kanten $E \subseteq \text{inc}(y)$, seien $E' = \{e_1, \dots, e_n\} \subseteq E$ auf y zeigende Kanten und $Z = \{z_1, \dots, z_n\} \subseteq V$ eine Knotenmenge der selben Größe wie E' . Die Menge Z ist ein **partial-instrumental set** für E auf den Kanten E' , falls Mengen $W_1, \dots, W_n \subseteq V \setminus \text{desc}(y)$ und Pfade $p(z_1), \dots, p(z_n) \subseteq D \cup B$ existieren, sodass

1. für jedes $k \in [n]$ mit $z_k \leftrightarrow y \notin B$ gilt, dass W_k die Knoten y und z_k in dem Graphen $(V, D \setminus E, B)$ d-separiert,
2. für jedes $k \in [n]$ mit $z_k \leftrightarrow y \in B$ gilt, dass E $z_k \leftrightarrow y$ enthält, $W_k \cap \text{desc}(z_k)$ leer ist und W_k die Knoten y und z_k in dem Graphen $(V, D \setminus E, B \setminus \{z_k \leftrightarrow y\})$ d-separiert,
3. für jedes $k \in [n]$ der Pfad $p(z_k)$ die Endknoten z_k und y hat, nicht von W_k blockiert wird und e_k enthält, und

4. für alle $1 \leq l < k \leq n$ der Knoten z_k nicht auf dem Pfad $p(z_l)$ liegt und für jeden gemeinsamen Knoten v von $p(z_k)$ und $p(z_l)$ die erste Kante von $p(z_l)[v \sim y]$, sowie die letzte Kante von $p(z_k)[z_k \sim v]$ auf v zeigen.

Existiert ein solches partial-instrumental set, so können basierend auf dieser Definition $|E'|$ lineare Gleichungen in den Parametern der Kanten in E konstruiert werden (Zhang et al. (2021), Lemma 1). Diese Gleichungen können nun verwendet werden, um externe Gleichheiten zwischen den mit E assoziierten Parametern zu ergänzen und so die rationale Identifizierbarkeit dieser Parameter zu ermöglichen (Zhang et al. (2021), Thm. 1). Für LSEMs, auf denen Abhängigkeiten zwischen den Parametern nur durch die Kantenfärbung des zugrundeliegenden Graphen induziert werden, erhält man damit folgendes Kriterium.

Proposition 12. *Sei $G = (V, D, B)$ ein gemischter Graph mit zugehörigem LSEM $\mathcal{M}_G(\mathcal{C})$ und Kantenfärbung \mathcal{C} . Sei außerdem $y \in V$, sodass alle Kanten in $V \setminus \text{inc}(y)$ in einer atomaren Komponente von \mathcal{C} liegen. Für y existiere eine Kantenmenge $E \subseteq \text{inc}(y)$, für die auf einer Teilmenge $E' \subseteq E$ ein partial-instrumental set Z gegeben ist. Angenommen die Parameter sämtlicher Kanten in $(D \cap E) \setminus E'$ sind bekannt, es existiert für jede Komponente $E_k \in \mathcal{C}$ mit $E_k \cap E \neq \emptyset$*

1. eine Kante $e_k \in E_k \cap E$, welche auch in E' enthalten ist oder
2. eine Kante $e_k \in E_k \cap E$, deren Parameterwert bekannt ist,

und man hat keine Informationen über Parameterwerte, welche nicht zu Kanten in E gehören. Dann sind die Parameter sämtlicher Kanten in E rational identifizierbar.

Beweis. Sei $m := |\mathcal{C}|$. Setze $E'_k := E_k \setminus (E' \cup \{e_k\})$ für jedes $k \in [m]$ und wähle die e_k so, dass e_k nur dann in E' liegt, wenn kein zu E_k gehörender Parameter bekannt ist. Dann erhält man aus der Kantenfärbung für jede Kante $i \leftrightarrow j \in E'_k$ eine Gleichung der Form $\omega_{ij} - \omega_k = 0$, wobei ω_k der Parameter der Kante e_k ist. Sei \tilde{E} die Menge aller e_k , deren Parameterwerte bekannt sind. Ohne Einschränkung ist e_k das einzige Element in E_k , welches eine der beiden Bedingungen in 1. und 2. erfüllt. Existiert nämlich eine bigerichtete Kante $u \leftrightarrow v \neq e_k$ in $E_k \cap E'$, so sind die dadurch erzeugten Gleichungen redundant, da ω_{uv} bereits in $\omega_{uv} - \omega_k = 0$ vorkommt. $u \leftrightarrow v$ kann also aus E' entfernt werden. Liegt e_k in \tilde{E} , so aus der Kantenfärbung unmittelbar, dass alle Kanten der selben Farbe den selben Wert haben. Weitere Gleichungen, welche einzelne Parameterwerte innerhalb von E_k festlegen, können daher ignoriert werden. Es wurde also sichergestellt, dass keine überflüssigen Gleichungen auftreten. Nach Definition gilt nun $\tilde{E} \cap E' = \emptyset$ und $E'_k \cap (\tilde{E} \cup E') = \emptyset$ für alle E'_k . Wegen $D \cap E \subseteq E' \cup \tilde{E}$ folgt somit

$$|E'| + |\tilde{E}| + \left| \bigcup_{k=1}^m E'_k \right| = |E|$$

und die Voraussetzungen von Theorem 1 in Zhang et al. (2021) sind erfüllt, d.h. es existieren $|E|$ lineare Gleichungen in den zu E zugehörigen Parametern, welche für generische Parameterwahlen linear unabhängig sind. Somit sind alle Parameter von Kanten in E rational identifizierbar. \square

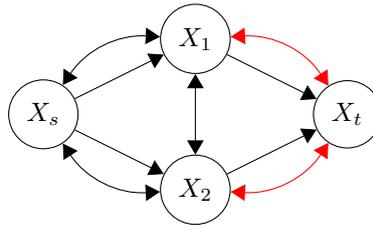


Abbildung 7: Graph aus Bsp. 5 mit $|V| = 4$; nur die roten Kanten haben dieselbe Farbe.

Beispiel 5. Sei $G = (V, D, B)$ ein Graph mit Kantenfärbung \mathcal{C} , für den zwei Knoten s und t existieren, sodass D für alle $i \in V \setminus \{s, t\}$ die Kanten $s \rightarrow i$ und $i \rightarrow t$ enthält und sonst keine weiteren. Außerdem enthalte B alle möglichen bigerichteten Kanten außer $s \leftrightarrow t$, d.h. es treten fast überall Störeinflüsse auf. Sei nun $E_0 := B \cap \text{inc}(t)$ die einzige nicht-atomare Komponente von \mathcal{C} . Für $|V| = 4$ erhält man dann den Graph in Abb. 7. Setze $E := (D \cup B) \cap \text{inc}(t)$ und $E' := (D \cap \text{inc}(t)) \cup \{v \leftrightarrow t\}$, wobei v ein beliebiger Knoten in $V \setminus \{s, t\}$ ist. Dann ist mit $Z = V \setminus \{t\}$ ein partial-instrumental set für E auf E' gegeben, indem für jedes $z_k \in Z \setminus \{v, s\}$ der Pfad $p(z_k) = \{z_k \rightarrow t\}$ und $W_k = \emptyset$ gewählt wird, bzw. $p(z_k) = \{v \leftrightarrow t\}$ und $W = \emptyset$ für $z_k = v$. Für $z_k = s$ setze $p(z_k) = \{s \rightarrow v, v \rightarrow t\}$ und $W_k = \emptyset$. Die $z_k \in Z \setminus \{s\}$ erfüllen die Bedingungen aus Def. 11, da E sämtliche gerichtete Kanten in $\text{inc}(y)$ enthält und für jedes z_k die d -Separation auf $(V, D \setminus E, B \setminus \{z_k \leftrightarrow t\})$ somit gegeben ist. Ebenso sind z_k und t auf $(V, D \setminus E, B)$ durch W_k d -separiert, falls $z_k = s$ gilt. Die einzigen Pfade mit gemeinsamen Knoten sind $s \rightarrow v \rightarrow t$ und $v \leftrightarrow t$, welche Eigenschaft 5. aus der Definition erfüllen. Z ist also ein gültiges partial-instrumental set. Da $v \leftrightarrow t \in E_0 \cap E$ in E' liegt, sind mit Proposition 12 alle Parameter der Kanten in E rational identifizierbar.

Neben der Einschränkung, dass alle nicht-atomaren Komponenten der Kantenfärbung in E enthalten sein müssen, ist das Kriterium aus Proposition 12 zusätzlich dadurch limitiert, dass es zum Teil selbst in vollständig rational identifizierbaren LSEMs nicht anwendbar ist.

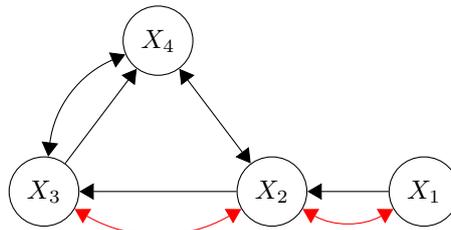


Abbildung 8: Rational identifizierbarer Graph, bei dem das graphische Kriterium versagt.

Beispiel 6. Sei $G = (V, D, B)$ der Graph aus Abb. 8 mit einer Kantenfärbung \mathcal{C} , in der nur $1 \leftrightarrow 2$ und $2 \leftrightarrow 3$ die selbe Farbe besitzen. Aus der Anwendung von Algorithmus 1 folgt, dass das zugehörige LSEM rational identifizierbar ist. Die einzigen Parameter, die potentiell mittels Proposition 12 identifiziert werden können, sind die in $\text{inc}(2)$. Für $2 \leftrightarrow 3$ kann keine instrumentelle Variable gefunden werden, da die Knoten 3 und 4 in

$\text{desc}(2)$ liegen. Es gilt also $2 \leftrightarrow 3 \notin E'$. Die beiden Kanten $1 \rightarrow 2$ und $1 \leftrightarrow 2$ müssen gemeinsam in E liegen, da sonst in keinem Fall d -Separation auf $(V, D \setminus E, B \setminus \{1 \leftrightarrow 2\})$ gegeben wäre. Da allerdings mit 1 nur eine potentielle instrumentelle Variable existiert, folgt somit $|E'| = 1$ und es kann mit Prop. 12 kein Gleichungssystem vollen Rangs in den Parametern von E gefunden werden.

Die strengen Anforderungen an die Kantenfärbung in Prop. 12 sind notwendig, da es vorkommen kann, dass nicht verwendete externe Abhängigkeiten zwischen Parametern deren Identifizierbarkeit unmöglich machen, wie im folgenden Abschnitt beleuchtet wird. Aus dem selben Grund wird in dem Beweis von Proposition 12 auch das Auftreten redundanter Gleichungen ausgeschlossen.

3.4 Durch Kantenfärbungen verursachte Nichtidentifizierbarkeit

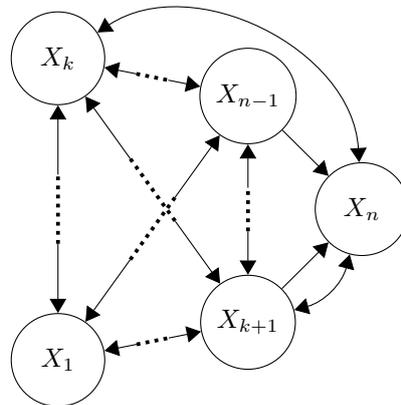


Abbildung 9: Nur die Knoten $k + 1$ bis $n - 1$ besitzen eine gerichtete Kante zu Knoten n .

Angenommen, in einem Graph $G = (V, D, B)$ mit $V = \{1, \dots, n\}$ haben alle gerichteten Kanten die Form $i \rightarrow n$ mit $i \in V$ und für $1 \leq i < j < n$ existieren alle bigerichteten Kanten $i \leftrightarrow j$, wie in Abb. 9 skizziert. Dann gilt für die Parametermatrix Λ

$$I_n - \Lambda = \begin{pmatrix} I_{n-1} & -\lambda \\ 0 & 1 \end{pmatrix},$$

wobei I_k die $k \times k$ Einheitsmatrix ist und λ der $(n-1) \times 1$ Parametervektor der gerichteten Kanten. Schreibe außerdem

$$\Sigma = \begin{pmatrix} \Sigma' & \tilde{\sigma} \\ \tilde{\sigma}^T & \sigma_{nn} \end{pmatrix}, \quad \Omega = \begin{pmatrix} \Omega' & \tilde{\omega} \\ \tilde{\omega}^T & \omega_{nn} \end{pmatrix}$$

mit den $(n-1) \times (n-1)$ Matrizen Σ', Ω' und den $(n-1) \times 1$ Vektoren $\tilde{\sigma}, \tilde{\omega}$. Dann erhält man für das Gleichungssystem in 7

$$0 = (I_n - \Lambda)^T \Sigma (I_n - \Lambda) - \Omega = \begin{pmatrix} \Sigma' & \tilde{\sigma} - \Sigma' \lambda \\ \tilde{\sigma} - \Sigma' \lambda & \lambda^T \Sigma' \lambda - 2\tilde{\sigma}^T \lambda + \sigma_{nn} \end{pmatrix} - \Omega. \quad (8)$$

Nach Proposition 7 sind die erzeugenden Polynome von $\mathcal{I}_C(G) \cap \mathbb{R}[\sigma, \lambda]$ im ungefärbten Fall daher durch die Elemente der linken Seite des Gleichungssystems $\tilde{\sigma} - \Sigma' \lambda = \tilde{\omega}$ gegeben, für die die rechte Seite Null ist. Sei Y die Menge der Indizes dieser Elemente. Man erhält also ein System aus linearen Gleichungen in λ der Form

$$[\tilde{\sigma} - \Sigma' \lambda]_k = 0 \text{ mit } k \in Y. \quad (9)$$

Nun sollen B und D so gewählt werden, dass λ rational identifizierbar ist und Gleichheiten unter den Parametern bigerichteter Kanten die Lösbarkeit des Gleichungssystems in 9 beeinflussen. Da wegen Gleichung 8 $\Sigma' = \Omega'$ gilt, soll dazu kein Diagonalelement σ_{ii} von Σ' in den Koeffizienten von λ vorkommen. Sei X die Menge der Indizes derjenigen Einträge λ_k von λ , welche nicht gleich Null sind. Es muss also $X \cap Y = \emptyset$ gelten, d.h. die Matrix $\Sigma'_{X,Y}$, welche nur aus den Elementen Σ'_{ij} mit $i \in X, j \in Y$ besteht, enthält kein $\sigma_{ii}, i \in V$. Für die Existenz einer eindeutigen Lösung der Gleichungen in 9 ist es außerdem notwendig, dass $\Sigma'_{X,Y}$ quadratisch ist, also $|X| = |Y|$ gilt. Dann existiert für λ_X die rationale Formel

$$\lambda_X = \Sigma'_{X,Y}{}^{-1} \tilde{\sigma}_Y = \frac{\hat{\Sigma}'_{X,Y} \tilde{\sigma}_Y}{\det(\Sigma'_{X,Y})} \quad (10)$$

wobei $\hat{\Sigma}'_{X,Y}$ die Adjunkte von $\Sigma'_{X,Y}$ ist. λ ist somit rational identifizierbar. Die Einträge von $\Omega'_{X,Y}$ sind aufgrund der Wahl von X und Y alle unterschiedlich, es können also beliebige Gleichheiten auf ihnen eingeführt werden, welche sich wegen Gleichung 8 auf $\Sigma'_{X,Y}$ übertragen. Verschwindet dadurch $\det(\Sigma'_{X,Y})$, so ist das Gleichungssystem in 9 nicht mehr eindeutig lösbar und somit keine rationale Identifizierbarkeit gegeben. Dies kann zum Beispiel erreicht werden, indem zwei Spalten oder Zeilen gleichgesetzt werden.

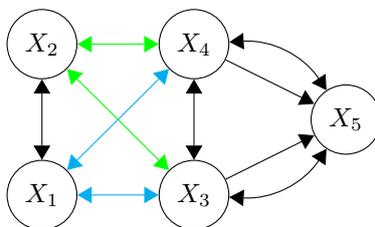


Abbildung 10: Graph mit $|V| = 5$, bei dem die Kantenfärbung Identifizierbarkeit verhindert.

Beispiel 7. Der Graph in Abb. 10 ist das einzige Beispiel für einen Graphen von oben beschriebener Form auf fünf Knoten. Die Indexmengen X und Y sind hierbei gegeben durch $Y = \{1, 2\}$, bzw. $X = \{3, 4\}$. Daraus ergibt sich das Gleichungssystem

$$\begin{pmatrix} \sigma_{13} & \sigma_{14} \\ \sigma_{23} & \sigma_{24} \end{pmatrix} \begin{pmatrix} \lambda_{35} \\ \lambda_{45} \end{pmatrix} = \begin{pmatrix} \sigma_{15} \\ \sigma_{25} \end{pmatrix}.$$

Ohne Berücksichtigung der Kantenfärbung sind $\lambda_{15}, \lambda_{25}$ also rational identifizierbar mit der Formel in 10. Wählt man allerdings C wie in Abb. 10, also mit den einzigen nicht-atomaren Komponenten $E_1 = \{2 \leftrightarrow 4, 2 \leftrightarrow 3\}$ und $E_2 = \{1 \leftrightarrow 4, 1 \leftrightarrow 3\}$, so geht diese Eigenschaft verloren. Dann ist nämlich wegen Gleichung 8 $\sigma_{13} = \omega_{13} = \omega_{14} = \sigma_{14}$ und

analog $\sigma_{23} = \sigma_{24}$, $\det \begin{pmatrix} \sigma_{13} & \sigma_{14} \\ \sigma_{23} & \sigma_{24} \end{pmatrix}$ verschwindet also für alle Kovarianzmatrizen Σ der betrachteten LSEMs.

Im Anhang sind in Abschnitt 5.2 weiter Beispiele für Graphen aufgelistet, deren Identifizierbarkeit durch eine Kantenfärbung verhindert wird. Die betrachtete bigerichtete Struktur aus Abb. 11 entspricht annähernd der in Abb. 10, nur die Kanten $1 \leftrightarrow 2$ und $3 \leftrightarrow 4$ sind nicht in B enthalten. Die kleinste Knotenmenge $V = [n]$, für die Indextmengen $X, Y \subseteq V \setminus \{n\}$ mit $X \cap Y = \emptyset$ und $|X| = |Y|$ gefunden werden kann, enthält fünf Elemente. Darüber hinaus gilt sogar folgendes Resultat.

Proposition 13. *Sei $G = (V, D, B)$ ein Graph mit $|V| \leq 4$, dessen zugehöriges LSEM rational identifizierbar ist. Dann existiert keine Kantenfärbung \mathcal{C} auf G , mit der das LSEM von G nicht mehr rational identifizierbar ist.*

Die Überprüfung von Proposition 13 findet experimentell statt, wie in Abschnitt 3.2 beschrieben. Der einzige Unterschied ist, dass nun zusätzlich zur Betrachtung sämtlicher gerichteter Strukturen einer einzigen Kantenfärbung auch noch über alle möglichen bigerichteten Strukturen und deren Kantenfärbungen iteriert werden muss.

Anhand des Graphen in Abb. 10 ist auch erkennbar, wieso redundante Gleichungen in Proposition 12 zum Versagen des Kriteriums führen können. Wie sich leicht überprüfen lässt, ist mit $Z = \{1, 2, 3, 4\}$ ein partial-instrumental set für $E = \{4 \leftrightarrow 5, 4 \rightarrow 5, 3 \leftrightarrow 5, 3 \rightarrow 5\}$ auf $E' = E$ gegeben (also ist Z sogar eine normale instrumentelle Menge). Um das zu erkennen, wähle für jedes W_k die leere Menge und setze $p(z) = \{z \leftrightarrow 5\}$, falls $z \in \{3, 4\}$, sowie $p(2) = \{2 \leftrightarrow 4, 4 \rightarrow 5\}$ und $p(1) = \{1 \leftrightarrow 3, 3 \rightarrow 5\}$. Somit sind eigentlich genügend strukturelle Gleichungen gegeben, um die Parameter der Elemente in E rational identifizieren zu können. Wie in Beispiel 7 beschrieben, besitzen jedoch aufgrund der zusätzlichen Einschränkungen durch die Kantenfärbung weder λ_{35} noch λ_{45} eine rationale Formel. Das Kriterium ist somit in diesem Fall tatsächlich nicht anwendbar.

4 Fazit

Für die Untersuchung von Homogenitäten unter den Kovarianzen der Fehlerterme eines LSEMs sind gefärbte Graphen ein nützliches Hilfsmittel. Die hier verwendete Notation der Kantenfärbung bietet dabei die Möglichkeit, Kanten mit gleichen Parametern intuitiv zu visualisieren und das zugrundeliegende Modell zu quantifizieren. In diesem Zusammenhang ist vor allem Algorithmus 1 zusammen mit Prop. 7 ein wichtiges Werkzeug, welches es erlaubt, gefärbte Graphen auf rationale Identifizierbarkeit zu überprüfen. Darüber hinaus sind mit Prop. 8 und Prop. 12 weitere Kriterien gegeben, mit denen zumindest in einigen Fällen Rückschlüsse auf die Identifizierbarkeit eines Graphen bzw. einzelner Kantenparameter möglich sind.

Eine wichtige Erkenntnis ist, dass Gleichheiten von Parametern bei der Bestimmung kausaler Zusammenhänge im Allgemeinen nicht ignoriert werden können, selbst wenn sie nicht zur Identifikation eines Parameters beitragen. Wie im letzten Abschnitt gezeigt wurde, kann es bei der Verwendung von Kantenfärbungen zu Mehrdeutigkeiten bei den Werten struktureller Parameter kommen, welche die rationale Identifizierbarkeit unmöglich

machen. Daher ist es generell notwendig, bei der Anwendung von Kriterien zur Identifizierbarkeit sämtliche durch die Kantenfärbung induzierten Gleichungen zu berücksichtigen. Gleichzeitig ist jedoch auch ersichtlich, dass Gleichheiten unter den Kantenparametern zu einer erheblichen Verbesserung der Identifizierbarkeit eines LSEMs führen können. Unklar bleibt, mit welchen Methoden sich die Auswirkungen von Kantenfärbungen effizient feststellen lassen.

5 Anhang

5.1 Hilfsaussagen

Lemma 14. Sei $G = (V, D, B)$ ein Graph mit Kantenfärbung \mathcal{C} . Dann ist die Menge $PD_{\mathcal{C}}(B)$ semialgebraisch.

Beweis. Setze $m = |V|$. Definiere die Abbildung

$$\rho : \mathbb{R}_+^m \times O(m) \rightarrow PD_m, \quad (v, Q) \mapsto Q^T \text{diag}(v) Q.$$

$O(m)$ sind hier die orthogonalen $m \times m$ Matrizen und $\text{diag}(v)$ diejenige Diagonalmatrix, deren Einträge durch die Komponenten von v gegeben sind. Dann gilt $PD_m = \text{im}(\rho)$. Wegen

$$\mathbb{R}_+^m \times O(m) = \{(v, Q) \in \mathbb{R}^m \times \mathbb{R}^{m \times m} \mid v_i > 0 \text{ für alle } i \in [m] \text{ und } Q^T Q = I\}$$

ist $\mathbb{R}_+^m \times O(m)$ eine semialgebraische Menge. Als das Bild der polynomiellen Abbildung ρ ist PD_m somit ebenfalls semialgebraisch (Bochnak et al. (1998), Prop. 2.2.7.). Definiere die Menge

$$S_{\mathcal{C}}(B) := \{\Omega \in \mathbb{R}^{m \times m} \mid \Omega_{ij} - \Omega_{kl} = 0, \text{ falls } i \neq j, k \neq l \text{ und } c(i, j) = c(k, l)\}.$$

Offenbar ist $S_{\mathcal{C}}(B)$ semialgebraisch. Die Behauptung folgt, da $PD_{\mathcal{C}}(B)$ der Schnitt von $S_{\mathcal{C}}(B)$ und PD_m ist. \square

Lemma 15. Für einen Graphen $G = (V, D, B)$ mit Kantenfärbung \mathcal{C} seien $o_k, k \in [|\mathcal{C}|]$ wie in Proposition 7. Dann ist keines der o_k in $\mathcal{I}_{\mathcal{C}}(G) \cap \mathbb{R}[\sigma, \lambda]$ enthalten.

Beweis. Betrachte das Ideal $\tilde{I} \subseteq \mathbb{R}[\sigma]$, welches entsteht, wenn man sämtliche Variablen λ in den Polynomen von $\mathcal{I}_{\mathcal{C}}(G) \cap \mathbb{R}[\sigma, \lambda]$ gleich Null setzt, also

$$\tilde{I} := \{p(\sigma, 0) \in \mathbb{R}[\sigma] \mid p \in \mathcal{I}_{\mathcal{C}}(G) \cap \mathbb{R}[\sigma, \lambda]\}.$$

Für jedes o_{ij} aus Prop. 7 erhält man $o_{ij}(\sigma, 0) = \sigma_{ij}$. Wäre o_k in $\mathcal{I}_{\mathcal{C}}(G) \cap \mathbb{R}[\sigma, \lambda]$ enthalten, so läge $\sigma_{i_k j_k}$ in \tilde{I} und damit in dem von den Polynomen

$$\begin{aligned} \tilde{a}_{ij} &:= \sigma_{ij} - \sigma_{i_l j_l}, \text{ falls } i \leftrightarrow j \in E_l \\ &\sigma_{ij}, \text{ falls } i \leftrightarrow j \notin B \end{aligned}$$

erzeugten \mathbb{R} -Vektorraum V . Die Variablen σ_{ij} mit $i \leftrightarrow j \in E_l$ treten nur in den zur Komponente E_l gehörenden Polynomen \tilde{a}_{ij} auf, es ist also ausreichend zu zeigen, dass $\sigma_{i_k j_k}$ nicht in dem Untervektorraum $V_k := \langle \{\tilde{a}_{ij} \mid i \leftrightarrow j \in E_k\} \rangle$ enthalten ist. Fasst man V_k über den natürlichen Isomorphismus als Unterraum von $\mathbb{R}^{|E_k|}$ auf, dann hat V_k das orthogonale Komplement $\langle (1, \dots, 1)^T \rangle$ und $\sigma_{i_k j_k}$ liegt genau dann in V_k , wenn $(1, 0, \dots, 0)^T \perp (1, \dots, 1)^T$ gilt. Da dies offenbar nicht der Fall ist, kann $\sigma_{i_k j_k}$ nicht in \tilde{I} liegen und o_k somit nicht in $\mathcal{I}_{\mathcal{C}}(G) \cap \mathbb{R}[\sigma, \lambda]$ enthalten sein. \square

5.2 Abbildungen

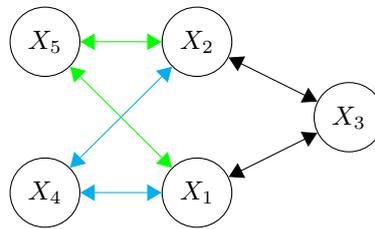


Abbildung 11: Bigerichteter Graph mit bestimmter Kantenfärbung

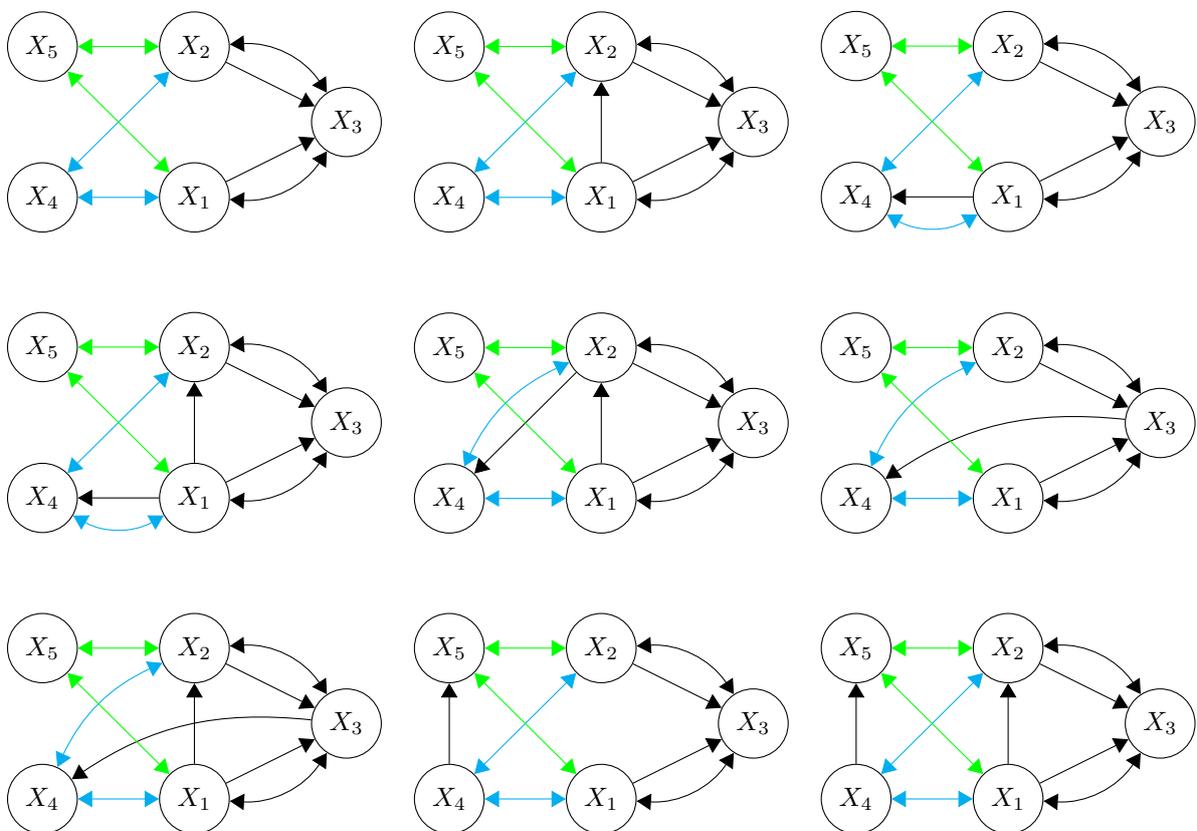


Abbildung 12: Sämtliche Graphen mit der bigerichteten Struktur aus Abb. 11, deren rationale Identifizierbarkeit durch die Kantenfärbung verhindert wird.

6 Literaturverzeichnis

- Bochnak, J., Coste, M., and Roy, M.-F. (1998). *Real algebraic geometry*, volume 36 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*. Springer-Verlag, Berlin. Translated from the 1987 French original, Revised by the authors.
- Bowden, R. J. and Turkington, D. A. (1990). *Instrumental Variables*. Cambridge University Press.
- Brito, C. and Pearl, J. (2002). Generalized instrumental variables. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, UAI'02, page 85–93, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- Chalakov, K. (2013). *Identification Without Exogeneity Under Equiconfounding in Linear Recursive Structural Systems*, pages 27–55. Springer New York, New York, NY.
- Drton, M., Foygel, R., and Sullivant, S. (2011). Global identifiability of linear structural equation models. *Ann. Statist.*, 39(2):865–886.
- Evans, W. N. and Ringel, J. S. (1999). Can higher cigarette taxes improve birth outcomes? *Journal of Public Economics*, 72(1):135–154.
- Foygel, R., Draisma, J., and Drton, M. (2012a). Half-trek criterion for generic identifiability of linear structural equation models. *Ann. Statist.*, 40(3):1682–1713.
- Foygel, R., Draisma, J., and Drton, M. (2012b). Supplement to "half-trek criterion for generic identifiability of linear structural equation models". *Ann. Statist.*, 40(3).
- Foygel, R., Drton, M., Sturman, N., and Weihs, L. (2022). Half-trek criterion for identifiability of latent variable models. <<https://arxiv.org/abs/2201.04457>>.
- García-Puente, L. D., Spielvogel, S., and Sullivant, S. (2010). Identifying causal effects with computer algebra. *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence (UAI 2010)*.
- Gut, A. (2009). *The Multivariate Normal Distribution*, pages 117–145. Springer New York, New York, NY.
- Højsgaard, S. and Lauritzen, S. L. (2008). Graphical gaussian models with edge and vertex symmetries. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 70(5):1005–1027.
- Pearl, J. (2009). *Causality and Structural Models in Social Science and Economics*, page 133–172. Cambridge University Press, 2 edition.
- Sullivant, S. (2018). *Algebraic statistics*, volume 194 of Graduate Studies in Mathematics. American Mathematical Society, Providence, RI.

- Weihls, L., Robinson, B., Dufresne, E., Kenkel, J., Kubjas, K., McGee, II, R., Nguyen, N., Robeva, E., and Drton, M. (2018). Determinantal generalizations of instrumental variables. J. Causal Inference, 6(1):Art. No. 20170009, 21.*
- Zhang, C., Cinelli, C., Chen, B., and Pearl, J. (2021). Exploiting equality constraints in causal inference. In Banerjee, A. and Fukumizu, K., editors, Proceedings of The 24th International Conference on Artificial Intelligence and Statistics, volume 130 of Proceedings of Machine Learning Research, pages 1630–1638. PMLR.*