

Causal Inference from Statistical Data

zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

von der Fakultät für Informatik
der Universität Fridericiana zu Karlsruhe (TH)

genehmigte

D i s s e r t a t i o n

von

Xiaohai Sun

aus Shanghai, China

Tag der mündlichen Prüfung: 15. April 2008
Erster Gutachter: PD Dr. Dominik Janzing
Zweiter Gutachter: Prof. Dr. Bernhard Schölkopf

【易经】

易简，而天下之理得矣；
天下之理得，而成位乎其中矣。

By means of the easy and the simple we grasp the laws of the whole world.

When the laws of the whole world are grasped, therein lies perfection.

Book of Changes

Abstract

“Automatic causal discovery” is a rather young research area, to which increasing attention is paid in recent years as more and better data have become available. Until the early nineties, most researchers still shunned away from discussing formal methods for inferring causal structure from purely observational statistical data without using controlled experiments, i.e., interventions. The seminal works of Spirtes, Glymour, and Scheines [153] and the works of Pearl [125] in the last fifteen years have established a promising basis of learning causality from such data. Bayesian networks are used as a concrete vehicle, where the corresponding directed acyclic graph can be interpreted causally. The test of statistical (conditional) independence between observed random variables provides a primary tool for learning such causal graphs. The theory and the practical applications of their approach, however, are far from fully developed. The essential shortcomings are the following. For the one thing, the test of independence is based on the strict assumption of multivariate Gaussian distribution. Moreover, if very few independence relationships are present, only few causal directions can be determined. The contribution of this thesis includes a direct attempt to address these problems.

A so-called kernel-based test of independence is further developed, which is conducted without making any specific assumption about the distribution. The kernel method maps data into an appropriate feature space by a nonlinear transformation, where also the nonlinear relations in the original space can be captured by correlations in the feature space. The singular values of the inherent covariance matrix provide a measurement of the magnitude of statistical dependences, which serves as a very useful additional tool for learning causal structures.

A new inference principle of determining the causal directions is developed for the case when no statistical independence relations are present. The complexity of conditional distributions gives hints on the causal direction in such situations that are rarely examined.

Experiments with many simulated and real-world data show that the proposed methods surpass in certain aspects other state-of-the-art approaches to the same problem.

Zusammenfassung

“Automatisiertes Erkennen von kausalen Zusammenhängen” ist ein noch recht junges Forschungsgebiet, das seit den letzten Jahren immer mehr Aufmerksamkeit bekommt, weil mehr und bessere Daten zur Verfügung stehen. Bis zum Anfang der neunziger Jahre zögerten noch die meisten Wissenschaftler sich mit dem Lernen von Ursache-Wirkungs-Beziehungen anhand von statistischen Daten zu beschäftigen, die lediglich auf Beobachtungen beruhen, d.h. ohne Zuhilfenahme von Interventionen. In den vergangenen fünfzehn Jahren sind vielversprechende Grundlagen für das maschinelle Lernen von Kausalstrukturen von Spirtes, Glymour und Scheines [153] sowie von Pearl [125] geschaffen worden. Diese beruhen auf Bayesschen Netzen, bei denen der zugehörige gerichtete azyklische Graph kausal interpretiert werden kann. Wichtigstes Hilfsmittel zum Lernen von solchen Kausalgraphen bilden dabei Tests auf (bedingte) statistische Abhängigkeiten zwischen den betrachteten Zufallsvariablen. Die Theorie und die praktische Umsetzung dieser Ansätze sind allerdings bei weitem nicht ausgereift. Die wichtigsten Unzulänglichkeiten sind folgende zu nennen: Zum einen basieren die Unabhängigkeitstests auf der starken Annahme multivariater Gauß-Verteilungen. Zum anderen lassen sich nur wenige kausale Richtungen identifizieren, wenn wenige Unabhängigkeitsbeziehungen vorliegen. Der Beitrag dieser Arbeit setzt gerade bei diesen beiden Nachteilen an.

Es wird ein sogenannter kern-basierter Unabhängigkeitstest weiter entwickelt, der ohne die Annahme einer speziellen Verteilung auskommt. Die Kernmethode bildet Daten durch eine nicht-lineare Transformation in einen geeigneten Merkmalsraum ab, in dem sich auch ursprünglich nicht-lineare Zusammenhänge als Korrelationen im Merkmalsraum manifestieren. Die Singulärwerte der Kovarianzmatrix liefern eine Quantifizierung der Stärke der statistischen Abhängigkeiten, die sich sehr gut als zusätzliches Hilfsmittel zum Lernen von Kausalstrukturen einsetzen ließ.

Es wird ein neues Inferenzprinzip entwickelt zum Schätzen von Kausalrichtungen für den bisher kaum betrachteten Fall dass keine statistischen Unabhängigkeiten vorliegen. Dabei liefert die Komplexität bedingter Verteilungen Hinweise auf die kausalrichtung.

Experimente mit simulierten und realen Daten zeigen, dass die vorgeschlagenen Methoden in mancher Hinsicht die aktuell bestehenden, anerkannten Ansätze zur Lösung des selben Problems übertreffen.

Contents

1. Introduction and Motivation	1
1.1. Causal modeling framework	1
1.2. Task of causal inference	6
1.3. State-of-the-art causal inference algorithms	9
1.4. Inductive causation	11
1.5. Thesis goal and outline	15
2. Kernel Dependence Measure	17
2.1. Linear and nonlinear dependence	17
2.2. Positive definite kernel and RKHS	21
2.3. Cross-covariance operator and independence	22
2.4. Conditional cross-covariance operator and conditional independence	23
2.5. Hilbert-Schmidt dependence measure	25
2.6. Empirical estimation of Hilbert-Schmidt dependence measure	27
2.7. Computation of empirical Hilbert-Schmidt dependence measure	30
3. Kernel Statistical Test of Independence	32
3.1. State-of-the-art tests of independence	32
3.2. Statistical test via kernel dependence measure	34
3.3. Simulated experiments with kernel independence test	36
3.3.1. Examples for kernel independence test on continuous domains	37
3.3.2. Examples for kernel independence test on time series	38
3.3.3. Numerical comparison of independence tests on continuous domain	41
3.3.4. Numerical comparison of independence tests on discrete domain	48
3.4. Real-world experiments with kernel independence test	50
3.4.1. Digoxin clearance	51
3.4.2. Rats' weights	52
3.4.3. Doctor visits and age/gender	53
4. From Independence Relations to Causal Structure	58
4.1. Logic of independence relations in DAG	58
4.2. Conflicts of representing independence relations	59
4.2.1. Relevant Independence constraints	59
4.2.2. Non-transitivity conflicts	60
4.2.3. Non-intersection conflicts	62

4.2.4.	Non-chordality conflicts	67
4.3.	Constraint-based clustering algorithm	70
4.4.	Constraint-based orientation algorithm	71
4.5.	Robust causal learning algorithm (RCL)	73
4.6.	Real-world Experiments with RCL	76
4.6.1.	College plans	76
4.6.2.	Egyptian skulls	77
4.6.3.	Montana outlook poll	78
4.6.4.	Caenorhabditis elegans	80
4.6.5.	Metastatic melanoma	85
5.	From Magnitude of Dependences to Causal Structure	90
5.1.	Problems of learning structure via independence tests	90
5.2.	Identifying colliders via magnitude of dependences	93
5.3.	Orientation heuristics via collider identification	95
5.4.	Simulated experiments with orientation heuristics	97
5.4.1.	Simulated data from noisy OR gates	97
5.4.2.	Simulated data from models with hidden common causes	101
5.4.3.	Simulated data from Asia network	102
5.4.4.	Simulated data from functional models	107
5.5.	Kernel-based causal learning algorithm (KCL)	110
5.5.1.	Some implementation issues of KCL	112
5.6.	Real-world experiments with KCL	115
5.6.1.	Ceramic surface	116
5.6.2.	Montana outlook poll	117
5.6.3.	Egyptian skulls	117
5.6.4.	Cheese data	117
5.6.5.	Smoking and cancer	119
5.6.6.	Brain size and IQ	120
5.6.7.	US crime data	122
5.6.8.	US economy data	123
6.	Discovering Causal Order by Properties of Conditionals	125
6.1.	Motivational example	125
6.2.	Plausible Markov kernel assumption	126
6.3.	Plausible Markov kernels via low-order interactions	128
6.3.1.	Smoothest Markov kernel of cause	129
6.3.2.	Smoothest Markov kernel of effect given single cause	129
6.3.3.	Smoothest Markov kernel of effect given multiple causes	130
6.4.	Examples of smoothest Markov kernels	131
6.4.1.	Numerical solution on continuous domain	131
6.4.2.	Analytical solution on hybrid (binary and real-valued) domain	134
6.4.3.	Analytical solution on binary domain	135

6.4.4.	Identifying causal order of OR/AND gates by Markov kernels	136
6.5.	plMK causal order discovery algorithm	138
6.6.	Experiments with data on binary domains	140
6.6.1.	Simulated noisy OR data	140
6.6.2.	Personal income data	141
6.7.	Combining plMK with constraint-based algorithm	142
6.8.	Experiments with data on continuous domains	145
6.8.1.	Demographic data	145
6.8.2.	Temperature data	148
7.	Discovering Causal Direction by Complexity Measure of Distributions	152
7.1.	Defining complexity measure by Hilbert space seminorms	152
7.2.	Calculation of seminorm using kernel methods	156
7.3.	Estimating densities from finite data with kernels	157
7.4.	Experiments with simulated and real-world data	159
7.4.1.	Unconditional densities	159
7.4.2.	Conditional densities	161
8.	Summary and Outlook	164
A.	Appendix	i
A.1.	Denseness of RKHS given by Gaussian RBF kernels	i
A.2.	Proof of Theorem 2	ii
A.3.	Proof of Theorem 4	iii
A.4.	Proof of Theorem 6	iv
A.5.	Plausible Markov kernels between binary and real-valued variable	vii
B.	Appendix	xi
B.1.	Pseudocode of Orientation Procedure A	xi
B.1.1.	Procedure FixOrientation	xi
B.1.2.	Procedure ProposeCollider	xii
B.1.3.	Procedure Graph2Matrix	xiii
B.2.	Pseudocode of Orientation Procedure B	xiii
B.2.1.	Procedure FixOrientation*	xiii
B.2.2.	Procedure ProposeCollider*	xiv
B.2.3.	Procedure Matrix2Graph*	xv
B.3.	Orientation Rules to Make Graphs Maximally Oriented	xv
C.	Appendix	xvii
C.1.	Numerical evidence of power increase of multiple testing	xvii
C.2.	Comparison of learning algorithms on categorical domains	xx
C.3.	Statistics of experiments with Asia Network	xxi
	Bibliography	xxv

Acknowledgments

xxxviii

Declaration of Academic Honesty

xxxix

List of Figures

1.1.	Causal structure represented by DAG	2
1.2.	Graphical representations for modeling prevalence of trisomy 21	6
1.3.	Number of possible DAGs and their Markov equivalence classes	7
1.4.	Graphical representation of v -structure	12
1.5.	Three-step-scheme of IC algorithm	14
1.6.	Stepwise results of IC algorithm	14
1.7.	Graphical representation of non- v -structure	15
2.1.	Correlation under linear dependency	18
2.2.	Correlation under quadratic dependency	18
2.3.	Correlation under periodical dependency	19
2.4.	Correlation vanishes under dependency	19
3.1.	Statistical hypothesis test of independence	33
3.2.	Permutation test of independence via kernel measures	36
3.3.	Data sampled from nonlinear functional model	37
3.4.	Dynamic Bayesian network of coupled time series	38
3.5.	Coupled Hénon maps	39
3.6.	Data sampled from coupled Hénon maps	40
3.7.	Graphical representation of functional model: Meander	41
3.8.	Meander data sampled from model in Fig. 3.7	42
3.9.	Meander data of sample size 20	43
3.10.	Results of independence tests on meander data	43
3.11.	Generating models: fork and collider structure	44
3.12.	2-dimensional plots of data sampled from fork structure	45
3.13.	3-dimensional plots of data sampled from fork structure	45
3.14.	2-dimensional plots of data sampled from collider structure	46
3.15.	3-dimensional plots of data from collider structure	47
3.16.	Graphical representation of 2-bit noisy OR	48
3.17.	Data sampled from model in Fig. 3.16	49
3.18.	Digoxin clearance data	51
3.19.	Rats' weight data	54
3.20.	Q-Q plot of p-values on doctor visit data	54
3.21.	Q-Q plot of p-values on doctor visit data with different subgroups	57
4.1.	Logic rules of faithful Bayesian networks	59

4.2.	Logic rules between constraints of different orders	60
4.3.	Rats' weight data represented by a DAG	62
4.4.	Graphical representation of Digoxin clearance data	64
4.5.	Graphical representation of Rats' weight data	64
4.6.	Data of endoderm of <i>Caenorhabditis elegans</i>	66
4.7.	Data of mesoderm of <i>Caenorhabditis elegans</i>	66
4.8.	Faithfulness in fine- and coarse-grained structures	71
4.9.	Constraint-based clustering procedure	72
4.10.	Orientation using only marginal independence	72
4.11.	Constraint-based orientation procedure	74
4.12.	Robust causal learning algorithm (RCL)	75
4.13.	Results of RCL (with χ^2 tests) on college plan data	77
4.14.	Output of RCL on Egyptian skull data	78
4.15.	Output of RCL on Montana data	79
4.16.	Output of RCL on endodermal data of <i>Caenorhabditis elegans</i>	80
4.17.	Data of maternal of <i>Caenorhabditis elegans</i>	81
4.18.	Metastatic melanoma data	81
4.19.	Output of RCL on maternal data of <i>Caenorhabditis elegans</i>	82
4.20.	Graphical representations of mesodermal data of <i>Caenorhabditis elegans</i>	83
4.21.	Output of RCL on data of <i>Caenorhabditis elegans</i>	84
4.22.	Results of RCL (with kernel measures) on metastatic melanoma data	87
4.23.	Results of RCL (with mutual information) on metastatic melanoma data	88
4.24.	Outputs of PC applied to metastatic melanoma data	88
5.1.	Q-Q plot of p-values of resampling-based test on taste score data	92
5.2.	Graphical representation of taste score data	92
5.3.	Orientation procedure A (OPA)	96
5.4.	Orientation procedure B (OPB)	98
5.5.	Graphical representation of Asia network	104
5.6.	Stepwise results of OPB on Asia network	104
5.7.	Result of OPA+K2 on Asia network	106
5.8.	Functional model with shielded collider structure	108
5.9.	Data sampled from model in Fig. 5.8	108
5.10.	Kernel dependence measures with different regularizers	109
5.11.	Kernel-based causal learning algorithm (KCL)	113
5.12.	Stepwise results of learning structure by KCL	114
5.13.	Graphical representations of ceramic surface data	116
5.14.	Graphical representations of Montana outlook poll data	117
5.15.	Graphical representations of Egyptian skulls data	118
5.16.	Graphical representations of cheese data	118
5.17.	Graphical representations of smoking and cancer data	119
5.18.	Graphical representations of brain size and IQ data	121
5.19.	Graphical representations of brain size and IQ data with variable clustering	121

5.20.	Graphical representations of US crime data	122
5.21.	Graphical representation of US crime data with variable clustering	123
5.22.	Graphical representation of US economy data with variable clustering	124
6.1.	Example for inferring direction between two dependent variables	126
6.2.	Two time layers of a first order Markov process	127
6.3.	Smoothest Markov kernel $P(X)$	132
6.4.	Smoothest Markov kernel $P(Y X)$	132
6.5.	Graphical representation of OR model with $n-1$ inputs	136
6.6.	plMK causal order discovery algorithm	140
6.7.	Graphical representation of output generated by plMK on CPS data	143
6.8.	Output of PC on CPS 2001	144
6.9.	Output of PC and PC+plMK on CPS 1995	145
6.10.	Relation between age and marriage status	145
6.11.	Relation between gender and income	146
6.12.	Recognizability of mixture of two Gaussian distributions	147
6.13.	Relation between date and temperature	148
6.14.	Smoothest Markov kernel of $DATE \rightarrow TEMPERATURE$	149
6.15.	Joint measure of smoothest Markov kernels subject to $DATE \rightarrow TEMPERATURE$	149
6.16.	Smoothest Markov kernel of $TEMPERATURE \rightarrow DATE$	150
6.17.	Joint measure of smoothest Markov kernels subject to $TEMPERATURE \rightarrow DATE$	150
7.1.	Complexity measures of sampled data	160
7.2.	Daily average temperature data	160
B.1.	Orientation rules to make graph maximally oriented	xv
B.2.	Examples of orientation rules in Fig. B.1	xvi
C.1.	Resampling-based multiple independence test	xviii
C.2.	Multiple independence test for noisy OR Data	xviii

List of Tables

1.1.	Independence relations of v - and non- v -structure	13
3.1.	Type I and II errors made by hypothesis tests of independence	33
3.2.	Kernel independence test on time series of coupled Hénon maps	40
3.3.	Independence tests on continuous Meander data	42
3.4.	Functional models defined by pairs of functions	44
3.5.	Independence tests on data sampled from fork structure	46
3.6.	Independence tests on data sampled from collider structure	47
3.7.	Independence tests on Noisy OR data	50
3.8.	Independence tests on digoxin clearance data	52
3.9.	Independence tests on rats' weight data	53
3.10.	Multiple hypothesis testing on doctor visit data	56
4.1.	Genes involved in <i>Caenorhabditis elegans</i>	65
4.2.	Dependence measures of endoderm of <i>Caenorhabditis elegans</i>	67
4.3.	Dependence measures in mesoderm of <i>Caenorhabditis elegans</i>	67
4.4.	Handling non-chordality conflicts by different strategies	69
4.5.	Structures learned by marginal independence relations	73
4.6.	Genes involved in metastatic melanoma data	86
5.1.	Independence tests on taste score data	91
5.2.	2/3-bit noisy OR models	99
5.3.	Ratio of empirical kernel dependence measures of OR model	99
5.4.	Learning noisy OR gates by different algorithms	100
5.5.	Learning structures with hidden common causes by orientation heuristics	102
5.6.	Learning structure with hidden common causes by OPB (Tab. 5.5, column 4)	103
5.7.	Learning structure with hidden common causes by OPB (Tab. 5.5, column 5)	103
5.8.	Ratios of kernel dependence measures of Asia network data	105
5.9.	Ratio of empirical kernel dependence measures of functional model	110
5.10.	Identifying collider structure in functional models	111
6.1.	Constraint parameters for plots in Fig. 6.3	133
6.2.	Constraint parameters for plots in Fig. 6.4	133
6.3.	Results of pMK on noisy OR data	141
6.4.	Scheffé tournament on CPS data	142
6.5.	Result of pMK on CPS data	143

6.6.	Result of pMK by learning relation between age and marriage status	146
6.7.	Result of pMK by learning relation between gender and income	146
6.8.	First and second moments of temperature data	148
7.1.	Complexity of conditional densities in binary mixture models.	161
C.1.	Single and multiple independence test on discrete domain	xix
C.2.	Single and multiple independence tests on continuous Meander data	xx
C.3.	Performance of algorithms on 2-bit noisy OR	xxi
C.4.	Performance of algorithms on 3-bit noisy OR	xxii
C.5.	Performance of OPA and OPB on Asia network	xxiii
C.6.	Performance of OPA+K2 on Asia network	xxiv

1. Introduction and Motivation

The central aim of many studies in medicine, biology, sociology, and economics is the elucidation of cause-and-effect relations among variables or events. In many real-life situations, randomized controlled experiments or interventions cannot always be utilized to provide causal knowledge. Methods for finding a causal structure from purely observational data are of special interest. The first chapter introduces the framework of causal modeling and summarizes the seminal works of Spirtes et al. [153] and Pearl [125], which showed that, under reasonable assumptions, it is possible to get hints on causal relationships from non-experimental data.

1.1. Causal modeling framework

Since the early nineties, it has become popular to express causal relations by a graphical representation, the so-called causal structure or causal graph.

Definition 1 (Causal Structure) *A causal structure of a set of random variables \mathcal{V} is a directed acyclic graph (DAG) \mathcal{G} in which each node (vertex) corresponds to a distinct element X or a set of distinct elements $X := (X_1, X_2, \dots)$ of \mathcal{V} , and each arrow (a directed edge) represents direct causal relationship between the corresponding nodes.*

For example, Fig. 1.1 illustrates a causal structure with 7 nodes representing variables X_1, \dots, X_7 and each arrow from X_i to X_j is interpreted as a direct causal influence of X_i on X_j . When talking about the relations in a DAG, we use the wording of family relations: if there is a link from X_i to X_j , we say that X_i is a parent of X_j . Some authors [8, 68] call such a graph an “acyclic digraph” instead of DAG. The corresponding undirected graph of a DAG \mathcal{G} is called the adjacency structure (or skeleton) of \mathcal{G} .

Regarding notations, we will normally not sharply distinguish between a single random variable and a set of variables. The capitalized variables X, Y, Z, \dots are used to depict a single variable or a set of variables. A vertex in \mathcal{G} normally corresponds to a random variable in \mathcal{V} , but it can also represent a set of variables, if necessary. Therefore, X, Y, Z, \dots are also used to depict a vertice representing the corresponding variable or the set of variables. The context will make clear whether variables or vertices are meant.

As a graphical representation, a DAG is capable of displaying cause-and-effect relationships between variables intuitively and clearly. Furthermore, a DAG can handle uncertainty through the established theory of probability on graphical models. A DAG \mathcal{G} with the probabilistic interpretation represents a probability distribution P . The primary link between the topology of a DAG and the underlying probability distribution P is the independence relations between variables. We recall the formal definition of conditional independence.

1. Introduction and Motivation

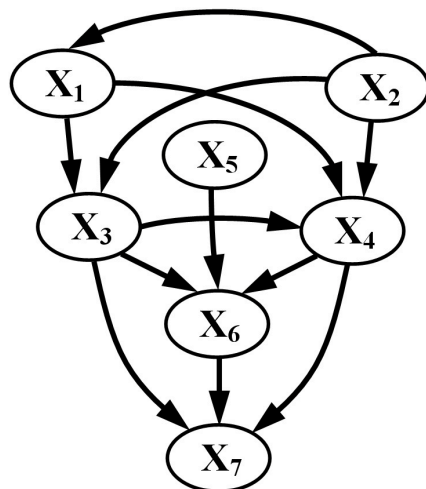


Figure 1.1.: Graphical representation of direct causal influences between 7 variables by DAG, i.e., causal structure.

Definition 2 (Independence Relation) Let $X := (X_1, \dots, X_m)$, $Y := (Y_1, \dots, Y_n)$, $Z := (Z_1, \dots, Z_k)$ be three disjoint subsets of the set \mathcal{V} of all measured variables. Conditional on Z , X , and Y are independently related to each other if and only if

$$P(x_1, \dots, x_m, y_1, \dots, y_n | z_1, \dots, z_k) = P(x_1, \dots, x_m | z_1, \dots, z_k) P(y_1, \dots, y_n | z_1, \dots, z_k)$$

for all possible values x_j of X_j , y_i of Y_i , and z_l of Z_l .

We use a notation of independence relations introduced by Dawid [46]: $X \perp\!\!\!\perp Y | Z$ meaning X and Y are independent conditional on Z . If Z is empty, X and Y are said to be marginally (unconditionally) independent when their joint probability can be factorized in the same way:

$$X \perp\!\!\!\perp Y \Leftrightarrow X \perp\!\!\!\perp Y | \emptyset \Leftrightarrow P(x, y) = P(x)P(y)$$

for all possible values x of X and y of Y . If X and Y are not unconditionally or conditionally independent, then they are said to be unconditionally or conditionally dependent, denoted as $X \not\perp\!\!\!\perp Y$ or $X \not\perp\!\!\!\perp Y | Z$ respectively. One of the essential alternatives to relate the graphical structures to the conditional independence relations is the so-called Markov condition [153, 125].

Definition 3 (Markov Condition) Let \mathcal{G} be a DAG and P the joint distribution over a set of variables \mathcal{V} . Let $X \subseteq \mathcal{V}$ be a variable or set of variables that is represented by a node in \mathcal{G} . The pair (\mathcal{G}, P) satisfies the Markov condition if and only if, conditional on all of X 's direct parents in \mathcal{G} , every X is independent in P of every other variable or set of variables that is represented by a node in \mathcal{G} , excepting its descendants.

The pair (\mathcal{G}, P) which satisfies the Markov condition is called a Bayesian network [118]. The Markov property can be found in many areas of research in order to approximate problems which

are too complex otherwise. For example, a (first order) Markov process assumes that knowing a system’s current state is relevant to its future, but knowing how it got to its current state is completely irrelevant. The intuition behind the Markov property for causal structures is that ignoring a variable’s effects, all the relevant probabilistic information about a variable that can be obtained from a model is contained in its direct causes. Therefore, the Markov condition is used as a bridge principle linking the causal interpretation of a DAG to its probabilistic interpretation. Variants of the Markov properties of causal structures have been discussed by many philosophers [131, 163, 146, 135, 27]. Lauritzen [98] distinguished among the pairwise, local and global Markov properties. However, it can be shown that all three Markov properties are equivalent for a strictly positive probability distribution.

A more generally useful graphical relation in DAGs: d-separation [124] (“d” for “directed” or “dependence”) turns out to be equivalent to the global Markov property of Bayesian networks [99], and hence also to the other Markov properties provided that the probability distribution is strictly positive.

Definition 4 (d-Separation) *In a DAG \mathcal{G} , two disjoint sets of nodes X and Y are d-separated by a set of nodes S_{XY} (excluding X and Y), if and only if along every path between a node in X and a node in Y there is a node Z (distinct from X and Y) satisfying one of the following two conditions:*

- (1) Z is a collider on the path and none of Z or its descendants are in S_{XY} , or
- (2) Z is not a collider and Z is in S_{XY} .

A node Z in DAG is a collider on a path if two arrow heads meet at Z , i.e., $\rightarrow Z \leftarrow$, otherwise Z is called a non-collider on the path. An unshielded collider on Z (also called v -structure) in a DAG \mathcal{G} is a substructure $X \rightarrow Z \leftarrow Y$ in \mathcal{G} for three distinct nodes, where X and Y are not adjacent to each other (see Fig. 1.7). If X and Y are adjacent, we call it a shielded collider. The corresponding adjacency structure is called unshielded or shielded triple respectively. Throughout this thesis, collider can be unshielded or shielded, unless explicitly stated otherwise.

The notation of d-separation is, in particular, defined for two distinct nodes X and Y . Definition 4 implies that a path between two nodes X and Y in \mathcal{G} is blocked when one of the conditions is fulfilled, and activated otherwise. X and Y are d-separated by a set S_{XY} , when all paths between them are blocked, otherwise, we call that X and Y are d-connected. By choosing d-separation to link DAGs to probability distributions, one assumes that the disjoint subsets of variables $X \subset \mathcal{V}$ and $Y \subset \mathcal{V}$ are independent conditional on $Z \subseteq \mathcal{V} \setminus \{X \cup Y\}$ in all of the distributions P that a DAG \mathcal{G} can represent, if vertices X and Y are d-separated by a set of vertices Z in \mathcal{G} .

Given the Markov condition or the d-separation criterion, several different DAGs may determine the same set of conditional independence restrictions on the set of measured variables.

Definition 5 (Markov Equivalence) *Two DAGs \mathcal{G}_1 and \mathcal{G}_2 on a set of nodes are Markov equivalent if and only if*

- (1) \mathcal{G}_1 and \mathcal{G}_2 have the same adjacencies, and
- (2) \mathcal{G}_1 and \mathcal{G}_2 have the same unshielded colliders.

1. Introduction and Motivation

If no further assumptions are made, the Markov condition and d-separation are just mathematics connecting DAGs and probability distributions and need not involve causation at all. One might use this mathematical theory solely to produce a compact and elegant representation of independence structures, i.e., Bayesian networks (\mathcal{G}, P) . The probability distribution P can be specified by the conditional distributions with respect to the corresponding DAG \mathcal{G} . More specifically, suppose N random variables X_1, \dots, X_N are measured. We denote the joint distribution P by

$$P(X_1, \dots, X_N) \quad \text{or} \quad P_{X_1 \dots X_N},$$

which is described by the values

$$P(X_1 = x_1, \dots, X_N = x_N) \quad \text{or} \quad P(x_1, \dots, x_N),$$

where (x_1, \dots, x_N) runs over all possible N -tuples. Since we assume that all probability measures are represented by densities, P is interpreted as a probability density throughout this thesis. According to an iterated application of Bayes' rule one may factorize the joint probability measure into

$$P(x_1, \dots, x_N) = P(x_1) P(x_2|x_1) \dots P(x_N|x_1, \dots, x_{N-1}) = \prod_{j=1}^N P(x_j|an_j). \quad (1.1)$$

The rightmost term in Eq. (1.1) is just a short notation, since $an_j := (x_1, \dots, x_{j-1})$ denotes the values of all $j-1$ ancestors $AN_j := (X_1, \dots, X_{j-1})$ of X_j . Obviously, any reordering $X_{\pi(1)}, X_{\pi(2)}, \dots, X_{\pi(N)}$, where $\pi \in \mathbf{S}_N$ is a permutation, defines a distinct corresponding factorization into some other conditional probability measures.

Furthermore, if P satisfies the Markov condition with respect to a DAG \mathcal{G} , the joint measure can be decomposed into

$$P(x_1, \dots, x_N) = \prod_{j=1}^N P(x_j|pa_j), \quad (1.2)$$

where pa_j depicts the tuple of values of all k_j parents $PA_j \subseteq \mathcal{V} \setminus \{X_j\}$ of X_j in \mathcal{G} . If \mathcal{G} can be indeed interpreted causally, each term $P_\pi(X_j|PA_j)$ ($j = 1, \dots, N$) formalizes the distribution of an effect given the values of all its direct causes. The conditional probability measures $P(X_j|PA_j)$ for each node X_j are called the Markov kernels corresponding to \mathcal{G} . All these N Markov kernels together define uniquely a joint measure over the N variables.

Definition 6 (Causal Model) *A causal model is a pair $(\mathcal{G}, \mathcal{P}_{\mathcal{G}})$ consisting of a causal structure \mathcal{G} and a set of parameters $\mathcal{P}_{\mathcal{G}}$ compatible with \mathcal{G} . The parameters $\mathcal{P}_{\mathcal{G}}$ assign a probability measure $P(X_j|PA_j)$ (the so-called Markov kernel) to each node $X_j \subseteq \mathcal{V}$, where $PA_j \subseteq \mathcal{V} \setminus \{X_j\}$ are the parents of X_j in \mathcal{G} .*

In other words, a causal structure serves as a blueprint for forming a causal model, which specifies how each variable is influenced by its parents in the structure. Due to the acyclicity, a DAG entails an ancestral ordering on the variables. Typically, a DAG does not determine a unique total

ordering, but merely a partial ordering. If there are no valid (conditional) independence relations, we only need to focus on complete acyclic causal graphs $\mathcal{G}_\pi^{\text{complete}}$ which are defined by an ordering π of the nodes and drawing arrows from each node to all its successors. In the general case, i.e., some (conditional) independence relations are valid, one can easily identify any causal graph \mathcal{G} as an proper subgraph embedded in a suitable $\mathcal{G}_\pi^{\text{complete}}$ by checking for each node X_j the set of its parents in $\mathcal{G}_\pi^{\text{complete}}$ which can be dropped without changing the Markov kernels $P(x_j|pa_j)$ and consequently the joint probability measure P . More generally, we may consider the factorization of Eq. (1.2) as the special case where \mathcal{G} is the unique complete (fully connected) acyclic graph that corresponds to the ordering X_1, \dots, X_N , i.e., \mathcal{G} has arrows from each X_i to every X_j with $i < j$. Likewise, we call $P_\pi(x_j|an_j)$ Markov kernels corresponding to an ordering π .

Since a Bayesian network represents the joint distribution, learning Bayesian networks can reveal insights into the underlying causal model that the observed data come from. But, the causal model should have more power than only representing the underlying joint measure. The principal quality and power that distinguish a causal structure from the graphical representation of a closely related Bayesian network is the assigned ability to exhibit causal knowledge from data instead of merely representing dependences. In doing so, a causal model becomes suitable for predicting the effect of potential interventions or actions in different circumstances. However, there are some fundamental difficulties to interpret a Bayesian network causally (called causal Bayesian network).

First, causality itself is yet not a well-understood concept. Whether a causal relation is a property of the real world or rather a concept in our minds helping us to organize our perception of the real world is a very philosophical question rather than a very scientific one. Even though the intuitive meaning of cause and effect in real life often is quite clear (not always obvious), there is a lack of widely accepted clear notation of cause-and-effect relationship in scientific research. For these reasons, we preferably treat causality as a primitive throughout this thesis.

Nevertheless, we propose to keep the concept of manipulation criterion [153] in mind, since the main benefit of having causal knowledge is being able to predict the effect of a manipulation or intervention. The manipulation criterion characterizes the causal influence of X_i on X_j in the manner that if one had a way of setting just the values of X_i and then measuring X_j , the causal influence of X_i on X_j will be reflected as a change in the distribution of X_j . That is, there exist states $x_i^{(1)}$ and $x_i^{(2)}$ of X_i which can be set, formalized as the so-called do-calculus [125, 183], such that

$$P(X_j | \mathbf{do} X_i = x_i^{(1)}) \neq P(X_j | \mathbf{do} X_i = x_i^{(2)}). \quad (1.3)$$

Roughly speaking, if one can manipulate something and something else changes, then the former causally influences the latter. The impact of a manipulation or an intervention will spread in the causal direction, but not opposite to the causal direction. If a Bayesian network does not reflect the real causal directions, it cannot be used to simulate the impact of interventions and consequently should not be interpreted causally.

By means of a real-life example, we clarify the difference between the graphical representation of a Bayesian network and the causal structure. Suppose that the leftmost plot of Fig. 1.2 illustrates the graphical representation of a Bayesian network for having Down's syndrome (prevalence of trisomy 21) with maternal age, two blood markers, i.e., free β -HCG (β -human chorionic

1. Introduction and Motivation

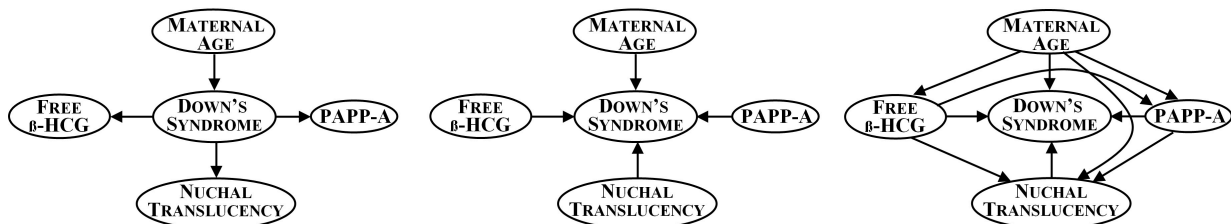


Figure 1.2.: Graphical representations for having Down's syndrome (prevalence of trisomy 21): sparse Bayesian network (leftmost), diagnostic model with wrong independence relations (middle) and fully connected Bayesian network (rightmost).

gonadotrophin) and PAPP-A (pregnancy associated plasma protein-A), and nuchal translucency (NT). It is well known that the risk of having a term pregnancy with Down's syndrome increases with maternal age and the blood markers and NT are highly correlated with Down's syndrome [151, 148]. The arrows in the Bayesian network can be interpreted causally due to a thought experiment of the hypothetical interventions: if Down's syndrome really could be treated with success, the measurement of blood markers and NT would change, but not maternal age. In contrast, although a fully connected Bayesian network (rightmost plot) is also able to represent the observed distribution perfectly (we will elaborate on the reason later), we cannot interpret it causally, because it is not capable of simulating the impact of interventions.

Another problem of interpreting a Bayesian network causally is the existence of potential hidden common causes (confounders) of measured variables. For instance, if there is some variable that is a cause of both maternal age and Down's syndrome, e.g., a genetic factor, then the arrow between them in the Bayesian network (leftmost plot of Fig. 1.2) is not an accurate depiction of the causal relationships for Down's syndrome. One possibility is to manage the measuring so that the set of variables \mathcal{V} include all of the common causes of pairs in \mathcal{V} , the so-called causal sufficiency assumption. Unless explicitly stated otherwise, we assume throughout this thesis that no common causes of any pair of variables in the graph is left out. Another more general possibility to enable one to focus on the structure over the measured variables that results from the presence of unmeasured variables without explicitly including them in the model is the concept of ancestral graphs [133] permitting undirected and bi-directed edges, which indicates sampling bias and confounding respectively.

1.2. Task of causal inference

The situation that we would like to focus on in this thesis is the following. An underlying process generates entities that share the same causal structure \mathcal{G} over a set of variables $\mathcal{V} := \{X_1, X_2, \dots\}$. The entities may have different parameters, i.e, probabilities $\mathcal{P}_{\mathcal{G}}^{(i)}$. We assume that each entity independently samples the joint distribution $P^{(i)}$ defined by its causal model $(\mathcal{G}, \mathcal{P}_{\mathcal{G}}^{(i)})$ to generate data points $(x_1^{(i)}, x_2^{(i)}, \dots)$ of all variables in the model. Admittedly, one cannot be sure that in

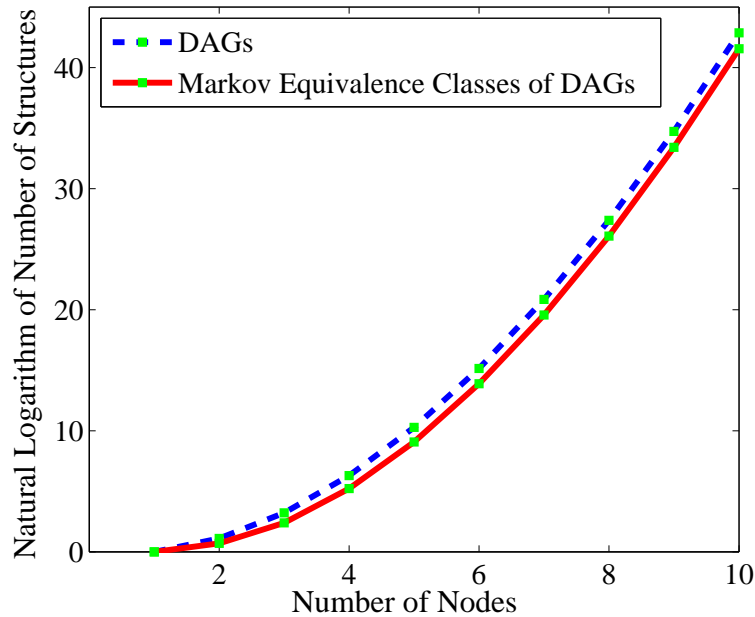


Figure 1.3.: The number of nodes against the natural logarithm of the number of possible DAGs and Markov equivalence classes of DAGs.

the real world the observed data are indeed sampled from an “underlying model”. Nonetheless, we assume in addition that the observed data fairly reflect the probabilities determined by the underlying model, i.e., the relative frequencies from data is very close to the actual underlying probabilities, provided that the sample size is large enough. Causal inference copes with the task to estimate the underlying structure representing causal relationships from finite data.

In principle this task can be done by performing parameter learning for all possible structures, and then selecting those structures for which the joint probability measure over the domain is sufficiently close to the observed measure. Unfortunately, by following such brute force approach we will be faced with the essential difficulties of structural learning. The space of all DAGs or Markov equivalence classes of DAGs is extremely large. In fact, it is known [134, 156, 83, 111] that given N labeled vertices the number of DAGs can be counted a recurrence equation. Moreover, Gillispie et al. [68] wrote a computer program to count the equivalence classes of DAGs up to 10 nodes. Fig. 1.3 shows the natural logarithm of both numbers, which indicate a super-exponential growth of possible structures in the number of nodes. For instance, there exists nearly 4.18×10^{18} different DAGs and approximately 1.12×10^{18} different Markov equivalence classes of DAGs with 10 nodes.

The other problem of the brute force search strategy is that we may end up the search through the structures with several equally good candidates. In particular, a Bayesian network over all complete graphs can represent any probability measure over its domain, consequently represents the observed measure exactly. Although such a causal model over complete graphs could be considered as the generating model, it will not be a preferable answer, when the data could be

1. Introduction and Motivation

sampled from a sparse network. Moreover, human beings generally prefer simple answers to things. Therefore, some kind of simplicity principle (called Ockham’s razor) in causal inference is desirable. The simplicity of a causal model $(\mathcal{G}, \mathcal{P}_{\mathcal{G}})$ is primarily reflected in the causal structure, i.e., the number of links in the DAG \mathcal{G} . The parameters $\mathcal{P}_{\mathcal{G}}$ of a model, i.e., the set of Markov kernels, provide an additional criterion for the simplicity.

To make the goodness of a candidate for the causal structure apparent, we consider the graphical representation of having Down’s syndrome (Fig. 1.2) as an example again. The blood markers and NT are widely accepted as identification tests for Down’s syndrome, because they are directly influenced by Down’s syndrome. When modeling the medical diagnosis of having Down’s syndrome, trained physicians are usually inclined to provide conditional probabilities in diagnostic directions, e.g., $P(\text{Down’s Syndrome} \mid \text{Indicator})$. The shorthand “Indicator” stands for free β -HCG, PAPP-A, and NT. A model reflecting this might look like the one in the middle plot of Fig. 1.2, which has some arrows opposite to directions in leftmost plot. However, according to this diagnostic model, maternal age, free β -HCG, PAPP-A, and NT are mutually independent, which is inconsistent with the Bayesian network as shown in leftmost plot. If we would like to correct the model (in the sense that the Bayesian network in the leftmost plot is the underlying model) to be a generating model in form of a Bayesian network, one must add some extra structure making maternal age, blood markers, and NT dependent, for example the rightmost plot of Fig. 1.2. Although the observed probability measure (actually any probability measure) can be perfectly represented by such a fully connected DAG (rightmost plot), one would not consider this structure as a good candidate for the underlying causal model, since there is a simpler structure (leftmost plot) which can represent the observed probability measure as well as the much more complicated structure. This example makes a main feature of a good candidate for the causal structure apparent, namely somewhat minimality in the structure with respect to links. In other words, if for some reason one wishes to represent a hypothetical relation by a DAG with some links directed opposite to the true causal direction, the total number of links in the hypothetical DAG that correctly represents the independence relations can not decrease, and most likely it will increase.

From the viewpoint of parameterization, we expect that a good candidate for the causal model $(\mathcal{G}, \mathcal{P}_{\mathcal{G}})$ should be stable or simple in the set of parameters $\mathcal{P}_{\mathcal{G}}$, i.e., the set of Markov kernels. Suppose the leftmost and rightmost plots of Fig. 1.2 are two Bayesian networks that can correctly describe the observed joint probability measure. For the sake of simplicity, we assume that the maternal age is given. The models based on the leftmost DAG have advantage over the models based on the rightmost DAG, namely that the conditional probabilities $P(\text{Indicator} \mid \text{Down’s Syndrome})$ (Markov kernel with respect to the leftmost structure) are more stable than the conditional probabilities $P(\text{Down’s Syndrome} \mid \text{Indicator})$ (Markov kernel with respect to the rightmost structure), in the sense that the latter would be changed and the former would remain unchanged, if we could intervene or manipulate the variable “Down’s Syndrome”. This is because the conditional probabilities for the leftmost model reflect general properties of the relation between Down’s syndrome and tests, and they are the ones that a developer of tests can publish, whereas the conditional probabilities for the rightmost model are a mixture of Down’s syndrome-test relations and prior frequencies of the Down’s syndrome. Due to Bayes’

rule on conditionals, we have

$$P(\text{Down's Syndrome} \mid \text{Indicator}) = \frac{P(\text{Indicator} \mid \text{Down's Syndrome}) P(\text{Down's Syndrome})}{P(\text{Indicator})}.$$

In other words, if a new drug is developed to prevent some pregnant women from Down's Syndrome, i.e., $P(\text{Down's Syndrome})$ is changed, the distribution $P(\text{Down's Syndrome} \mid \text{Indicator})$ will consequently be changed, but the distribution $P(\text{Indicator} \mid \text{Down's Syndrome})$ will remain the same. We intend to utilize such inherent differences in properties of Markov kernels with respect to different structures to make causal inference. This way we will have some simplicity criterion even though the generating model indeed has a fully connected structure.

In summary, the task of causal inference is finding simple models that represent the observed data, in the sense of requiring less links and having stable parameters. The underlying graphical structure of such models provides a good candidate for the representation of causal relationships.

1.3. State-of-the-art causal inference algorithms

As mentioned previously, without restrictive assumptions, a brute force search over all possible structures requires super-exponential time in the number of variables in the model. Over the last years, a large amount of work has been dedicated to formulating reasonable assumptions and feasible search strategies to identify a good causal structure. A detailed discussion of the complexity of causal inference with different conceptions can be found in [53].

In general, two basic search strategies, constraint-based and model-based approaches, are typically employed. The constraint-based approaches mainly focus on the structure of the model, while the model-based approaches take the parameters of the model into account. The model-based approaches often base the search strategy on a Bayesian score. Therefore, it is also called in many literature score-based search. Surely, a non-Bayesian model-based search can be designed.

A Bayesian score-based search assigns a score to each candidate model, characterizing how well that model describes the data, and maximizes this score [86, 59]. Cooper [38] and Chickering et al. [30, 33] showed that given a complete dataset and no hidden variables, locating the Bayesian network structure that has the highest posterior probability is NP-hard, which suggests the use of heuristic strategies for finding close-to-optimum solutions. Particularly for purely discrete networks, various search strategies for models with the maximum score are proposed, e.g., greedy search by Chickering [32], and MCMC by Herskovits [89]. One of the challenges of applying score-based methods is the assessment of informative priors on possible causal models and on parameters for those models. On the one side, the ability to represent prior information is a great advantage of score-based approaches. On the other side, the choice of priors is not trivial. It is currently common to specify some form of non-informative priors on models, e.g., uniform prior over all possible models. Even though non-informative priors typically require only a few parameters to be specified, it is sometimes not obvious how to set them. In addition, there are both theoretical and computational difficulties in calculating scores for models with hidden vari-

1. Introduction and Motivation

ables, if a general framework, i.e., ancestral graphs [133], are used. Finally, it is noteworthy that Cowell [41] showed that if a node ordering, e.g., the causal order, is given the score-based and constraint-based learning for complete data (no missing values in observed data) are equivalent under the assumption of no hidden variables, in the sense that both prefer the same structures as output.

Constraint-based approaches [153, 125] carry out independence tests on the database and build a Bayesian network in agreement with the obtained independence restrictions. They make weak commitments as to the nature of causal relationships. The best-known example of this kind of approaches is the so-called inductive causation (IC) algorithm [125]. The IC algorithm can be broken into an adjacency phase and an orientation phase. The main drawback of IC is that one makes binary decisions about the relations between variables in the adjacency phase when conditional independence is tested. These decisions, which are based on some statistical test, may be erroneous and affect the subsequent behavior of the algorithm, which makes the whole algorithm unstable. Moreover, testing conditional independence, especially in a continuous domain, is a challenging task in its own right. Standard refinements of IC are the PC [152] and FCI algorithms [153]. PC excludes hidden common causes, while FCI allows them. The usual implementation of PC/FCI employs standard statistical tests which are based upon partial correlations (Fisher’s Z) in continuous domains and χ^2 -tests in discrete/categorical domains. The limitations of both tests are obvious: the former relies on the strict assumption of a multivariate Gaussian distribution and the latter leads to a combinatorial explosion of the contingency table, especially if the cardinality of the conditioning set is large. Another shortcoming of such tests is that without discretizing or embedding data, hybrid models, i.e., models of both continuous and discrete/categorical variables, can not be treated by PC/FCI at all.

A first attempt to modify PC by measuring dependences via mutual information is made by Cheng et al. [28], the so-called BN-PC algorithm. Unfortunately, Chickering et al. [34] showed that the “monotone faithfulness assumption” made by BN-PC could not be generally valid. Furthermore, the current implementation of BN-PC can only be applied to purely discrete domains. The essential difficulty is that usual methods for estimation of mutual information from continuous data involve the explicit estimation of the densities, which is hard for high-dimensional data, unless suitable smoothness assumptions are made.

For purely continuous domains, Margaritis [107, 109] proposed a distribution-free independence test for structural learning via constraint-based approaches. His test of independence is Bayesian, because it calculates the exact posterior probability of independence by using Bayesian integration based on a sophisticated iterative procedure of discretization over observed domains.

Apart from the difficulty of testing independence, another weakness of constraint-based approaches is that there are some distinct DAGs that represent exactly the same set of independence relations, the Markov equivalence class. Some interesting empirical results for the size of Markov equivalence classes with up to 10 nodes can be found in [68]. A straightforward consequence is that one cannot determine the causal direction between two dependent variables X and Y if only these two are measured, because hypothetical DAGs $X \rightarrow Y$ and $X \leftarrow Y$ are Markov equivalent to each other. This problem is traced back to the fact that although a cause normally changes the probability of a direct effect when controlling the direct effect’s other causes, such a dependency may be symmetric. Causation, however, is asymmetric even in case of two dependent variables.

Kano et al. [96] have also recognized this problem and proposed a causal inference rule using non-normality via structural equation models. They utilized the fact that linear causal relations can induce non-Gaussian joint measures. The non-Gaussian measure would require nonlinear relations for hypothetical models that are inconsistent with the generating causal order. Based on this observation, their inference algorithm, the so-called LiNGAM algorithm [143], which is based on independent component analysis (ICA) selects models for which linear cause-effect relations are sufficient whenever such causal hypotheses are possible for an observed distribution. However, the underlying idea is only justified for real-valued variables and how to extend their algorithm to other kinds of domains is not straightforward. Note that LiNGAM is an algorithm of non-Bayesian model-based approaches.

1.4. Inductive causation

Since we intend to propose new methods of constraint-based approaches for causal inference, we would like to have a close look at the inference principle of a constraint-based approach, in particular the details of the so-called inductive causation (IC) algorithm.

First, IC makes the Markov assumption. Any population produced by a Bayesian network (output of IC) implies the independence relations entailed by the Markov condition. However, it does not follow that the population induces exactly these and no additional independence relations in the population. The faithfulness [153] (also called stability by Pearl [125]) condition formulates this converse principle.

Definition 7 (Faithfulness Condition) *Let \mathcal{G} be a DAG and P a probability distribution generated by \mathcal{G} . The pair (\mathcal{G}, P) satisfies the faithfulness condition if and only if no conditional independence implied by P holds unless entailed by the Markov condition applied to \mathcal{G} .*

The pair (\mathcal{G}, P) which satisfies both Markov and faithfulness conditions is called a faithful Bayesian network. Because faithful Bayesian networks share the feature of simplicity in graphical structures \mathcal{G} , they are good candidates for causal structures. To see how the faithfulness condition leads to simplicity of structures, we will once again consider the causal relations among the maternal age, Down’s syndrome and various diagnostic tests in the example of having Down’s syndrome. In the case of the fully connected Bayesian network as shown in the rightmost plot of Fig. 1.2, the Markov assumption alone puts no restrictions on the distributions that this structure could produce, because one obtains no independence relations whatsoever from applying the Markov condition or the d-separation criterion to the corresponding DAG. But in some of the distributions that this structure could produce, maternal age might be independent of each diagnostic test “by coincidence”. If the observed distribution induces indeed such independence, this fully connected Bayesian network contradicts the faithfulness assumption and cannot be accepted as a good candidate for the causal model.

Roughly speaking, faithfulness requires that the verified constraints are not implied by accident, but by the structure. If two effects, e.g., a direct influence $X \rightarrow Y$ and an indirect influence via Z , $X \rightarrow Z \rightarrow Y$ happen to exactly balance and thus cancel, then there might be no association

1. Introduction and Motivation

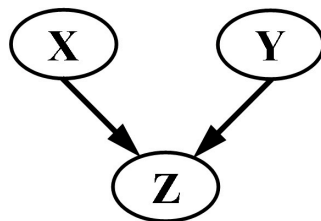


Figure 1.4.: Graphical representation of a v -structure (unshielded collider on Z). As a causal structure, Z is the common effect of mutually independent causes X and Y .

at all between causally connected variables X and Y . In such a case the population is unfaithful to the graph that generated it. In other words, under faithfulness assumption, we can say that whatever structure generated the data, it implies by d-separation exactly the independence relations that are present in the population. Admittedly, it could happen that particular choices of the parameters entail probability distributions which imply additional independence relations not represented in the DAG. However, it can be shown that almost all probability distributions described by Bayesian networks, in a measure-theoretic sense, imply a conditional independence if and only if the DAG represents the corresponding d-separation [113].

If the probability distribution is perfectly known, i.e., without error, and the faithfulness assumption is fulfilled, it can be shown that the constraint-based approaches yield the correct Bayesian network [152, 169], i.e., the Markov equivalence class of DAGs. Note that such a nice property is not guaranteed by the other approaches of optimizing a scoring function, as they can get stuck at local optima. That is the reason why we prefer the constraint-based approach to a score-based search. It should be mentioned that even though the faithfulness condition is not explicitly used by a Bayesian score-based approach, one obtains an implicit preference for faithful structures provided that the priors are strictly positive densities on the space of all conditional distributions [113, 129].

From the algorithmic viewpoint, the Markov and faithfulness assumption leads in some situations to a unique causal structure. Suppose we have a population involving only three distinct variables X , Y , and Z . The only independence relation in this population is $X \perp\!\!\!\perp Y$ (see the first column of Tab. 1.1 for all non-trivial relations). The question is what structure might have produced data with these independence restrictions. The graphs that satisfy the Markov condition share the feature that each has a direct link between X and Z and a direct link between Y and Z , i.e., 6 fully connected DAGs, 3 DAGs with only two arrows as shown in Fig. 1.7 and one DAG as in Fig. 1.4. Adding faithfulness assumption reduces the set of ten to a singleton, i.e., the so-called v -structure as shown in Fig. 1.4. The so-called “explaining away” phenomenon [174] gives a typical example of such v -structures in real-life situations. This phenomenon is also known as Berkson’s paradox, or “selection bias” in statistics. After all, the v -structure is the only structure satisfying the given set of independence constraints, otherwise the Markov or faithfulness condition would be violated.

Having embedded such triples into a larger network, the consideration above leads to an essential strategy for learning structure by v -structure identification. Such kind of structural learning

Independence btw. X, Y and Z	v -Structure	Non- v -Structure
$X \perp\!\!\!\perp Y$	Present in Population	Absent in Population
$X \perp\!\!\!\perp Z$	Absent in Population	Absent in Population
$Y \perp\!\!\!\perp Z$	Absent in Population	Absent in Population
$X \perp\!\!\!\perp Y \mid Z$	Absent in Population	Present in Population
$X \perp\!\!\!\perp Z \mid Y$	Absent in Population	Absent in Population
$Y \perp\!\!\!\perp Z \mid X$	Absent in Population	Absent in Population

Table 1.1.: The configurations of all possible non-trivial independence relations between three distinct nodes X, Y and Z imply the v -structure (Fig. 1.4) and non- v -structure (Fig. 1.7) respectively, when Markov and faithfulness assumptions are made. All structures have the adjacency $X - Z - Y$.

is efficient, because the maximum number of v -configurations is generally less than $\binom{N}{3}$, where N is the number of nodes. For instance, a model with 10 nodes can only have maximum 100 v -configurations (see [68]). The key properties for discovering structures among more than three variables can be summarized as follows:

- Any two distinct variables X and Y are directly connected by an edge (with yet unknown orientation) if and only if there exists no set of variables $S_{XY} \subseteq \mathcal{V} \setminus \{X \cup Y\}$ such that $X \perp\!\!\!\perp Y \mid S_{XY}$. This is due to the faithfulness condition that every edge in the resulting graph induces a dependence that cannot be screened off by conditioning on any subset of variables. An induced dependence can always be screened off if the underlying relation is indirect.
- For subgraphs of the form $X \rightarrow Z \rightarrow Y$, $X \leftarrow Z \rightarrow Y$, and $X \leftarrow Z \leftarrow Y$, where X and Y are nonadjacent, the dependence between X and Y can only be screened off by conditioning on subsets that contain Z .
- For subgraphs of the form $X \rightarrow Z \leftarrow Y$, where X and Y are nonadjacent, conditioning on any subset which contains Z will induce dependence between X and Y .

Based on these thoughts, the IC algorithm conducts three main steps, which are itemized in Fig. 1.5, to learn the causal structure. Fig. 1.6 demonstrates an example, where the underlying true causal model has the structure as shown in Fig. 1.1. After the test of conditional independence (step 1), the underlying skeleton as shown in the leftmost plot of Fig. 1.6 is obtained. In the posterior orientation phase, step 2 provides a partially directed graph whose orientation is only given by the detected v -structures (Fig. 1.6, middle). The remaining edge can be directed (Fig. 1.6, rightmost) in step 3, since the reverse direction would produce new v -structures or directed cycles. Note that the output in this example is not completely directed.

If the distribution P is indeed faithful to some graph \mathcal{G} and we have a perfect way of determining S_{XY} for all pairs (X, Y) , it is then guaranteed that IC produces a graph that is Markov

1. Introduction and Motivation

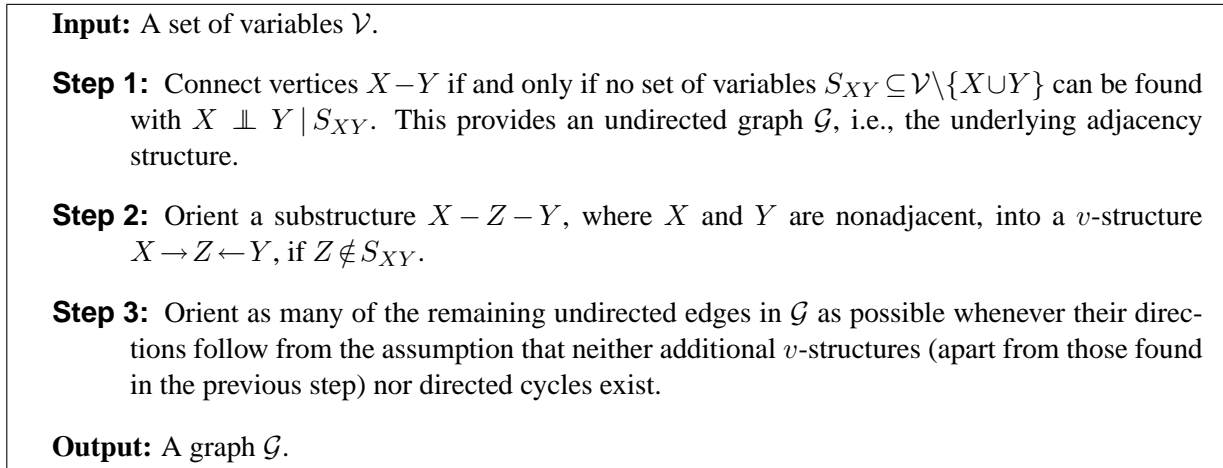


Figure 1.5.: Three-step-scheme of the IC algorithm. Step 1 searches for the underlying adjacency structure. Steps 2 and 3 orient the edges.

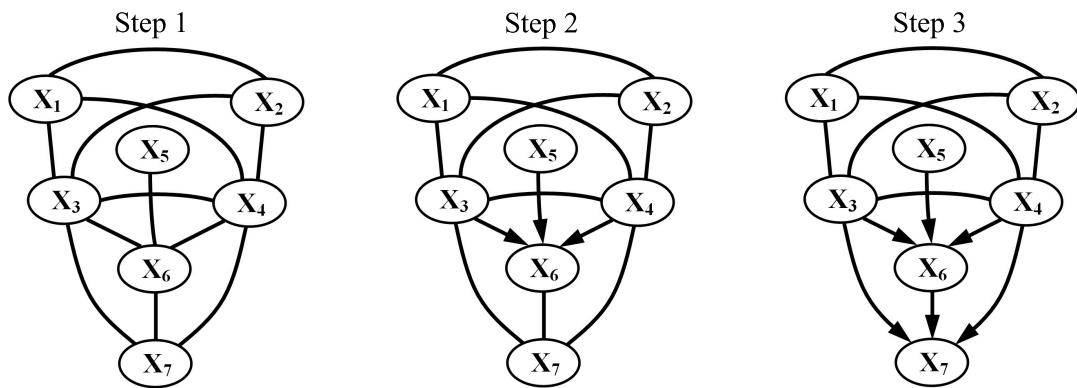


Figure 1.6.: Stepwise results of the IC algorithm for learning the structure as shown in Fig. 1.1. Step 1 learned the adjacency structure. Step 2 identified the v -structures. Step 3 oriented $X_6 \rightarrow X_7$ due to the assumption that no additional v -structures are present. Step 3 further oriented $X_3 \rightarrow X_7$ and $X_4 \rightarrow X_7$ under the assumption that no directed cycles are present. The final output is a partially directed graph.

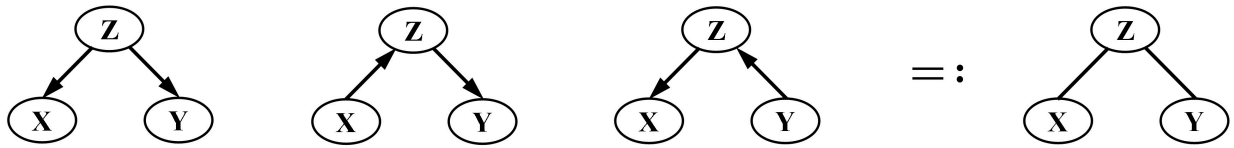


Figure 1.7.: A Markov equivalence class of three different structures. $X \perp\!\!\!\perp Y \mid Z$ is the only independence relation between X , Y and Z . The maximally oriented graph learned by a constraint-based approach, e.g., IC, is an undirected structure, where X and Y are nonadjacent (rightmost plot). It should be stressed that the representation of this undirected structure (rightmost plot) excludes the v -structure (Fig. 1.4), because a constraint-based approach represents such v -structures explicitly. We call the structure as shown in the leftmost plot “fork”, and the structures as shown in the middle two plots “chain”.

equivalent to the original one. Even though IC leaves the details of the tests of conditional independence in step 1 unspecified, it is well-known that a test of independence can fail if it is based on finite sample sizes. In addition, an error made in the early phase of orientation can have cascading effects in the orientation of the output. Due to this instability, the resulting structure is sensitive to the order in which conditional independence relations are checked.

Step 3 of IC can be systematized in several ways. Meek [113] showed that the four rules proposed by Verma et al. [170] are sufficient, so that repeated application will eventually orient all edges that are common to the Markov equivalence class. We call such partially oriented graph a maximally oriented graph. These four rules can be found in [125] p. 51. Note that rule 4 is not required if the starting orientation is limited to v -structures. The first three rules, which are actually needed for step 3 of IC are summarized in Appendix B.3.

The output of IC, a maximally oriented graph, is often not completely directed, e.g., the three structures in Fig. 1.7 are equivalent with respect to the imposed independence constraints (see the second column of Tab. 1.1) and thus indistinguishable by IC. In particular, if no independence can be verified, the usual IC algorithm provides a fully connected and completely undirected graph as output. In particular, IC is not capable of learning direction between two dependent variables, when only these two are measured. In such cases, additional inference rules are desirable.

1.5. Thesis goal and outline

The goal of this thesis is to develop new techniques for recovering the causal relationships from statistical data and demonstrate their utility by applying them to real-world problems. Aside of the contributions of each individual chapter, the thesis introduces mainly two novel techniques for learning causal structures: a so-called kernel dependence measure (Chapter 2 to 5), and an inference rule via properties of Markov kernels (Chapter 6 to 7). The former tool uses the main ideas of a constraint-based approach and goes beyond it (as the magnitude of dependences is used), while the latter is a non-Bayesian model-based approach. The remainder of the thesis

1. Introduction and Motivation

document is structured as follows.

In Chapter 2, we first introduce the so-called kernel dependence measures. Based on the kernel measures, Chapter 3 presents a kernel statistical test of independence. Such a statistical hypothesis test is not only able to capture the conditional independence between continuous variables, without assumption of a specific kind of distribution, but also to deal with hybrid models containing both continuous and discrete/categorical variables in a straightforward way.

By means of the independence constraints obtained by the kernel tests, Chapter 4 elaborates on the problems of inferring the causal structure and presents a so-called robust causal learning (RCL) algorithm. RCL learns the graphical representation of data in favor of constraints of small conditioning sets, under the reliability assumption of these constraints. Since the independence relations are essential for learning directions in the structure, RCL is especially suited to learn sparse models.

Chapter 5 copes with the problem of inferring causal structure using kernel dependence measures. First part of this chapter is spent on introducing an orientation heuristics under some assumption about the magnitude of dependences measured by kernel methods. After that, a fast kernel-based causal learning (KCL) algorithm is presented. KCL uses an auxiliary graph which is obtained by the orientation heuristics to explore the adjacency structure by kernel test of independence. The use of the degree of dependences radically reduces the search space of possible DAGs. The algorithm is particularly suitable for dense models, since the degree of dependences give some hints about the direction of edges without independence relations.

Chapter 6 strives for an inference principle, which goes beyond constraint-based approaches. We aim at the challenging problem how to make causal inference between structures within a Markov equivalence class. In particular, we try to infer causal direction between only two dependent variables. The main idea is to capture the asymmetry between the causal and effect by evaluating the plausibility or complexity of parameters of a causal model, i.e., the set of Markov kernels. The chapter introduces a first concept of plausible Markov kernels (pMK) via low-order (first and second) statistical moments and presents the so-called pMK algorithm to discover the causal order.

Chapter 7 introduces a kernel-based norm to define the complexity of Markov kernels and shows its application to causal inference between only two dependent variables. In the last chapter, the main results of this thesis are summarized and an outlook to future work is given.

2. Kernel Dependence Measure

Conditional independence relationships between variables play a crucial role in constraint-based approaches of structural learning. However, measuring independence or dependence is a non-trivial problem currently unsolved in its generality. In this chapter, we will introduce the so-called kernel dependence measures, which can generally capture both linear and nonlinear relationships.

2.1. Linear and nonlinear dependence

In probability theory and statistics, correlation indicates the magnitude of a linear relationship between two random variables. There are a number of different coefficients for different situations. The most popular is the Pearson correlation coefficient ρ , which is obtained by dividing the covariance of two variables by the product of their standard deviations, namely

$$\rho_{YX} := \frac{\text{Cov}[X, Y]}{\sqrt{\text{Var}[X]\text{Var}[Y]}} = \frac{\text{E}[(X - \text{E}[X])(Y - \text{E}[Y])]}{\sqrt{\text{E}[(X - \text{E}[X])^2]\text{E}[(Y - \text{E}[Y])^2]}} \in [-1, 1], \quad (2.1)$$

where “ $\text{E}[\cdot]$ ”, “ $\text{Var}[\cdot]$ ” and “ $\text{Cov}[\cdot]$ ” depict the expectation, variance and covariance of corresponding random variables.

Suppose variable Y is linearly related to standard normally distributed variable $X \propto \mathcal{N}(0, 1)$ added with a standard normally distributed noise $\epsilon \propto \mathcal{N}(0, \sigma_i^2)$, i.e., $Y = X + \epsilon$. Throughout this thesis, $\mathcal{N}(\mu, \sigma^2)$ denotes a normal (Gaussian) distribution with mean μ and variance σ^2 (standard deviation $\sigma > 0$). Fig. 2.1 illustrates the correlation coefficients between X and Y in the case of noises with different variances. It shows that the larger the variance of the noise, the noisier the relation, the smaller the correlation coefficient ρ . In such linear case, correlation is capable of characterizing the dependence between X and Y .

In general, dependences in nature can be generated by relationships of various forms. In a nonlinear case, the correlation coefficient is not suited to measure the dependence, for instance, samples based on the relation $Y = X^2 + \epsilon$ (see the first plot of Fig. 2.2 for one sample) or $Y = \sin(\pi X) + \epsilon$ (see the first plot of Fig. 2.3 for one sample). Although X and Y are obviously strongly dependent in samples, the correlation coefficient is small. It is noteworthy that the correlation coefficient of transformed data would be large if the nonlinear transform of the original variable were known (see the second, the third and the fourth plot of Fig. 2.2 and Fig. 2.3).

In some situation, the correlation coefficient is not able to capture dependences at all. To make this apparent, we sampled data as shown in the first plot of Fig. 2.4. The original sample (X_0, Y_0) is transformed by a rotation of angle ω (in degree), denoted (X_ω, Y_ω) . The second and

2. Kernel Dependence Measure

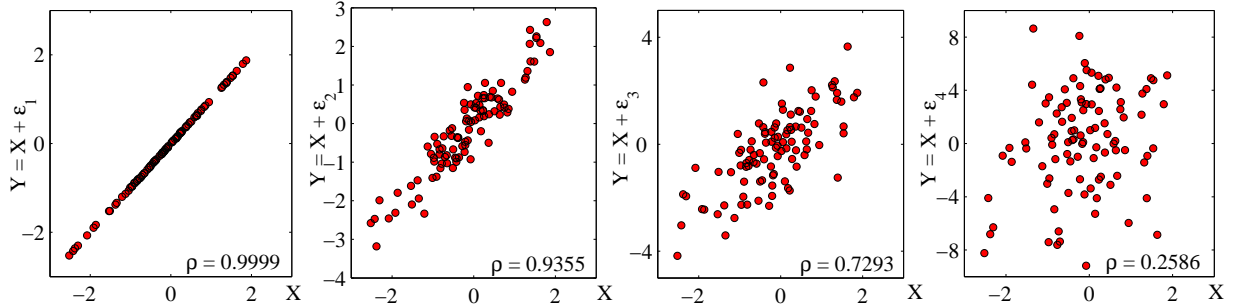


Figure 2.1.: Correlation coefficients indicate the strength of linear relation between two variables X and Y . 100 data points of X are sampled from the standard Gaussian distribution $\mathcal{N}(0, 1)$. Y is linearly related to X , i.e., $Y = X + \epsilon$, with an independent Gaussian noise $\epsilon_i \propto \mathcal{N}(0, \sigma_i^2)$, where $(\sigma_1, \sigma_2, \sigma_3, \sigma_4) = (0.01, 0.5, 1, 4)$ are the standard deviations. The larger the variance σ_i^2 of the noise, the noisier the relation, the smaller the correlation coefficient ρ .

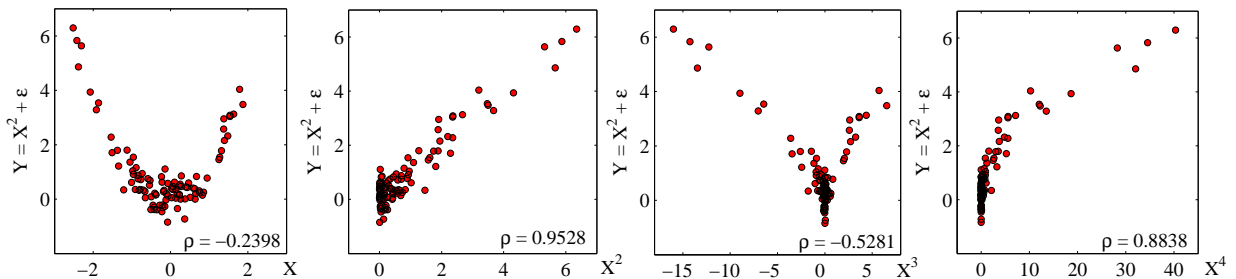


Figure 2.2.: Correlation coefficients are not suitable for measuring the strength of a quadratic relation between two variables. 100 data points of X are sampled from the standard Gaussian distribution $\mathcal{N}(0, 1)$. Y is quadratically related to X , i.e., $Y = X^2 + \epsilon$ with a Gaussian noise $\epsilon \propto \mathcal{N}(0, 0.25)$. The correlation coefficient is small in the original data (leftmost plot). However, if the functional form X^2 of the quadratic relation were known, the correlation coefficient of transformed data (X and X^2) would be large (second plot from left). The correlation coefficient would be also large with a transform of similar polynomial function, e.g., X^3 or X^4 (the last two plots).

2.1. Linear and nonlinear dependence

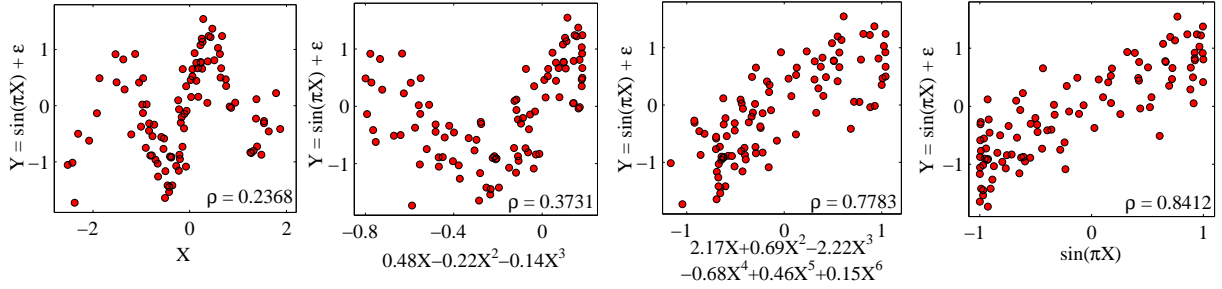


Figure 2.3.: Correlation coefficients are not suitable for measuring the strength of a periodical relation between two variables. 100 data points of X are sampled from the standard Gaussian distribution $\mathcal{N}(0, 1)$. Y and X has a periodical relation, i.e., $Y = \sin(\pi X) + \epsilon$ with a Gaussian noise $\epsilon \propto \mathcal{N}(0, 0.25)$. The correlation coefficient is small in the original data (leftmost plot). However, if the functional form $\sin(\pi X)$ of the periodical relation were known, the correlation coefficient of transformed data (X and $\sin(\pi X)$) would be large (rightmost plot). The correlation coefficient would be also large if an appropriate polynomial function could be found for the transform (the second and third plot from left).

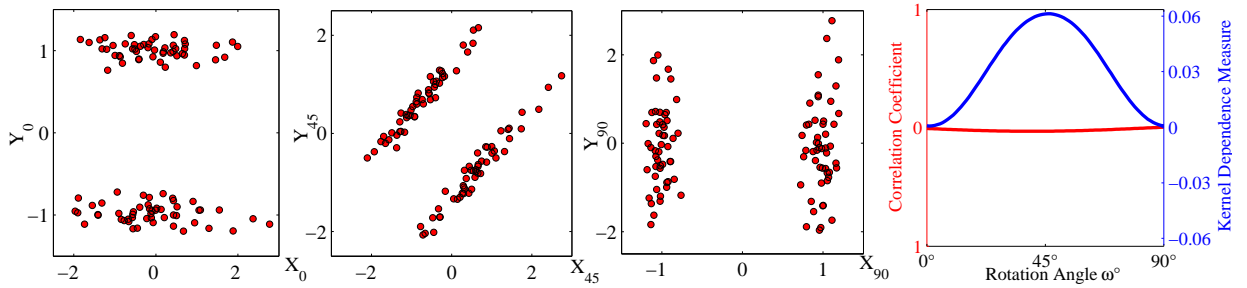


Figure 2.4.: An example of a nonlinear dependence, where the correlation coefficient vanishes, although a strong dependence is present. The half of the total 100 data points of Y is sampled from $\epsilon_1 \propto \mathcal{N}(1, 0.01)$ and the other half from $\epsilon_2 \propto \mathcal{N}(-1, 0.01)$. Data points for X are sampled according to the distributions $P(X|Y < 0), P(X|Y \geq 0) \propto \mathcal{N}(0, 1)$. (X_ω, Y_ω) denotes the original data (X, Y) transformed by a rotation of angle ω in an anticlockwise direction. X_0 and Y_0 (leftmost plot) as well as X_{90} and Y_{90} (second plot from right) are mutually independent, whereas X_{45} and Y_{45} (second plot from left) are strongly dependent. Theoretically, the correlation coefficient vanishes for all ω . The rightmost plot visualizes the typical curve of the estimated correlation coefficient (red line) for $\omega \in [0, 90]$, and a typical curve of the estimated kernel-based dependence measures (blue line), which will be discussed in Section 2.6.

2. Kernel Dependence Measure

the third plot of Fig. 2.4 visualize the transformed data (X_{45}, Y_{45}) and (X_{90}, Y_{90}) . According to the underlying model with $P(X_0|Y_0 < 0) = P(X_0|Y_0 \geq 0)$ (see Fig. 2.4 for description of the generating model), X_0 and Y_0 are independent. It is obvious that X_ω and Y_ω are independent for $\omega = 0, 90, 180, \dots$

In this example, it is easy to see that the correlation matrix ρ_0 of the data matrix

$$\mathcal{D}_0 := \begin{pmatrix} X_0 \\ Y_0 \end{pmatrix}$$

is a unit matrix, namely

$$\rho_0 := \begin{pmatrix} \rho_{X_0X_0} & \rho_{X_0Y_0} \\ \rho_{Y_0X_0} & \rho_{Y_0Y_0} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Further, it is well known that the rotation matrix R_ω of angle ω in an anticlockwise direction has the form of

$$R_\omega = \begin{pmatrix} \cos(\frac{\pi}{180}\omega) & \sin(\frac{\pi}{180}\omega) \\ -\sin(\frac{\pi}{180}\omega) & \cos(\frac{\pi}{180}\omega) \end{pmatrix}.$$

Hence, data transformed by a rotation angle ω can be calculated by

$$\begin{pmatrix} X_\omega \\ Y_\omega \end{pmatrix} =: \mathcal{D}_\omega = R_\omega \mathcal{D}_0 = R_\omega \begin{pmatrix} X_0 \\ Y_0 \end{pmatrix},$$

and the corresponding correlation matrix ρ_ω is given by

$$\rho_\omega = \text{E}[R_\omega (\mathcal{D}_0 - \text{E}[\mathcal{D}_0])(\mathcal{D}_0 - \text{E}[\mathcal{D}_0])^\text{T} R_\omega^\text{T}] = R_\omega \rho_0 R_\omega^\text{T} = \rho_0.$$

This means that the correlation coefficient indeed vanishes for all values of ω , while the dependence actually vanishes only for few specific rotation angles $\omega = 0, 90, 180, \dots$. In this example, correlation coefficient fails to capture the dependence completely. In other words, “uncorrelated” does not mean “independent”. For this reason, it is not very surprising that the performance of solving real-world problems by the PC/FCI algorithm is sometimes unsatisfactory, since it takes only correlation, i.e., the linear dependences into account.

The examples as shown in Fig. 2.3 and Fig. 2.4 suggest that an appropriate nonlinear transform with e.g., polynomials $X \mapsto (X, X^2, X^3, \dots)$ of the original variable might be generally useful for revealing various kinds of dependence. In various applications, however, it is very hard to find the proper parameters. And a well-fitting transform, if possible, may cause high-dimensionality of parameters, since we do not have a direct access to the underlying real relationship. In order to transform the non-linear relationship into a linear one in the feature space, we will employ the so-called kernel method [137] as the general framework.

2.2. Positive definite kernel and RKHS

The idea of “kernelization” is to transform the original data with a “feature map” [3]. The kernel method maps variables into an appropriate feature space by a nonlinear transformation, where the nonlinear dependences in the original space are captured by correlations between variables in the feature space. Hence, the conditional independence in the original space has clear statistical or probabilistic meaning in the feature space. Although the word “kernel” is traditionally used in statistics in a different meaning, which does not impose positive definiteness, e.g., kernel density estimation of Parzen window approach [122], “kernel” means “positive definite kernel” [10] throughout this thesis.

Definition 8 (Positive Definite Kernel) A positive definite kernel on a nonempty set \mathcal{X} is defined by a symmetric function $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ such that for arbitrary $n \in \mathbb{N}$ and $x^{(1)}, \dots, x^{(n)} \in \mathcal{X}$ the matrix K with $(K)_{ij} := k(x^{(i)}, x^{(j)})$ is positive definite, i.e.,

$$\sum_{i,j=1}^n c_i c_j k(x^{(i)}, x^{(j)}) \geq 0$$

for all $c_1, \dots, c_n \in \mathbb{R}$.

A popular positive definite kernel on a subset \mathcal{X} of \mathbb{R}^m is the so-called Gaussian radial basis function (RBF) kernel,

$$k_\sigma(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right), \quad (2.2)$$

with $x, x' \in \mathcal{X}$ and parameter $\sigma \in \mathbb{R}^+$. Every positive definite kernel defines a map Φ from \mathcal{X} into a feature space, i.e., an RKHS \mathcal{H} on \mathcal{X} :

$$\begin{aligned} \Phi : \mathcal{X} &\rightarrow \mathcal{H} \\ x &\mapsto k(x, \cdot). \end{aligned}$$

Here, $\Phi(x)$ denotes the function that assigns the value $k(x, x')$ to $x' \in \mathcal{X}$, i.e., $\Phi(x)(\cdot) = k(x, \cdot)$. Given the inner product $\langle k(x, \cdot), k(x', \cdot) \rangle = k(x, x')$, the feature space \mathcal{H} is defined by the completion of an inner product space spanned by the functions $k(x, \cdot) \in \mathcal{H}$ for all $x \in \mathcal{X}$. Due to the reproducing property

$$\langle k(x, \cdot), f \rangle = f(x) \quad (2.3)$$

for all $f \in \mathcal{H}$, positive definite kernels k are also called reproducing kernels. In view of the map Φ , the reproducing property amounts to

$$\langle \Phi(x), \Phi(x') \rangle = k(x, x').$$

Therefore, the inner product space \mathcal{H} on \mathcal{X} constructed in this way is a possible instantiation of the feature space associated with a kernel k . In some situation, we write $(\mathcal{H}_\mathcal{X}, k_\mathcal{X})$ explicitly to make clear that the RKHS $\mathcal{H}_\mathcal{X}$ on \mathcal{X} is induced by $k_\mathcal{X}$.

2. Kernel Dependence Measure

The main benefit of mapping a random variables $X = (x^{(1)}, \dots, x^{(n)})$ on \mathcal{X} into an RKHS $(\mathcal{H}_{\mathcal{X}}, k_{\mathcal{X}})$, i.e., $X \mapsto \Phi(X)$, is that one can do linear statistics in the feature space $\mathcal{H}_{\mathcal{X}}$. The random variable $\Phi(X)$ (feature representation of random variable X) on the RKHS $\mathcal{H}_{\mathcal{X}}$ is useful to represent the distribution of X .

2.3. Cross-covariance operator and independence

Following the lines of Baker [12], Gualtierotti [80] and Fukumizu et al. [61], we introduce now the cross-covariance operator, expressing correlations between variables in the feature space.

Suppose we have a random vector (X, Y) taking values on $\mathcal{X} \times \mathcal{Y}$. The base spaces \mathcal{X} and \mathcal{Y} are topological spaces. The measurability of spaces is defined with respect to the Borel σ -field. The joint probability measure of (X, Y) is denoted by P_{XY} , and the marginal probability measure by P_X and P_Y . We make the following assumption for a kernel and a random variable throughout this thesis.

Assumption 1 *A positive definite kernel k and a random variable X on measurable space \mathcal{X} satisfy*

$$\mathbb{E}_X [k(X, X)] < \infty.$$

Using the reproducing property of Eq. (2.3), it is easy to see that Assumption 1 guarantees $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Y}}$ are included in $L^2(P_X)$ and $L^2(P_Y)$, respectively, where $L^2(\mu)$ denotes the Hilbert space of square integrable functions with respect to a measure μ (see [62], p. 6).

Definition 9 (Cross-Covariance Operator) *Let $(\mathcal{H}_{\mathcal{X}}, k_{\mathcal{X}})$ and $(\mathcal{H}_{\mathcal{Y}}, k_{\mathcal{Y}})$ be RKHSs of functions on measurable spaces \mathcal{X} and \mathcal{Y} , respectively, which satisfy Assumption 1. It is known that there exists a unique operator Σ_{YX} from $\mathcal{H}_{\mathcal{X}}$ to $\mathcal{H}_{\mathcal{Y}}$ such that*

$$\langle g, \Sigma_{YX} f \rangle_{\mathcal{H}_{\mathcal{Y}}} = \mathbb{E}_{X,Y} [f(X)g(Y)] - \mathbb{E}_X [f(X)] \mathbb{E}_Y [g(Y)] = \text{Cov} [f(X), g(Y)] \quad (2.4)$$

holds for all $f \in \mathcal{H}_{\mathcal{X}}$ and $g \in \mathcal{H}_{\mathcal{Y}}$. It is called the cross-covariance operator.

As the operator Σ_{YX} is a linear map on RKHSs (from $\mathcal{H}_{\mathcal{X}}$ to $\mathcal{H}_{\mathcal{Y}}$), the above definition means that Σ_{YX} works as an analogue to the covariance matrix for finite dimensional random variables. From the definition, it is obvious that $\Sigma_{YX}^* = \Sigma_{XY}$, where Σ^* denotes the adjoint of an operator Σ . If Y is equal to X , the positive self-adjoint operator Σ_{XX} is called the covariance operator. Furthermore, let \mathcal{P}_X and \mathcal{P}_Y be the orthogonal projections which map $\mathcal{H}_{\mathcal{X}}$ onto $\overline{\mathcal{R}(\Sigma_{XX})}$ and $\mathcal{H}_{\mathcal{Y}}$ onto $\overline{\mathcal{R}(\Sigma_{YY})}$, respectively. $\mathcal{R}(\Sigma)$ denotes here the range of an operator Σ . It is known that Σ_{YX} has a representation of the form [12]

$$\Sigma_{YX} = \Sigma_{YY}^{1/2} V_{YX} \Sigma_{XX}^{1/2}, \quad (2.5)$$

where $V_{YX}: \mathcal{H}_{\mathcal{X}} \rightarrow \mathcal{H}_{\mathcal{Y}}$ is the unique bounded operator such that $\|V_{YX}\| \leq 1$ and $V_{YX} = \mathcal{P}_Y V_{YX} \mathcal{P}_X$. $\|\cdot\|$ is used for the operator norm of a bounded operator, i.e., $\|V\| = \sup_{\|f\|=1} \|Vf\|$. To evaluate various nonlinear correlation through the covariance operator, the RKHS should be

2.4. Conditional cross-covariance operator and conditional independence

“rich enough” to express such a variation of nonlinear functions. We make the following assumption for RKHSs throughout this thesis.

Assumption 2 *Let $\mathbf{1}$ denote the function with constant value 1 on \mathcal{X} . Then $\mathcal{H}_{\mathcal{X}} + \mathbb{R} \cdot \mathbf{1}$ is dense in $L^2(P_X)$, where “+” means the sum of RKHSs.*

The kernels that satisfy Assumption 2 are necessarily “characteristic”. The class of characteristic kernels is in general useful for inference on probabilities. Fukumizu et al. [61, 62] used the notation of “probability determining” for this class of kernels. Probability determining kernels mean that the associated RKHS determines a probability by the expectation of $k(x, \cdot)$. We prefer the term “characteristic” because of the analogy with the characteristic function (see [63] for more details).

The $L^2(P_X)$ -space is a rich class of functions including all bounded measurable functions, such as the index function of a measurable set. Thus, under the above assumptions, the following characterization of independence is easy to be proved (see [11], Theorem 2).

Theorem 1 *Under Assumptions 1 and 2, the random variables X and Y are independent if and only if the operator Σ_{YX} vanishes. That is,*

$$\Sigma_{YX} = O \iff X \perp\!\!\!\perp Y.$$

Many popular kernels satisfy Assumption 2. A famous class of such kernels is given by the so-called universal kernels, proposed by Steinwart [159]. A simple criterion for the universality, as well as various examples of universal kernels, are given by Steinwart [159].

Definition 10 (Universal Kernel) *A positive definite kernel $k_{\mathcal{X}}$ on a compact set \mathcal{X} is called universal if the associated RKHS $\mathcal{H}_{\mathcal{X}}$ is dense in the Banach space of bounded continuous functions.*

Since the Banach space of bounded continuous functions on a compact subset \mathcal{X} of \mathbb{R}^m is dense in $L^2(P_X)$ for any probability measure P_X on \mathcal{X} , any universal kernel on a compact subset of \mathbb{R}^m , e.g., the Gaussian RBF kernel and Laplacian kernel, satisfies Assumption 2, and thus can be used to capture independence. Another important example is the Gaussian RBF kernel on the entire Euclidean space. Assumption 2 holds also in this case, as shown by Lemma 4 in Appendix A.1.

In summary, Gaussian RBF kernels can be used to capture the independence between random variables either on a compact subset of \mathbb{R}^m or the entire \mathbb{R}^m . The former fact has been shown by Gretton et al. (see [75], Theorem 6), and the latter by Bach et al. [11] by a direct argument.

2.4. Conditional cross-covariance operator and conditional independence

Following the lines of Fukumizu et al. [61], we define the conditional cross-covariance operator and derive its relation to the conditional independence of random variables.

2. Kernel Dependence Measure

Definition 11 (Conditional Cross-Covariance Operator) Let (\mathcal{H}_X, k_X) , (\mathcal{H}_Y, k_Y) , (\mathcal{H}_Z, k_Z) be RKHSs on measurable spaces \mathcal{X} , \mathcal{Y} , \mathcal{Z} , respectively. And let (X, Y, Z) be a random vector on $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$. The conditional cross-covariance operator of (X, Y) given Z is defined as

$$\Sigma_{YX|Z} := \Sigma_{YX} - \Sigma_{YY}^{1/2} V_{YZ} V_{ZX} \Sigma_{XX}^{1/2}, \quad (2.6)$$

where V_{YZ} and V_{ZX} are the bounded operators in Eq. (2.5) for Σ_{YZ} and Σ_{ZX} , respectively.

If Σ_{ZZ}^{-1} exists, we sometimes rewrite $\Sigma_{YX|Z}$ as

$$\Sigma_{YX|Z} = \Sigma_{YX} - \Sigma_{YZ} \Sigma_{ZZ}^{-1} \Sigma_{ZX} \quad (2.7)$$

for convenience of calculation. Obviously, $\Sigma_{YX|Z}^* = \Sigma_{XY|Z}$. If X and Y coincide, the positive self-adjoint operator $\Sigma_{YY|Z}$ is called the conditional covariance operator. The conditional cross-covariance operator expresses the conditional covariance of $f(X)$ and $g(Y)$ given Z in the feature space, as shown in the following theorem, which generalizes the result on conditional covariance operator (see [62], Proposition 3). The same relation was proved by Fukumizu et al. (see [61], Proposition 5) with more sophisticated assumptions. A simpler proof is given in Appendix A.2.

Theorem 2 Under Assumption 2,

$$\langle g, \Sigma_{YX|Z} f \rangle_{\mathcal{H}_Y} = \mathbb{E}_Z [\text{Cov} [f(X), g(Y) | Z]]$$

for all $f \in \mathcal{H}_X$ and $g \in \mathcal{H}_Y$.

As with the connection between vanishing of the cross-covariance operator and the marginal independence, one would wish for an analogous relationship between vanishing of conditional cross-covariance operator, i.e., $\Sigma_{YX|Z} = O$ and the conditional independence, i.e., $X \perp\!\!\!\perp Y | Z$. Unfortunately, Fukumizu et al. [61] show the equivalence

$$\Sigma_{YX|Z} = O \iff P_{XY} = \mathbb{E}_Z [P_{X|Z} \otimes P_{Y|Z}],$$

which means that the condition $\Sigma_{YX|Z} = O$ is weaker than the conditional independence of X and Y given Z , since

$$X \perp\!\!\!\perp Y | Z \implies P_{XY} = \mathbb{E}_Z [P_{X|Z} \otimes P_{Y|Z}] \not\Rightarrow X \perp\!\!\!\perp Y | Z.$$

Nevertheless, Fukumizu et al. [61] also show that if Z is a part of either X and Y , one obtains the equivalence with the conditional independence. For notational simplicity, we define the shorthands $\check{X} := (X, Z)$ and $\check{Y} := (Y, Z)$. Due to the fact that $X \perp\!\!\!\perp Y | Z$ if and only if $(X, Z) \perp\!\!\!\perp (Y, Z) | Z$ (see [46], Lemma 4.1) we can characterize the conditional independence $X \perp\!\!\!\perp Y | Z$ by $\Sigma_{\check{Y}\check{X}|Z} = O$. To state the result more precisely, we need to introduce a technical assumption on the kernels. To avoid detailed mathematical discussion, we restrict our attention to the following class of spaces for the base spaces.

Assumption 3 The base space of a kernel admits a metric.

2.5. Hilbert-Schmidt dependence measure

The above assumption is satisfied by most of the sets that are used in our context, such as subsets in the Euclidean space and discrete sets, while more general cases may be discussed. Under Assumption 3, it is known (see [51], Lemma 9.3.2) that for a metric space the Banach space of the bounded continuous functions is characteristic (or probability-determining). Since the Banach space is contained in $L^2(P)$ for any probability measure P , it is easy to derive Theorem 3 in the same manner as Theorem 7 in [61]. Recall that for two RKHSs \mathcal{H}_X and \mathcal{H}_Y on \mathcal{X} and \mathcal{Y} , respectively, the tensor product $\mathcal{H}_X \otimes \mathcal{H}_Y$ is the RKHS on $\mathcal{X} \times \mathcal{Y}$ with the positive definite kernel $k_X \otimes k_Y$ (see [10] for details) where

$$(k_X \otimes k_Y)((x, x'), (y, y')) := k_X(x, x') k_Y(y, y').$$

Note that if the two RKHSs both have Gaussian kernels, their direct product is also an RKHS with a Gaussian kernel, which satisfies Assumptions 1 and 2.

Theorem 3 *Let (\mathcal{H}_X, k_X) , (\mathcal{H}_Y, k_Y) , and (\mathcal{H}_Z, k_Z) be RKHSs on measurable spaces \mathcal{X} , \mathcal{Y} , and \mathcal{Z} , respectively, and let (X, Y, Z) be a random vector on $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$. We further define $\ddot{X} := (X, Z)$ and $\ddot{Y} := (Y, Z)$. If kernels $k_X \otimes k_Z$, $k_Y \otimes k_Z$, k_Z , and $(k_X \otimes k_Z) \otimes (k_Y \otimes k_Z)$ satisfy Assumptions 1, 2 and 3, we have*

$$\Sigma_{\ddot{Y}\ddot{X}|Z} = O \iff X \perp\!\!\!\perp Y | Z.$$

Fukumizu et al. (see [61], Corollary 9) have proved $\Sigma_{Y\ddot{X}|Z} = 0 \Leftrightarrow X \perp\!\!\!\perp Y | Z$. Based on this corollary, the proof of Theorem 3 is trivial. Since in general $\Sigma_{\ddot{X}Y|Z} \neq \Sigma_{X\ddot{Y}|Z}$, we prefer a definition which is inherently symmetric with respect to exchanging X and Y .

2.5. Hilbert-Schmidt dependence measure

To derive dependence measures based on the previous results, we need to evaluate how far the operator is from zero. Although there are other choices for measuring the ‘‘size’’ of an operator, such as the largest eigenvalue or determinant (see e.g., [61]), we focus on the Hilbert-Schmidt norm in this thesis. This norm, when applied to the cross-covariance operator, was proposed by Gretton et al. [73] as an independence criterion. For our purpose, we extend it also to the conditional cross-covariance operator.

Definition 12 (HS Norm of Operators) *The Hilbert-Schmidt (HS) norm of $\Sigma: \mathcal{H}_X \rightarrow \mathcal{H}_Y$ is defined, provided that the sum below is finite, by*

$$\|\Sigma\|_{\text{HS}}^2 := \sum_{i,j=1}^{\infty} \langle \varphi_j, \Sigma \phi_i \rangle_{\mathcal{H}_Y}^2,$$

where $\{\phi_i\}_{i=1}^{\infty}$ and $\{\varphi_j\}_{j=1}^{\infty}$ are complete orthonormal systems of separable RKHSs \mathcal{H}_X and \mathcal{H}_Y , respectively.

An RKHS (\mathcal{H}_X, k_X) is separable, when the topological space \mathcal{X} is separable and k_X is continuous on $\mathcal{X} \times \mathcal{X}$ [102]. It is easy to see that this definition generalizes the Frobenius norm on matrices,

2. Kernel Dependence Measure

i.e., the trace of the squared matrix: $\|\Sigma\|_{\text{HS}}^2 = \text{Tr}(\Sigma\Sigma^*)$. The squared HS norms of the cross-covariance and conditional cross-covariance operators will be our dependence measures. This is justified by Theorem 1 and Theorem 3, which imply

$$\begin{aligned} \|\Sigma_{YX}\|_{\text{HS}}^2 = 0 &\iff X \perp\!\!\!\perp Y, \\ \|\Sigma_{\check{Y}\check{X}|Z}\|_{\text{HS}}^2 = 0 &\iff X \perp\!\!\!\perp Y | Z, \end{aligned}$$

under Assumptions 1, 2, and 3.

It is known that the absolute or square value of a correlation coefficient is bounded from above by 1 and indicates the strength of correlation. The sign indicates the manner of correlation. In contrast, the value of a kernel dependence measure is always nonnegative and not bounded from above by 1. The value of the afore-introduced kernel measure eventually depends on the choice of kernels. Fukumizu et al. [63] currently proposed a normalization of the (conditional) cross-covariance operator that makes the afore-introduced dependence measure independent of the choice of kernels. However, its computation requires even in the unconditional case already regularization coefficients, which causes trouble for the empirical estimation in practical applications. Fortunately, various experiments later will show that the issue of kernel choice is not so crucial for our purpose as it seems to be, provided that a normalization factor for the kernel measure is introduced, which makes the measures of unconditional and conditional dependence comparable.

To motivate our normalization factor for the kernel measure, we show by means of graphical models why the comparability of unconditional and conditional dependence is desirable. Suppose a DAG \mathcal{G} including variables X , Y and Z is given. Imagine that the dependence between variables X and Y is partly induced by a direct relation from X to Y and partly by an indirect relation over Z , e.g., via a path $X \rightarrow Z \rightarrow Y$. According to the d-separation criterion (see Definition 4), conditioning on Z blocks the indirect connection from X to Y via Z and changes the dependence between X and Y . When the connections between X and Z and between Y and Z are very weak, the dependence between X and Y is dominated by the direct relation between them. In such situation, one would expect that the conditional (given Z) dependence measure achieve almost the same value as the unconditional one. In particular, the measure of unconditional (given empty set) and conditional (given Z) dependence measure of X and Y should have the same value, if $Z \perp\!\!\!\perp (X, Y)$. Actually mutual information, a popular dependence measure, fulfills this requirement automatically, because $\mathbb{I}(X, Y) = \mathbb{I}(X, Y|Z)$ always holds for $Z \perp\!\!\!\perp (X, Y)$. However, the norms $\|\Sigma_{YX}\|_{\text{HS}}^2$ and $\|\Sigma_{\check{Y}\check{X}|Z}\|_{\text{HS}}^2$ constructed above do not coincide in that case. To make the measure of unconditional and conditional dependence in some sense comparable, we have to renormalize the afore-introduced dependence measure appropriately. For this purpose, we show the following theorem.

Theorem 4 *Let X , Y and Z be random variables with $(X, Y) \perp\!\!\!\perp Z$. Then we have*

$$\Sigma_{\check{Y}\check{X}|Z} = \Sigma_{YX} \otimes T_Z,$$

2.6. Empirical estimation of Hilbert-Schmidt dependence measure

where $T_Z: \mathcal{H}_Z \rightarrow \mathcal{H}_Z$ is defined as

$$\langle h_2, T_Z h_1 \rangle := \mathbb{E}_Z [h_1(Z)h_2(Z)] . \quad (2.8)$$

for arbitrary $h_1, h_2 \in \mathcal{H}_Z$. Hence we obtain

$$\|\Sigma_{\tilde{Y}\tilde{X}|Z}\|_{\text{HS}}^2 = \|T_Z\|_{\text{HS}}^2 \|\Sigma_{YX}\|_{\text{HS}}^2 . \quad (2.9)$$

The proof of this theorem can be found in Appendix A.3. Eq. (2.9) suggests to rescale the dependence measure $\|\Sigma_{\tilde{Y}\tilde{X}|Z}\|_{\text{HS}}^2$ by $1/\|T_Z\|_{\text{HS}}^2$. By means of this rescaling, the conditional dependence measure equals the marginal one, if conditioning variable Z is independent of X and Y .

Definition 13 (Kernel Dependence Measure) *The kernel unconditional (marginal) and conditional dependence measure can be defined by*

$$\begin{aligned} \mathbb{H}_{YX} &:= \|\Sigma_{YX}\|_{\text{HS}}^2 , \\ \mathbb{H}_{YX|Z} &:= \|\Sigma_{\tilde{Y}\tilde{X}|Z}\|_{\text{HS}}^2 / \beta_Z , \end{aligned}$$

respectively, where the scalar $\beta_Z := \|T_Z\|_{\text{HS}}^2 > 0$ makes $\mathbb{H}_{YX|Z}$ and \mathbb{H}_{YX} comparable, in the sense that $(X, Y) \perp\!\!\!\perp Z$ implies $\mathbb{H}_{YX|Z} = \mathbb{H}_{YX}$.

It is straightforward to express the renormalization factor $\beta_Z = \|T_Z\|_{\text{HS}}^2$ in terms of kernels, since the operator T_Z is given by

$$T_Z = \sum_{z \in \mathcal{Z}} k(\cdot, z)k(z, \cdot) P_Z(z) ,$$

which implies

$$\text{Tr}(T_Z^2) = \sum_{z, z' \in \mathcal{Z}} k^2(z, z') P_Z(z)P_Z(z') .$$

For notational convenience, we henceforth drop the double-dots on the variables for the indices of the conditional cross-covariance operators. All conditional cross-covariance operators, e.g., $\Sigma_{YX|Z}$ for measuring conditional independence between X and Y hereafter should be interpreted with the implicit understanding that the conditioning variable Z is a part of both X and Y .

2.6. Empirical estimation of Hilbert-Schmidt dependence measure

Since we need to estimate the dependence measures from a finite number of samples in practical situations, we define the estimators and show the convergence to the population ones in this section.

2. Kernel Dependence Measure

Suppose $(x^{(1)}, y^{(1)}, z^{(1)}), \dots, (x^{(n)}, y^{(n)}, z^{(n)})$ is an independent and identically distributed sample from the joint probability P_{XYZ} . Define $\tilde{k}_{\mathcal{X}}^{(i)} \in \mathcal{H}_{\mathcal{X}}$ by

$$\tilde{k}_{\mathcal{X}}^{(i)} := k_{\mathcal{X}}(\cdot, x^{(i)}) - \frac{1}{n} \sum_{i=1}^n k_{\mathcal{X}}(\cdot, x^{(i)}),$$

and $\tilde{k}_{\mathcal{Y}}^{(i)} \in \mathcal{H}_{\mathcal{Y}}, \tilde{k}_{\mathcal{Z}}^{(i)} \in \mathcal{H}_{\mathcal{Z}}$ analogously.

First we consider an empirical estimator of $\|\Sigma_{XY}\|_{\text{HS}}^2$. By replacing the expectation with the empirical average in Eq. (2.4), the squared norm $\|\Sigma_{XY}\|_{\text{HS}}^2$ is approximated by

$$\begin{aligned} \left\| \widehat{\Sigma}_{YX}^{(n)} \right\|_{\text{HS}}^2 &= \sum_{l,m=1}^{\infty} \left\langle \varphi_m, \widehat{\Sigma}_{YX}^{(n)} \phi_l \right\rangle_{\mathcal{H}_{\mathcal{Y}}}^2 \\ &= \frac{1}{(n-1)^2} \sum_{l,m=1}^{\infty} \sum_{i,j=1}^n \left\langle \tilde{k}_{\mathcal{X}}^{(i)}, \varphi_m \right\rangle \left\langle \tilde{k}_{\mathcal{Y}}^{(i)}, \phi_l \right\rangle \left\langle \tilde{k}_{\mathcal{X}}^{(j)}, \varphi_m \right\rangle \left\langle \tilde{k}_{\mathcal{Y}}^{(j)}, \phi_l \right\rangle, \end{aligned}$$

where $\{\phi_l\}_{l=1}^{\infty}$ and $\{\varphi_m\}_{m=1}^{\infty}$ are complete orthonormal systems of $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Y}}$, respectively. Let \widehat{K} be the centralized kernel matrix (see e.g., [138]) defined as

$$\widehat{K}_X := \left(I_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^{\text{T}} \right) K_X \left(I_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^{\text{T}} \right)$$

where $(K_X)_{ij} = k_{\mathcal{X}}(x^{(i)}, x^{(j)})$ is the kernel matrix, and $\mathbf{1}_n = (1, \dots, 1)^{\text{T}}$ is the vector with all entries equal to 1. The matrix \widehat{K}_Y is defined analogously by using $y^{(i)}, y^{(j)}$. Then, it is easy to see that $\langle \tilde{k}_{\mathcal{X}}^{(i)}, \tilde{k}_{\mathcal{X}}^{(j)} \rangle = (\widehat{K}_X)_{ij}$ and $\langle \tilde{k}_{\mathcal{Y}}^{(i)}, \tilde{k}_{\mathcal{Y}}^{(j)} \rangle = (\widehat{K}_Y)_{ij}$. We have

$$\left\| \widehat{\Sigma}_{YX}^{(n)} \right\|_{\text{HS}}^2 = \frac{1}{(n-1)^2} \sum_{i,j=1}^n (\widehat{K}_Y)_{ji} (\widehat{K}_X)_{ij} = \frac{1}{(n-1)^2} \text{Tr} \left(\widehat{K}_Y \widehat{K}_X \right).$$

$\widehat{K}_X, \widehat{K}_Y$ are matrices of inner products between centered observations in respective feature spaces. The trace of their product can, in some sense, be interpreted as a measure of similarity between two kernel matrices \widehat{K}_X and \widehat{K}_Y measured by Frobenius inner product. It is easy to see that the inner product is always nonnegative, due to the fact that kernel matrices are positive definite. Another possible interpretation of the HS-norm for a cross-covariance operator is the following. According to [24], one can represent P_{XY} and $P_X P_Y$ as Hilbert space vectors in feature space, then the kernel maximum mean discrepancy, i.e., distance of the mean elements of P_{XY} and $P_X P_Y$, in feature space (see [24] for details). An empirical estimator of the HS-norm for a cross-covariance operator, which is also called the Hilbert-Schmidt Independence Criterion, or HSIC by Gretton et al. [73], can be defined as follows.

Definition 14 (Empirical Unconditional Dependence Measure) *An empirical estimate of the*

2.6. Empirical estimation of Hilbert-Schmidt dependence measure

Hilbert-Schmidt unconditional dependence measure is

$$\widehat{\mathbb{H}}_{YX}^{(n)} := \frac{1}{(n-1)^2} \text{Tr} \left(\widehat{K}_Y \widehat{K}_X \right). \quad (2.10)$$

Note that the normalization factor proposed in Section 2.5 is only used for the conditional dependence measure. The unconditional dependence measure will not be rescaled. As mentioned previously, the absolute value of the empirical measures as defined above actually depends on the choice of kernels and is only bounded from below by 0 but not bounded from above. It is not clear how to interpret the value of empirical dependence measures. Nevertheless, because of the smoothness assumption implicitly made by kernels, this kernel measure captures the dependence between two variables in a reasonable way.

As an example, we can calculate the empirical unconditional dependence measure for the sample as shown in Fig. 2.4 by using Gaussian kernels. The rightmost plot shows that the estimators are always larger than zero for all rotation angles ω , and the empirical measure is nearly zero when $\omega = 0$ or $\omega = 90$. The empirical kernel dependence measure reaches its maximum when $\omega = 45$. This behavior reflects the intuitive understanding of the underlying dependence. The degree of dependence achieves its maximum when $\omega = 45$.

The next step is to show how to estimate the HS-norm of a conditional cross-covariance operator, let $\widehat{\Sigma}_{YX}^{(n)}$, $\widehat{\Sigma}_{YZ}^{(n)}$, $\widehat{\Sigma}_{ZX}^{(n)}$, $\widehat{\Sigma}_{ZZ}^{(n)}$ denote the empirical estimators corresponding to the respective operators. Based on Eq. (2.7), the empirical conditional cross-covariance operator $\widehat{\Sigma}_{YX|Z}^{(n,\epsilon)}$ is defined as

$$\widehat{\Sigma}_{YX|Z}^{(n,\epsilon)} = \widehat{\Sigma}_{YX}^{(n)} - \widehat{\Sigma}_{YZ}^{(n)} \left(\widehat{\Sigma}_{ZZ}^{(n)} + \epsilon I \right)^{-1} \widehat{\Sigma}_{ZX}^{(n)}, \quad (2.11)$$

where $\epsilon > 0$ is a regularization constant that enables inversion.¹ It is analogous to Tikhonov regularization [77] or ridge regression [92]. The most natural unbiased estimator $\widehat{\beta}_Z^{(n)}$ for β_Z is given by a U-statistic

$$\widehat{\beta}_Z^{(n)} := \frac{n(n-1)}{\sum_{i \neq j} k_Z(z^{(i)}, z^{(j)})^2}$$

with $z^{(i)}, z^{(j)} \in \mathcal{Z}$. Henceforth, the corresponding estimator for the conditional dependence measure can be defined as follows.

Definition 15 (Empirical Conditional Dependence Measure) *An empirical estimate of the Hilbert-Schmidt conditional dependence measure is*

$$\begin{aligned} \widehat{\mathbb{H}}_{YX|Z}^{(n,\epsilon)} := & \frac{\widehat{\beta}_Z^{(n)}}{(n-1)^2} \text{Tr} \left(\widehat{K}_Y \widehat{K}_X - 2\widehat{K}_Y \widehat{K}_Z (\widehat{K}_Z + \epsilon I)^{-2} \widehat{K}_Z \widehat{K}_X \right. \\ & \left. + \widehat{K}_Y \widehat{K}_Z (\widehat{K}_Z + \epsilon I)^{-2} \widehat{K}_Z \widehat{K}_X \widehat{K}_Z (\widehat{K}_Z + \epsilon I)^{-2} \widehat{K}_Z \right). \end{aligned} \quad (2.12)$$

¹The regularizer is required as the observed data are finite, whereas the feature space could be infinite-dimensional. The regularization may be understood as a smoothness assumption on the eigenfunctions of \mathcal{H}_Z . Our experiments in Section 5.4.4 will give some numerical evidence that the empirical measures are insensitive to ϵ , if it is chosen in the interval $[10^{-10}, 10^{-2}]$. In our experiments, we always chose 10^{-5} .

2. Kernel Dependence Measure

The estimators $\widehat{\mathbb{H}}_{YX}^{(n)}$ and $\widehat{\mathbb{H}}_{YX|Z}^{(n,\epsilon)}$ are justified by the following two results on their statistical consistency.

Theorem 5 (Fukumizu et al. [60])

$$\widehat{\mathbb{H}}_{YX}^{(n)} - \mathbb{H}_{YX} = O_p(n^{-1/2}) \quad (n \rightarrow \infty).$$

The notation $O_p(n^{-1/2})$ means the convergence in probability (see [166] for more details) at rate $n^{-1/2}$, which means for all $\epsilon > 0$ there exists $c > 0$ such that

$$P\left(n^{1/2} \left| \widehat{\mathbb{H}}_{YX}^{(n)} - \mathbb{H}_{YX} \right| > c\right) < \epsilon$$

as n is sufficiently large.

Theorem 6 *If the regularization parameter ϵ in Eq. (2.12) satisfies*

$$\epsilon \rightarrow 0, \quad \epsilon n^{1/2} \rightarrow \infty \quad (n \rightarrow \infty),$$

then we have

$$\widehat{\mathbb{H}}_{YX|Z}^{(n,\epsilon)} - \mathbb{H}_{YX|Z} \xrightarrow{P} 0 \quad (n \rightarrow \infty).$$

The letter P over the arrow indicates a convergence in probability. The proof of Theorem 6 is given in Appendix A.4. The above theorem shows that if ϵ tends to zero sufficiently slowly, the empirical estimator $\widehat{\mathbb{H}}_{YX|Z}^{(n,\epsilon)}$ converges to $\mathbb{H}_{YX|Z}$. For notational convenience, we will henceforth omit the upper index of the empirical estimators and use $\widehat{\mathbb{H}}_{YX}$ and $\widehat{\mathbb{H}}_{YX|Z}$ to denote the empirical estimators of \mathbb{H}_{YX} and $\mathbb{H}_{YX|Z}$, respectively.

2.7. Computation of empirical Hilbert-Schmidt dependence measure

Kernel matrices allow us to capture dependence in a non-parametric setting. On the other hand, working with kernel matrices of n data points implies not only the storage of n^2 entries, but also the $O(n^3)$ complexity of matrix multiplication and inversion. A naive implementation would require $O(n^3)$ operations. If the sample size n is large, the computation will be inefficient.

Fortunately, the problem of time and memory requirements of large matrices is not a new one. Various methods have been developed to alleviate it. If the number of possible values of a variable X is smaller than the number of data points n , the rank of the kernel matrix \widehat{K}_X is smaller than n . Even when this is not the case, we may still have good low-rank approximations. For a positive definite matrix \widehat{K} , e.g., kernel matrices \widehat{K}_X , \widehat{K}_Y or \widehat{K}_Z in Eq. (2.10) or Eq. (2.12), we can use an incomplete version of the Cholesky decomposition $\widehat{K} = LL^T$ [55] where L is a lower triangular matrix determined uniquely by this equation. This may lead to considerably fewer columns than the original matrix. If k columns are returned, the storage requirements are

2.7. Computation of empirical Hilbert-Schmidt dependence measure

$O(kn)$ instead of $O(n^2)$, and the running time of many matrix operations reduces to $O(nk^2)$ instead of $O(n^3)$.

The other question is the choice of kernels. Actually, the choice of kernels $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ specifies the sets of functions that we use for characterizing the dependence, via the correlation between $f \in \mathcal{H}_{\mathcal{X}}$ and $g \in \mathcal{H}_{\mathcal{Y}}$. While in general we can apply different kernels to \mathcal{X} and \mathcal{Y} , for simplicity we restrict ourselves in this thesis to the case where the two kernels are the same. Certainly, the absolute values of both (marginal and conditional) dependence measures strongly depend on the choice of kernels. However, every statistical measure of dependence between continuous variables relies on explicit or implicit assumptions on properties of the probability distributions. We believe that one of the most natural assumptions which still leads to a feasible dependence measure is that dependence between continuous variables X, Y should be considered greater when correlations arise between smooth functions $f(x), g(y)$, but lesser when these correlations are only seen for non-smooth $f(x)$ and $g(y)$. That kernel dependence measures embody this assumption can be seen from the discussion in Section 4.2 in [76], which we now summarize. Let $\phi_i(x)$ be the i -th eigenfunction of the integral operator with kernel $k_{\mathcal{X}}(x, x')$, and let $\varphi_j(y)$ be the j -th eigenfunction corresponding to $k_{\mathcal{Y}}(y, y')$. We consider the case where $\text{Cov}(\phi_l(X), \varphi_l(Y))$ is large for some value of l , and small otherwise. In this case, subject to a mild condition on the kernel spectrum (see Lemma 4 in [76]. This is satisfied for Gaussian kernels, for instance), the spectral norm of the covariance operator decreases for increasing l . Since the spectral norm is the largest singular value of the covariance operator, it follows that the HS norm likewise decreases. In other words, as the nonlinear mapping required to obtain a high covariance becomes more “complex”, the dependence as measured by kernels decreases (see [76] for more details, and for a proof of the result).

If we choose identical Gaussian kernels (our default universal kernel) for each variable, the computation has two free parameters: the regularization parameter ϵ for the conditional dependence measure of Eq. (2.11) and Eq. (2.12), as well as the width σ of the kernel in Eq. (2.2). To see how the kernel width σ influences the value of the dependence measure, we consider the Fourier transform of an isotropic Gaussian kernel $\nu(\omega) = (\pi\sigma^2) \exp(-\|\omega\|^2/\sigma^2)$. The feature space \mathcal{F}_{σ} contains functions whose Fourier transform decays very rapidly. In the case of a too large σ^2 , all entries of kernel matrices are almost the same. Smaller σ^2 means greater sensitivity to dependence, although making σ^2 too small causes sensitivity to drop again, because an overly small σ^2 leads to diagonal kernel matrices and our criteria become trivial. From the computational point of view, the smaller σ^2 , the more complexity, since the kernel matrices contain more non-negligible eigenvalues. Admittedly, we have no principled way of choosing ϵ and σ^2 . In our experiments, unless otherwise noted, we used the regularizer $\epsilon = 10^{-5}$. In our experimental work, it turned out that the evaluation of dependence criteria that we used for structural learning was reasonably robust, if ϵ is chosen sufficient small (see Fig. 5.10 in Section 5.4.4). We set $2\sigma^2 = 1$ in Eq. (2.2), since all variables are independently rescaled to have unit variance in pre-processing.

3. Kernel Statistical Test of Independence

Causal inference by a constraint-based approach naturally includes inferring whether a causal relation between two variables is present or not. This involves the choice of a cut-off value for some kind of dependence measures, e.g., our kernel measures, which judges whether conditional dependence between them is present or not. However, a straightforward threshold does not work well in general, because the value of kernel dependence measures depends on the choice of kernels and could theoretically be small even under dependency (see Theorem 8 in [76]). In this chapter, we introduce statistical hypothesis tests to set the cut-off value in a principled way.

3.1. State-of-the-art tests of independence

Given some dependence measure, one wishes to make a decision whether two variables are dependent or not. A principled way of deciding whether a hypothesis is true or not is the statistical test. In such tests, there is a “null hypothesis” which corresponds to the state of independence and an “alternative hypothesis” which corresponds to the opposite situation, i.e., state of dependence. The goal is to determine, with high confidence, if the null hypothesis can be discarded in favor of the alternative. The result of an independence hypothesis test may be negative, i.e., independent, or positive, i.e., dependent.

If the null hypothesis, i.e., state of independence, is the truth, the dependence measure from sample should follow the null distribution, which can be simulated by random permutations (see [71] for permutation tests). If the alternative is true, the dependence measure will be, in the genetic case, “large”. To specify a “large” measure, a threshold with risk α (the so-called significance level, usually $\alpha = 5\%$) is pre-specified.

The p-value is the probability that the sample could have been drawn from the population being tested given the assumption that the null hypothesis is true. A p-value of 0.02, for example, indicates that one would have only a 2% chance of drawing the sample being tested if the null hypothesis was actually true. The further out the test statistic is in the tail, the smaller the p-value, and the stronger the evidence against the null hypothesis in favor of the alternative. If the p-value is larger than the significance level, the null hypothesis is accepted, otherwise the null hypothesis is rejected in favor of the alternative (see Figure 3.1).

Since the decision is made based on one sample, we can not be completely certain. If the result of the hypothesis test does not coincide with the ground truth, which might not be known, then an error has occurred. Statisticians speak of two sorts of statistical errors, classified as the type I and II error. The type I error, also known as a “false positive”: the error of rejecting a

3.1. State-of-the-art tests of independence

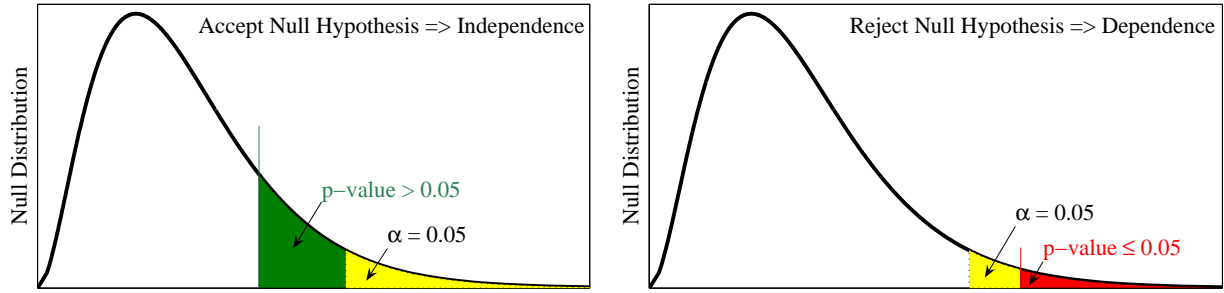


Figure 3.1.: One-sided statistical independence hypothesis test with a significance level $\alpha = 0.05$. If the p-value is larger than 0.05, null hypothesis, i.e., state of independence, is accepted, otherwise independence hypothesis is rejected.

	Declared Non-Significant (Accept Independence)	Declared Significant (Reject Independence \Rightarrow Dependence)
True Null Hypothesis (State of Independence)	True Negative	False Positive (Type I error)
True Alternative Hypothesis (State of Dependence)	False Negative (Type II error)	True Positive

Table 3.1.: Type I and type II error of independence hypothesis test.

null hypothesis when it is actually true. This is the error of rejecting independence although independence is true. The type II error, also known as a “false negative”: the error of accepting a null hypothesis when the alternative hypothesis coincides with the ground truth. This is the error of accepting independence when dependence is present. Tab. 3.1 summarizes the situation in a traditional form.

Based on the same fundamental concept of statistical tests, well-established statistical tests of independence vary in the way of capturing dependences, i.e., the test statistics. The popular χ^2 test is based on the contingency table for discrete/categorical domains. The Fisher’s Z test [56] is based on partial correlations and therefore only justified for continuous domains under the assumption that the variables are multivariate Gaussian distributed. Mutual information, which is based on the entropy concept of Shannon [142], can be generally considered as a distribution-free dependence measure. It can be shown (see [178], Appendix A) that mutual information is proportional to the χ^2 test based on maximum likelihood estimation, the so-called likelihood ratio χ^2 test. For this reason, a likelihood ratio χ^2 test and a permutation test by means of mutual information are expected to be similar in performance. The empirical estimation of (conditional) mutual information on discrete/categorical domains is well established, while the estimation of (conditional) mutual information on continuous domains is a non-trivial problem currently unsolved in its generality, unless suitable assumptions of smoothness are made. We are of the opinion that kernel methods provide a convenient tool to assume smoothness in an implicit way.

3. Kernel Statistical Test of Independence

The extension of kernel dependence measures to high-dimensional data is straightforward.

A totally different method of testing independence on continuous domains is proposed by Margaritis et al. [107, 109]. Their method is not based on a conventional hypothesis test but on the calculation of probability of independence given data by the Bayesian approach. To determine whether two variables are (conditionally) independent, the Margaritis' Bayesian method discretized the domains by maximizing the posterior probability of dependence given the data. If the probability of independence larger than $\frac{1}{2}$, the independence is verified, otherwise dependence. More precisely, the method determines the probability of dependence by calculating the likelihoods of modeling the data as dependent with a joint multinomial distribution or as independent with two marginal multinomial distribution. Margaritis' Bayesian method is impressive because it is the first practicable distribution-free learning of Bayesian network in continuous domains, although it involves a sophisticated process of domain discretization.

Note that χ^2 test, Fisher's Z test and Margaritis' Bayesian method share the property of good scalability with respect to sample size. They remain efficient, when the sample size becomes large. Unfortunately, work with kernel matrices of a large number of data points, which is required for the computation of empirical kernel measures, will be inefficient. However, the power of kernel measures is the ability of capturing linear and non-linear relations, without requiring the specification of any kind of dependence model. Moreover, kernel measures can be applied to discrete/categorical, continuous, vectorial, or even hybrid domains. For discrete domains, one can use integers $1, 2, \dots, d$ to specify d different categories, if the categories can be ordered in some intuitive sense. For strictly nominal-categorical domains, the natural way to represent the d nominal alternatives, namely d unit vectors in a d -dimensional Cartesian coordinate system

$$\{(1, 0, \dots, 0)^T, (0, 1, 0, \dots, 0)^T, \dots, (0, 0, \dots, 0, 1)^T\} \subset \mathbb{R}^d. \quad (3.1)$$

Note that, in the binary case, the representations of integers $\{0, 1\}$ or two-dimensional vectors $\{(0, 1), (1, 0)\}$ does not makes any difference at all.

3.2. Statistical test via kernel dependence measure

To design a statistical test of independence via kernel measure, we need the statistics of the dependence measure if the null hypothesis is true, i.e., the null distribution of \mathbb{H}_{YX} or $\mathbb{H}_{YX|Z}$. For this purpose, we employ the random permutation to simulate the state of independence.

Let us first consider the marginal case \mathbb{H}_{YX} . To simulate the null distribution of \mathbb{H}_{YX} under independency, we apply a set of random permutations $\pi := \{\pi_1, \dots, \pi_m\}$ to the X - or the Y -vector of the original data matrix (X, Y) , where $X = (x^{(1)}, \dots, x^{(n)})^T$ and so on. The marginal distribution $P(X)$ or $P(Y)$ of the original data does not change in the shuffled data (X, Y^{π_j}) with $Y^{\pi_j} = (y^{(\pi_j(1))}, \dots, y^{(\pi_j(n))})^T$ and $\pi_j \in \pi$. However, the relation between X and Y in the original data is released. For each shuffled dataset (X, Y^{π_j}) , we compute the empirical estimate of the kernel dependence measure $\widehat{\mathbb{H}}_j$ with $j = 1, \dots, m$. The null distribution of measure \mathbb{H}_{YX} can be simulated by $\{\widehat{\mathbb{H}}_1, \dots, \widehat{\mathbb{H}}_m\}$.

The way of using random permutations to simulate the null distribution of $\mathbb{H}_{YX|Z}$ under con-

3.2. Statistical test via kernel dependence measure

ditional independency from data matrix (X, Y, Z) is not straightforward. On the one side, the random permutation should release the connection between X and Y to simulate the independency between them. On the other side, it has to keep the mutual relation between X and Z and the relation between Y and Z , since Z is tied to a specific value. Applying a random permutation $\pi_j \in \pi$ to the two-dimensional (Y, Z) -vector of the original data matrix (X, Y, Z) , the conditional marginal distribution $P(Y|Z)$ of shuffled dataset $(X, Y^{\pi_j}, Z^{\pi_j})$ remains indeed the same as that of the original dataset (X, Y, Z) , since $P(y^{(i)}|z^{(i)}) = P(y^{(\pi_j(i))}|z^{(\pi_j(i))})$. But, the conditional marginal distribution $P(X|Z)$ changes, because the conditional probability $P(x^{(i)}|z^{(i)})$, in general, does not equal $P(x^{(i)}|z^{(\pi_j(i))})$. In particular, the conditional joint probabilities of $(X, Y^{\pi_j} | Z^{\pi_j})$ and $(X, Y^{\pi_j} | Z)$ are different:

$$P(x^{(i)}, y^{(\pi_j(i))} | z^{(\pi_j(i))}) \neq P(x^{(i)}, y^{(\pi_j(i))} | z^{(i)}) .$$

The only exception is the case when

$$z^{(\pi_j(1))} = z^{(1)}, z^{(\pi_j(2))} = z^{(2)}, \dots, z^{(\pi_j(n))} = z^{(n)} . \quad (3.2)$$

Therefore, we have to restrict the set of random permutations π to those that satisfy the condition of Eq. (3.2) to simulate the null distribution of $\mathbb{H}_{YX|Z}$ under the conditional independency. A related observation in the context of the conditional copula is made by Patton (see [123], p. 534).

If Z is categorical, the condition of Eq. (3.2) restricts π to random permutations within the same category of Z . In the case of a real-valued Z , the condition of Eq. (3.2) could be said to hold if $z^{(\pi_j(i))}$ and $z^{(i)}$ are “similar” in some sense. This suggests the use of clustering techniques to search for an appropriate partition of data points of Z .

In our experiments, we applied the standard K-means clustering to n data points $\{z^{(1)}, \dots, z^{(n)}\}$ and chose the number of clusters n_c so that $\frac{n}{n_c} = 3$. Various experiments showed that the decision of independence is robust with respect to the choice of n_c , if n_c is not too large, i.e., $\frac{n}{n_c} > 2$. If the number of clusters n_c is chosen as large as the number of data points n , every distinct data point builds a separate cluster and the condition of Eq. (3.2) restricts π to the identity. The null distribution of the dependence measures is degenerate, i.e., all its probability mass is concentrated on one point. With such a choice of n_c , permutations can not provide any information about the distribution of dependence measure under conditional independency.

Having chosen an appropriate parameter n_c (number of clusters of data points of Z), the null distribution of the conditional kernel measure $\widehat{\mathbb{H}}_{YX|Z}$ under conditional independency $X \perp\!\!\!\perp Y | Z$ can be simulated by applying a set of random permutations $\pi := \{\pi_1, \dots, \pi_m\}$ to the two-dimensional (Y, Z) -vector of data matrix (X, Y, Z) within the same cluster of data points of Z . Fig. 3.2 summarizes the three-step-schema of the hypothesis test by means of kernel dependence measures. The parameter m describes the number of permutations to simulate the null distribution. We chose $m = 1000$ in our experiments and set the significance level α to 0.05, unless explicitly stated otherwise.

Obviously, permutation tests are computationally time-consuming due to the m replications. An alternative kernel statistical test based on moment matching is currently proposed by Gretton et al. [74]. Instead of computing the HS-norm of the operator directly, they designed a test statis-

3. Kernel Statistical Test of Independence

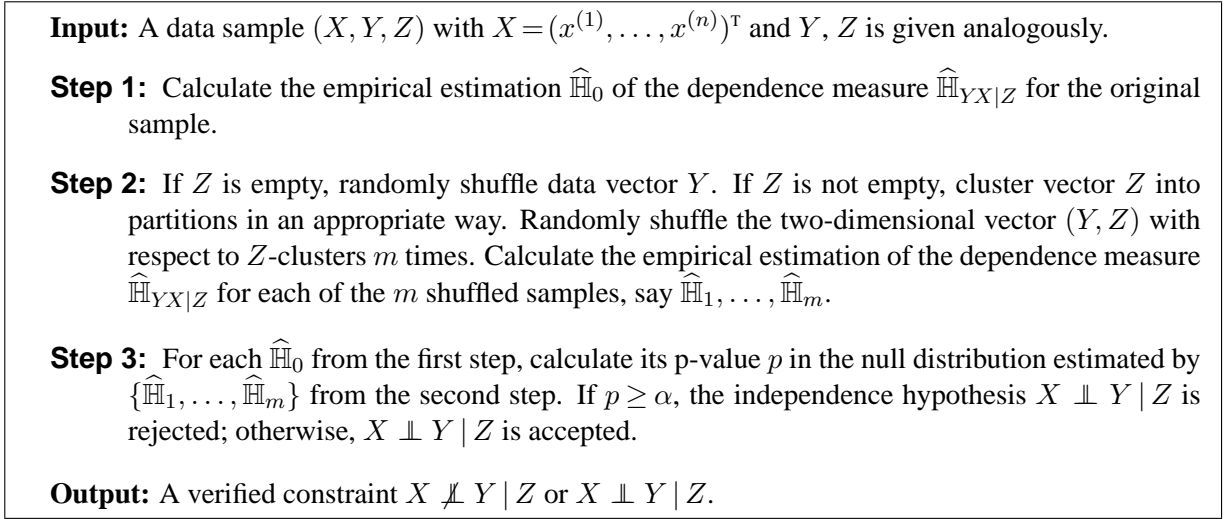


Figure 3.2.: Permutation test of (conditional) independence via kernel dependence measures. Typically, we set $m = 1000$ and $\alpha = 0.05$, unless explicitly stated otherwise.

tics based on entries of kernel matrices. But, for one thing, this alternative test is only designed for unconditional cases. For the other thing, we expect that the permutation test outperforms this alternative, particularly if the sample size is small (e.g., less than 200), since the estimation of second moments of entries of kernel matrices tends to be unreliable (see [74] for numerical experiments with text data). In practice, employing the incomplete Cholesky decomposition [55] for the computation of kernel measures makes the permutation tests efficient (see Section 2.7 for details).

3.3. Simulated experiments with kernel independence test

It is known that there is yet no general good way to test independence, especially between continuous variables. Theoretically, the kernel dependence measure can capture both linear and nonlinear dependences without assumptions of a specific dependence model. Therefore, the statistical test by means of the kernel measure provides a useful tool for handling the challenging task of testing independence. In this section, we demonstrate some simulated experiments with the kernel independence test, in particular some examples on continuous domains.

3.3. Simulated experiments with kernel independence test

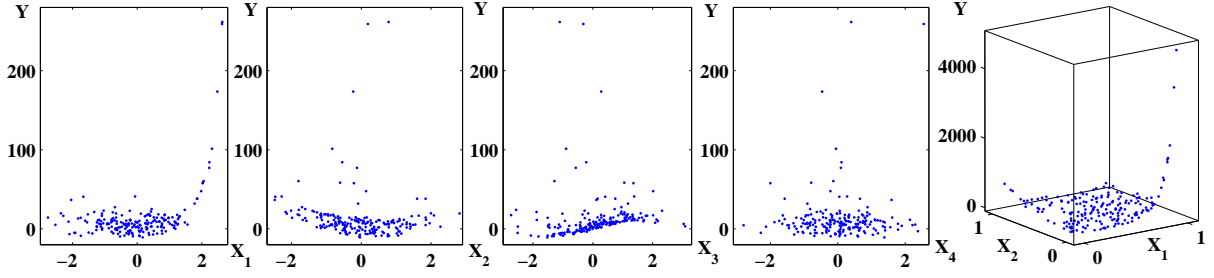


Figure 3.3.: One dataset of (X_1, X_2, X_3, X_4, Y) is sampled (of size 200) from the functional model, defined by Eq. (3.3). The first four plots visualize the relationship between X_i ($i = 1, \dots, 4$) and Y , respectively. One dataset of (X_1, X_2, Y) is sampled (of size 200) from the functional model, defined by Eq. (3.4). The rightmost plot visualizes the functional relationship between (X_1, X_2) and Y .

3.3.1. Examples for kernel independence test on continuous domains

First, we consider the following functional model:

$$y_i = 0.1 \exp(3x_{1i}) + (2x_{2i} - 1)^2 + 10 \sin(x_{3i}) + 0 x_{4i} + \epsilon_i, \quad i = 1, \dots, 200 \quad (3.3)$$

where X_1, \dots, X_4 and error term ϵ_i are randomly generated from a standard normal distribution. Variable Y has a nonlinear additive dependence on the first three variables and is independent of the last one. A simulation was performed for statistical independence tests of the mutual dependence between X_1, \dots, X_4 and Y . The first four plots in Fig. 3.3 visualize the relationship between X_1, \dots, X_4 and Y . An experiment consisting of 1000 replications shows that the dependence relation $X_1 \not\perp Y$ could be verified correctly in 80.8% of all cases, 99.4% for $X_2 \not\perp Y$, 100% for $X_3 \not\perp Y$, and 99.1% for $X_4 \perp Y$ with a sample size of 200.

The second example is also an artificial one, first introduced by Gu et al. [79]. (x_{1i}, x_{2i}) , $i = 1, \dots, 200$, are generated randomly from a uniform distribution in the unit square and set the response to

$$y_i = 40 \frac{\exp \{8 [(x_{1i} - 0.5)^2 + (x_{2i} - 0.5)^2]\}}{\exp \{8 [(x_{1i} - 0.2)^2 + (x_{2i} - 0.7)^2]\}} + \exp \{8 [(x_{1i} - 0.7)^2 + (x_{2i} - 0.2)^2]\} + \epsilon_i. \quad (3.4)$$

The errors ϵ_i were drawn from a standard normal distribution. We tested the following independence relations: $X_1 \perp X_2$, $X_1 \not\perp X_2 | Y$, and $(X_1, X_2) \not\perp Y$. 99.4% of 1000 replications verified $X_1 \perp X_2$ correctly, 99.8% for $X_1 \not\perp X_2 | Y$, and 100% for $(X_1, X_2) \not\perp Y$. This example makes one advantage of kernel test apparent, i.e., the kernel measures can be straightforwardly applied to quantifying dependence between variables of different dimensions.

In summary, the results of both functional models showed that the kernel independence test reliably detected the nonlinear dependence using only a moderate sample size. The kernel measure

3. Kernel Statistical Test of Independence

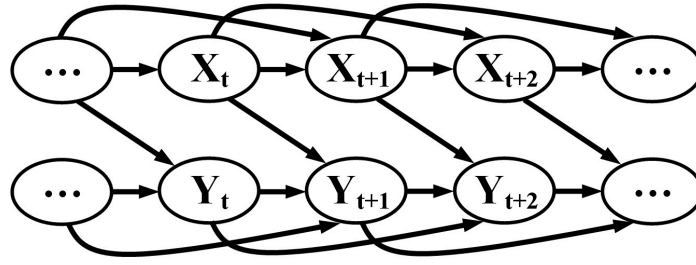


Figure 3.4.: Dynamic Bayesian network of a coupled time series.

provides a good alternative to capture the dependence between continuous variables.

3.3.2. Examples for kernel independence test on time series

Now, we consider a more challenging situation, namely time series. The difficulty of time series is that the assumption of the i.i.d. sample could be violated, even though it is stationary. Due to the additional temporal information, the causal direction is sometimes known in time-series data. DAGs, which capture the fact that time flows forward, can be naturally used for modelling time-series data. Arcs within a time-slice can be directed or undirected, since they model “instantaneous” dependence. If all arcs are directed, both within and between slices, the model is called “Dynamic Bayesian Networks” (DBN) [117].

In our experiments, we are interested in the case of a uni-directed influence $X \rightarrow Y$ between two times series $X = (\dots, X_t, X_{t+1}, X_{t+2}, \dots)$ and $Y = (\dots, Y_t, Y_{t+1}, Y_{t+2}, \dots)$ with point $t \in \mathbb{Z}$ in time. The graphical representation is given by a DBN as shown in Fig. 3.4. The independence constraints $(Y_{t+1} \perp\!\!\!\perp X_{t+2} \mid X_t, X_{t+1})$ and $(X_{t+1} \not\perp\!\!\!\perp Y_{t+2} \mid Y_t, Y_{t+1})$ characterize the causal direction $X \rightarrow Y$, because the dependence between Y_{t+1} and X_{t+2} is spurious, whereas the dependence between X_{t+1} and Y_{t+2} is generated by the direct causal influence from X to Y . The spurious dependence can be screened off by conditioning on the cause (X_t, X_{t+1}) , while the genuine dependence induced by the direct causal influence cannot be screened off by conditioning on the effect (Y_t, Y_{t+1}) .¹ We show a simulated experiment to demonstrate how well dependence or independence is captured by kernel tests.

We sample chaotic time series from coupled Hénon maps [87]. The parameters for the coupled Hénon maps are chosen as follows:

$$\begin{aligned} X_{t+2} &= 1.4 + 0.3 X_t - X_{t+1}^2, \\ Y_{t+2} &= 1.4 + 0.1 Y_t - (1 - \gamma) 0.4 Y_{t+1}^2 - \gamma X_{t+1} Y_{t+1}. \end{aligned} \quad (3.5)$$

This specific choice of parameters guarantees the dynamics in the times series $X = (X_0, X_1, \dots)$

¹Another well-known way to capture causality in bivariate time series is the so-called Granger causality [175, 72], which utilizes the temporal properties and is expressed in terms of predictability. The standard test of Granger causality developed by Granger [72] is based on a linear regression model. A test of Granger causality in kernel formalism will be interesting line of further research.

3.3. Simulated experiments with kernel independence test

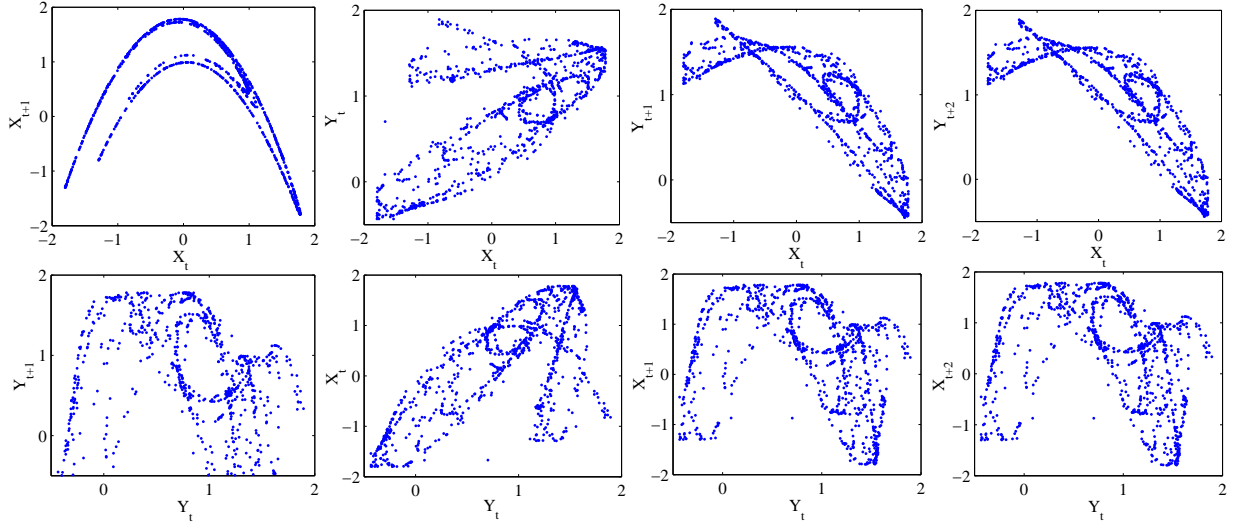


Figure 3.5.: Noiseless data sampled from coupled Hénon maps with coupling parameter $\gamma = 0.5$ as shown in Eq. (3.5) and sampling interval $k = 5$ (see text).

and $Y = (Y_0, Y_1, \dots)$, i.e., there exists no time step t so that for all $i \geq t$, X_i or Y_i takes a constant value.

Both X and Y are dynamical systems of the second order. We start (X, Y) with initial points $(X_0, Y_0) = (X_1, Y_1) = (0, 0)$ and collect data points for X and Y every k time steps, i.e., $X = (\dots, X_t, X_{t+k}, X_{t+2k}, \dots)$ and $Y = (\dots, Y_t, Y_{t+k}, Y_{t+2k}, \dots)$. The time step k is called sampling interval. The sampling interval k simulates the situation in real applications, e.g., in the study of biological data, where the exact time delay of influences might be unknown. This problem is known as temporal aggregation in some literatures [17].

X and Y are uncorrelated for $\gamma = 0$, while they are synchronized for $\gamma > 0$. Fig. 3.5 illustrates the dynamics between X and Y with coupling parameter $\gamma = 0.5$ and sampling interval $k = 5$.

Fig. 3.6 illustrates datasets of sample size 100 used in our experiments. All samples are added with an independent normally distributed noise $\mathcal{N}(0, 0.2^2)$. This noise simulated the noise of measurements in practice. We conducted the experiments with 1000 replications for different coupling factors γ and different sampling intervals k .

Tab. 3.2 shows the acceptance quota of independence hypothesis via the kernel independence tests (permutation tests with significance level $\alpha = 0.05$). In the cases of $\gamma = 0$, i.e., $X \perp\!\!\!\perp Y$, tests achieved consistent results. In the dependent cases, i.e., $\gamma > 0$, the underlying causal direction could be in most cases correctly identified when the sampling interval k is smaller than 7. If $k \geq 7$, i.e., a too large sampling rate, the dependences between X_{t+1} and Y_t and between Y_{t+1} and X_t vanish. Hence, the causal direction was not erroneously determined, but indeterminate. The best performance was achieved for $k = 5$. Interestingly, Yu et al. [182] found also that a sampling interval of 5 yielded the best results in their experiments with simulated biological data of time series.

3. Kernel Statistical Test of Independence

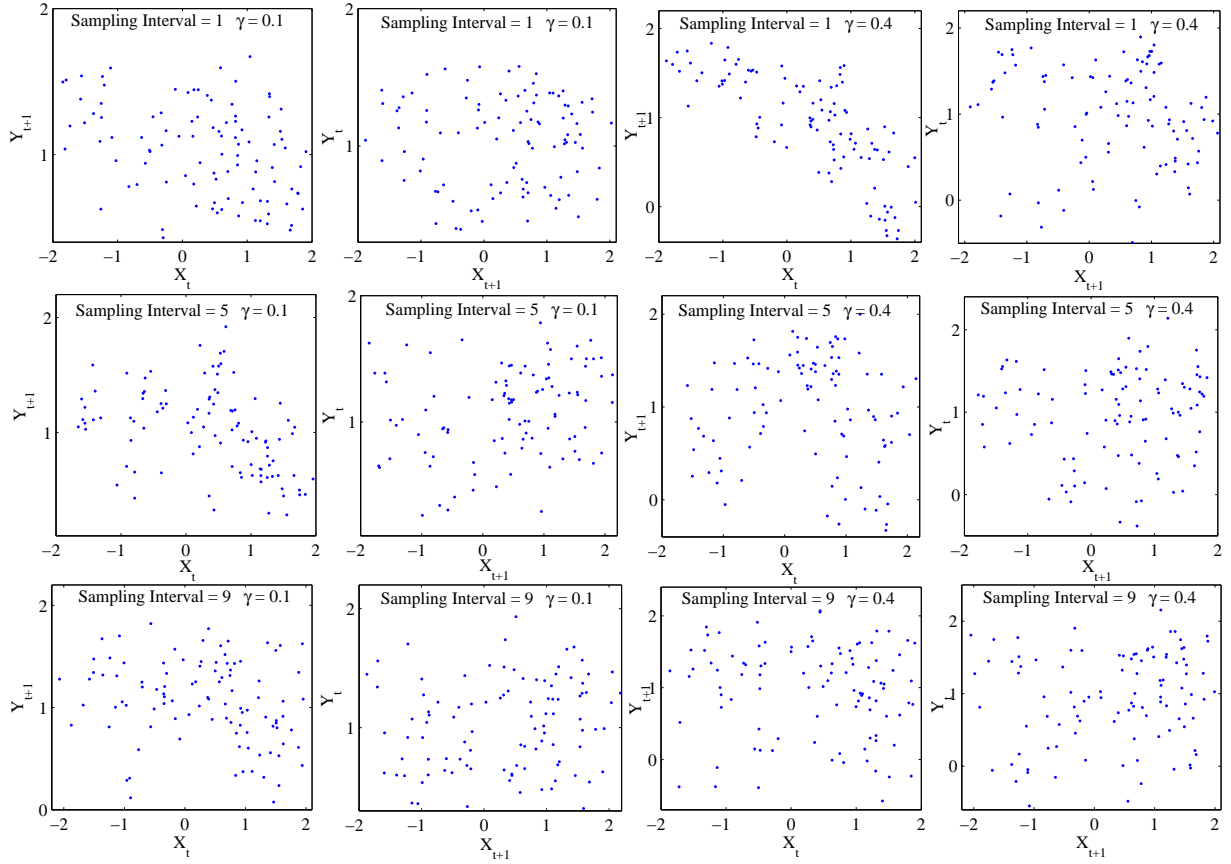


Figure 3.6.: Samples of 100 data points from coupled Hénon maps with different coupling parameters γ as shown in Eq. (3.5) and different sampling intervals k (see text).

γ (coupling)	Accepting $Y_{t+1} \perp\!\!\!\perp X_{t+2} \mid X_t, X_{t+1}$						Accepting $X_{t+1} \perp\!\!\!\perp Y_{t+2} \mid Y_t, Y_{t+1}$					
	0	0.1	0.2	0.3	0.4	0.5	0	0.1	0.2	0.3	0.4	0.5
$k = 1$	95.6	89.1	76.7	65.8	59.9	52.1	94.2	20.1	0	0	0	0
$k = 3$	93.5	88.9	89.6	83.3	77.8	64.9	93.3	1.3	0	0.1	0.2	0.2
$k = 5$	95.1	68.1	68.7	79.0	91.3	94.4	95.6	1.1	0	0.1	2.3	9.6
$k = 7$	94.5	90.1	80.9	79.8	89.6	97.4	95.1	2.4	5.0	33.8	52.2	59.9
$k = 9$	94.6	86.1	93.4	96.9	97.7	98.5	95.3	29.8	64.3	94.2	99.4	99.6

Table 3.2.: Kernel independence test on time series of coupled Hénon maps with different coupling parameters γ as shown in Eq. (3.5) and different sampling intervals k (see text). The entries show how often (in percentage) the independence is accepted.

3.3. Simulated experiments with kernel independence test

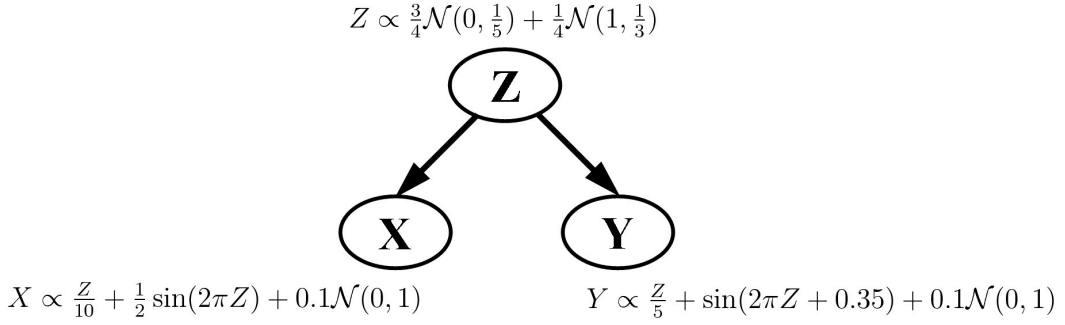


Figure 3.7.: Graphical representation of the underlying model of the meander dataset. The generating model implies $X \not\perp Y$ and $X \perp Y | Z$.

3.3.3. Numerical comparison of independence tests on continuous domain

In order to provide some numerical evidence of the performance of various tests on continuous domains, we conduct experiments with toy data generated by various functional models. Similar models are originally used by Margaritis in [108].

Three independence tests, i.e., Fisher’s Z test under multivariate Gaussian assumption, Margaritis’ Bayesian method [107, 109] and permutation test via kernel dependence measures, are evaluated on the so-called Meander dataset, shown in the left plot of Fig. 3.8. It resembles a spiral. This dataset is challenging because the joint distribution of X and Y given Z changes dramatically with the given value of Z . The data were generated by the model and equations shown in Fig. 3.7. According to the functional relation, X and Y are conditionally independent given Z , however, unconditionally dependent, in fact strongly correlated as seen from the right plot of Fig. 3.8.

We generated 1000 datasets of different sample sizes and ran independence tests. Tab. 3.3 shows the results for samples size ranging from 20 to 200. The dependence between X and Y can already be captured by the linear relation, as seen from the right plot of Fig. 3.8. For this reason, all methods achieved very good performance at testing $X \not\perp Y$ from merely 20 data points (see Fig. 3.9 for a sample of 20 data points).

Testing conditional independence $X \perp Y | Z$ is more challenging. Here, the kernel test clearly outperforms other two methods. The Fisher’s Z test fails completely due to the incorrect multivariate Gaussian assumption. The Margaritis’ Bayesian method or the kernel test performs better, as the sample size becomes larger and larger. The right plot in Fig. 3.10 shows the frequencies of p-values of testing $X \perp Y | Z$ from 1000 datasets of sample size 20 by the kernel independence test, while the left plot shows the probabilities of independence by the Margaritis’ Bayesian method. The kernel independence test made significantly less errors than the Margaritis’ Bayesian method.

In order to gain more numerical evidence of performance in learning models with various nonlinear relations, we sampled datasets of 200 data points by models shown in Fig. 3.11. The

3. Kernel Statistical Test of Independence

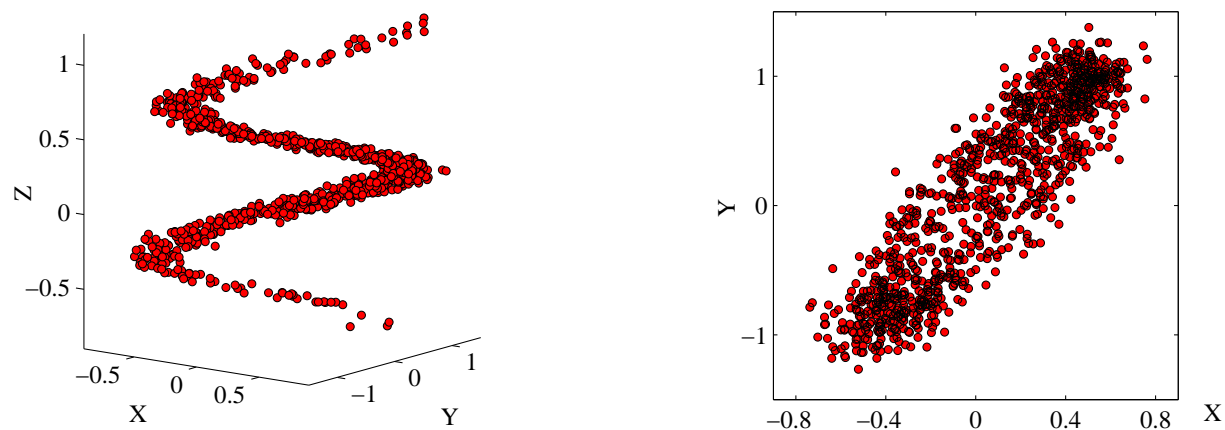


Figure 3.8.: Three-dimensional plot of the Meander dataset (left) and projection of data along Z axis (right). The generating model is shown in Fig. 3.7.

Sample Size	Rejecting $X \perp\!\!\!\perp Y$					Accepting $X \perp\!\!\!\perp Y Z$				
	20	50	100	150	200	20	50	100	150	200
Fisher's Z	100	100	100	100	100	0	0	0	0	0
Margaritis' Bayesian	94.3	100	100	100	100	4.8	15.1	21.2	23.2	33.2
Kernel Dependence	99.9	100	100	100	100	35.1	49.7	67.0	75.3	79.9

Table 3.3.: Numerical comparison of various independence tests on continuous domains, i.e., Fisher's Z test, Margaritis' Bayesian method, and permutation test via kernel dependence measures. The underlying model Meander is given by Fig. 3.7. One sample is illustrated in Fig. 3.8. The experiments are conducted with 1000 replications. The entries show how often (in percentage) the constraint $X \not\perp\!\!\!\perp Y$ and $X \perp\!\!\!\perp Y | Z$ are verified.

3.3. Simulated experiments with kernel independence test

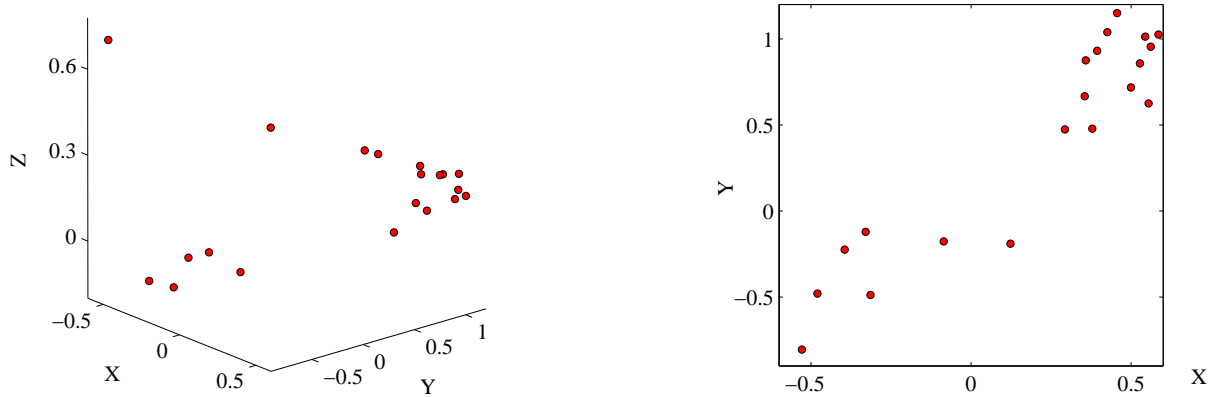


Figure 3.9.: A sample of Meander dataset of size 20. The underlying model as shown in Fig. 3.7 implies $X \not\perp Y$ and $X \perp Y | Z$.

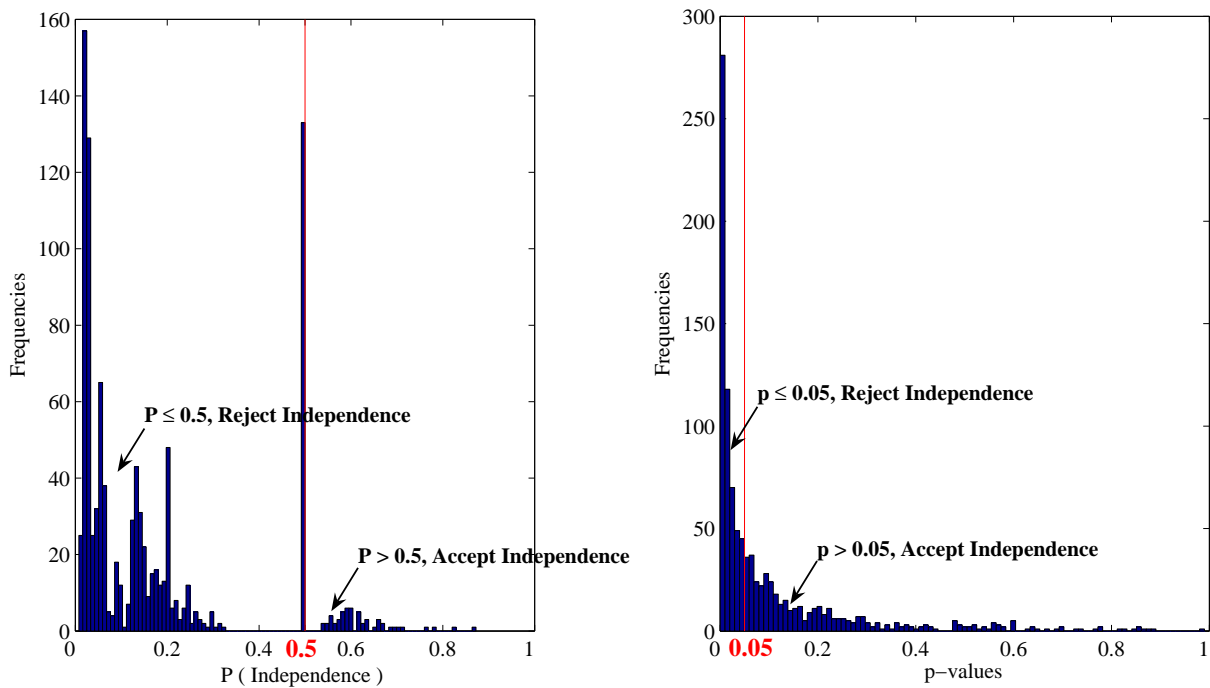


Figure 3.10.: Experimental results of Margaritis' Bayesian method and kernel test of independence, when $X \perp Y | Z$ is tested. Both methods are conducted with 1000 replications. The left plot is the histogram of the resulting $P(\text{Independence})$ of all 1000 samples, obtained by the Margaritis' Bayesian method. The independence hypothesis is accepted in only 4.8% of all cases. The right plot is the histogram of the resulting p-values of all 1000 samples, obtained by the kernel-based independence test. The independence hypothesis is accepted in 35.1% of all cases (see also Tab. 3.3).

3. Kernel Statistical Test of Independence

$M_k := (f_i, f_j)$	$f_2 := 2 \sin(x)$	$f_3 := \ln(x)$	$f_4 := \frac{1}{\frac{x}{5}+1}$	$f_5 := \exp(x)$
$f_1 := x$	M_1	M_2	M_3	M_4
$f_2 := 2 \sin(x)$	—	M_5	M_6	M_7
$f_3 := \ln(x)$	—	—	M_8	M_9
$f_4 := \frac{1}{\frac{x}{5}+1}$	—	—	—	M_{10}

Table 3.4.: 10 different pair of functions (f_i, f_j) with $i, j = 1, \dots, 5$ define the functional models M_1, \dots, M_{10} , which are used to generate data by two models as shown in Fig. 3.11.

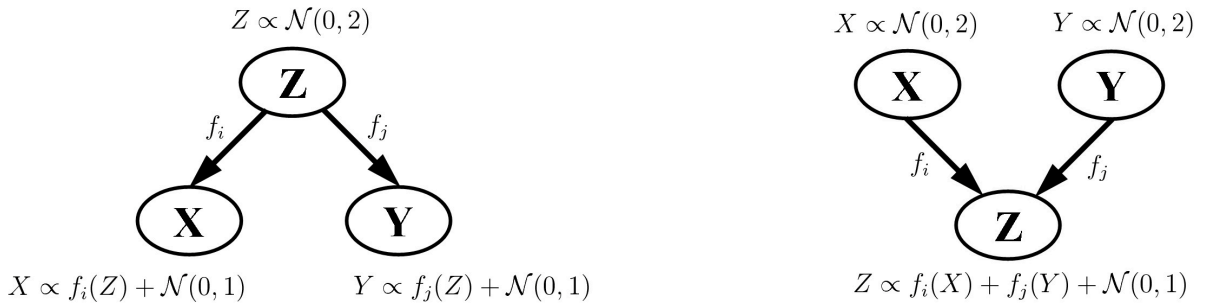


Figure 3.11.: Graphical representation of a fork (left) and a collider (right) structure. Models with a fork structure (non- v -structure) imply $X \not\perp\!\!\!\perp Y$ and $X \perp\!\!\!\perp Y \mid Z$, while models with a collider structure (v -structure) imply $X \perp\!\!\!\perp Y$ and $X \not\perp\!\!\!\perp Y \mid Z$. The pairs of functions $M_k = (f_i, f_j)$ for both models are defined in Tab. 3.4.

left plot of Fig. 3.11 is a non- v -structure: fork structure, while the right plot is a v -structure: (unshielded) collider structure. Models of a fork structure imply the independence relations $X \not\perp\!\!\!\perp Y$ and $X \perp\!\!\!\perp Y \mid Z$, while models of a collider structure imply the independence relations $X \perp\!\!\!\perp Y$ and $X \not\perp\!\!\!\perp Y \mid Z$.

We define $f_{1,\dots,5}$ in the same way as Margaritis proposed in [108] and use all pairs of the functions $M_k = (f_i, f_j)$, i.e., 10 different combinations M_1, \dots, M_{10} as shown in Tab. 3.4, added by a Gaussian noise as underlying ground-truth for the sampling. One sample of 200 data points for the fork structure (left plot in Fig. 3.11) with M_1, \dots, M_{10} (see Tab. 3.4) is visualized in Fig. 3.12 and Fig. 3.13. The performance of various independence tests 1000 replications on these datasets is summarized in Tab. 3.5. One sample of 200 data points for the collider structure (right plot in Fig. 3.11) with M_1, \dots, M_{10} (see Tab. 3.4) is visualized in Fig. 3.14 and Fig. 3.15. The performance of various independence tests of 1000 replications on these datasets is summarized in Tab. 3.6.

One can see that all three methods make relatively few errors at discovering independence, i.e., $X \perp\!\!\!\perp Y$ in a collider structure (see the left half of Tab. 3.6) and $X \perp\!\!\!\perp Y \mid Z$ in a fork structure (see the under half of Tab. 3.5). Only the Fisher's Z test performed very bad in the case of testing conditional independence $X \perp\!\!\!\perp Y \mid Z$ (see the first row of the under half of Tab. 3.5) on data sampled by models M_7 and M_9 . It is hard to evaluate the performance of these three

3.3. Simulated experiments with kernel independence test

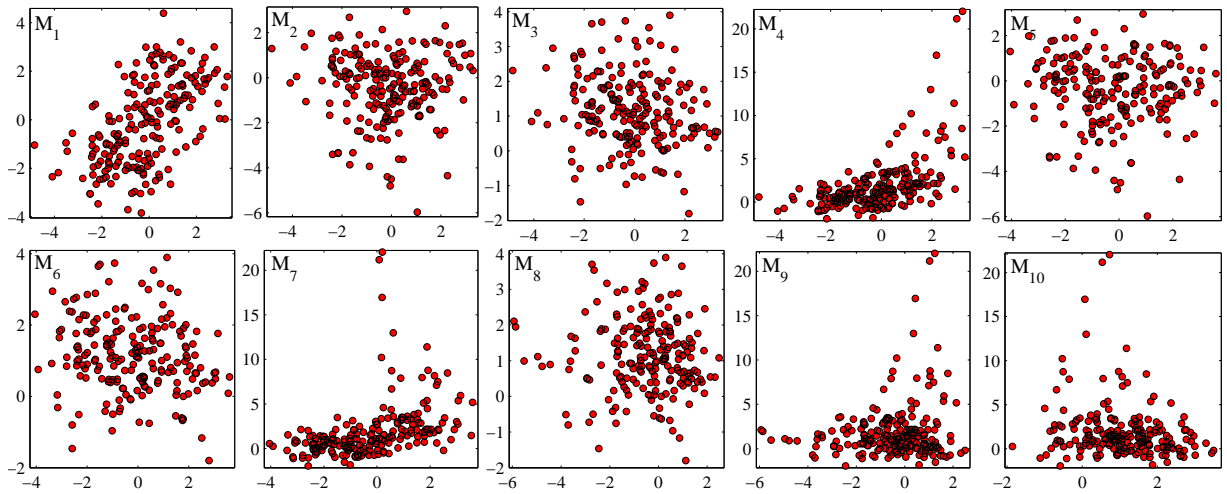


Figure 3.12.: The underlying model is a fork structure (Fig. 3.11, left), where X and Y have a functional relation $M_k = (f_i, f_j)$ (see Tab. 3.4) with Z , respectively. The fork structure implies that X and Y are unconditionally dependent. The illustrated sample contains 200 data points.

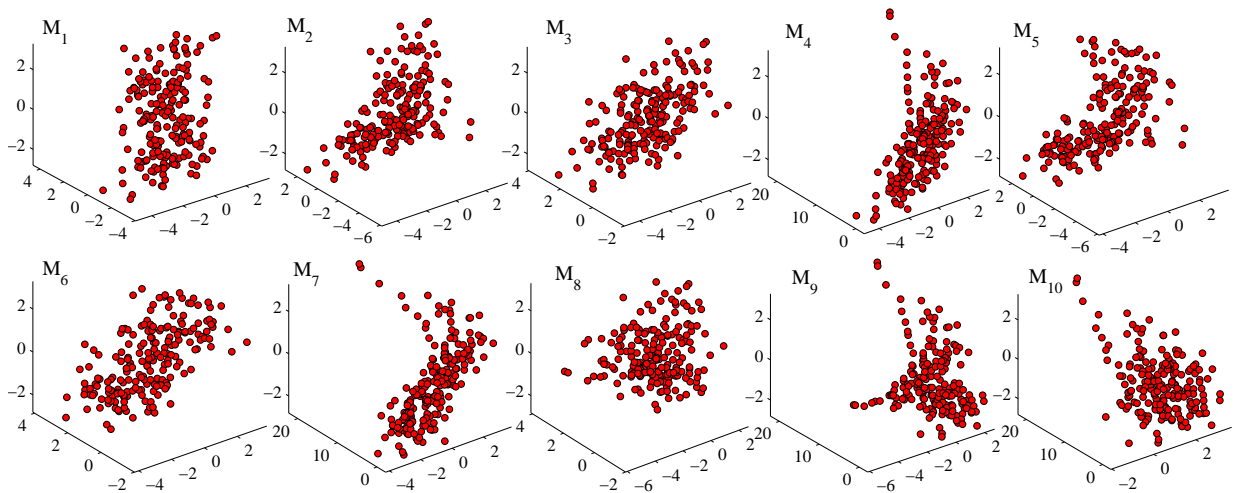


Figure 3.13.: The underlying model is a fork structure (Fig. 3.11, left), where X or Y has a functional relation $M_k = (f_i, f_j)$ (see Tab. 3.4) with Z , respectively. The fork structure implies that X and Y are independent, conditional on Z . The illustrated sample contains 200 data points.

3. Kernel Statistical Test of Independence

	Rejecting $X \perp\!\!\!\perp Y$									
$M_k = (f_i, f_j)$	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8	M_9	M_{10}
Fisher's Z	100	4.2	93.6	100	1.8	57.6	71.5	17.5	41.0	58.6
Margaritis' Bayesian	100	2.0	42.5	100	2.0	27.6	100	1.7	4.0	18.6
Kernel Dependence	100	95.6	63.9	100	63.9	56.1	100	11.5	97.8	68.7
	Accepting $X \perp\!\!\!\perp Y Z$									
Fisher's Z	94.0	95.6	94.1	95.6	95.5	72.2	10.6	81.5	1.2	64.3
Margaritis' Bayesian	97.0	97.6	97.9	98.7	97.0	97.9	98.9	98.3	98.7	98.8
Kernel Dependence	93.8	93.8	92.5	93.4	93.3	93.5	93.4	94.5	94.2	92.9

Table 3.5.: Numerical comparison of various independence tests, i.e., Fisher's Z test, Margaritis' Bayesian method, and test via kernel dependence measure, on continuous domains sampled by a fork structure (Fig. 3.11, left). The parameter $M_k = (f_i, f_j)$ of models is defined in Tab. 3.4. The entries show how often (in percentage) $X \not\perp\!\!\!\perp Y$ and $X \perp\!\!\!\perp Y | Z$ are verified after 1000 replications of simulations.

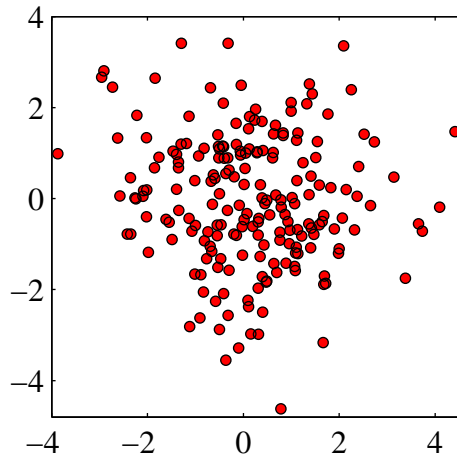


Figure 3.14.: The underlying model is a collider structure (Fig. 3.11, right), where Z has a functional relation $M_k = (f_i, f_j)$ (see Tab. 3.4) with X and Y . The collider structure implies that X and Y are unconditionally independent. The illustrated sample contains 200 data points.

3.3. Simulated experiments with kernel independence test

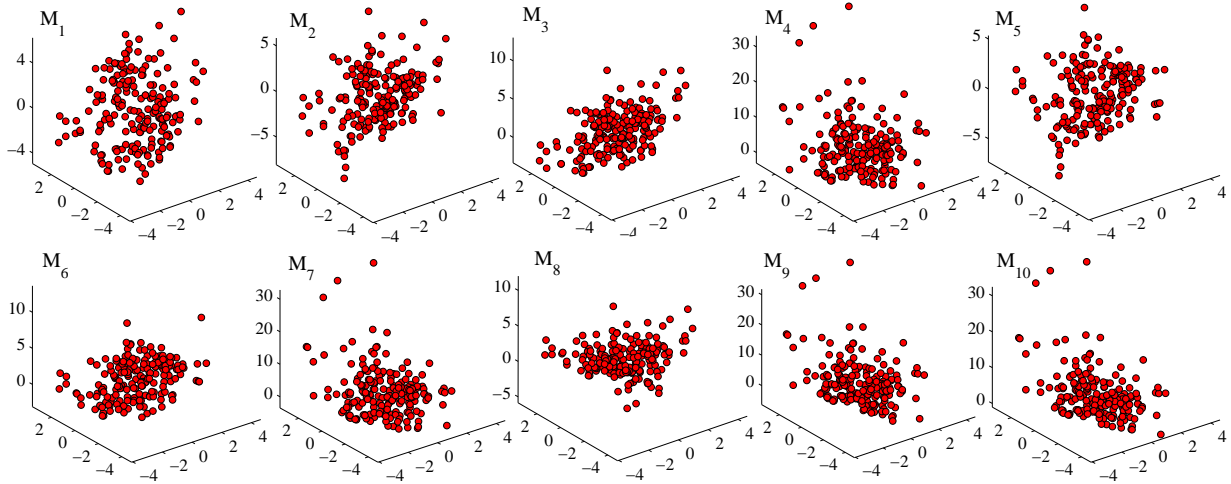


Figure 3.15.: The underlying model is a collider structure (Fig. 3.11, right), where Z has a functional relation $M_k = (f_i, f_j)$ (see Tab. 3.4) with X and Y . The collider structure implies that X and Y are dependent, conditional on Z . The illustrated sample contains 200 data points.

	Accepting $X \perp\!\!\!\perp Y$	Rejecting $X \perp\!\!\!\perp Y Z$									
$M_k = (f_i, f_j)$	$M_{1,\dots,10}$	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8	M_9	M_{10}
Fisher's Z	94.6	100	4.1	92.1	77.1	4.7	58.8	61.2	5.1	3.9	20.8
Margaritis' Bayesian	98.1	91.4	3.9	10.9	84.8	3.1	9.1	75.0	2.1	3.7	6.7
Kernel dependence	94.1	100	92.0	60.7	100	96.0	51.7	100	18.7	93.6	46.9

Table 3.6.: Numerical comparison of various independence tests, i.e., Fisher's Z test, Margaritis' Bayesian method, and test via kernel dependence measure, on continuous domains sampled by a collider structure (Fig. 3.11, right). The parameter $M_k = (f_i, f_j)$ of models is defined in Tab. 3.4. The entries show how often (in percentage) $X \perp\!\!\!\perp Y$ and $X \not\perp\!\!\!\perp Y | Z$ are verified after 1000 replications of simulations.

3. Kernel Statistical Test of Independence

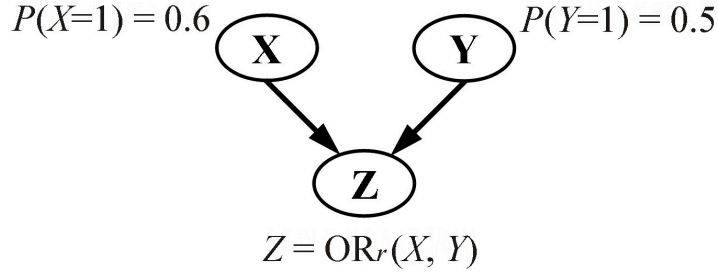


Figure 3.16.: Graphical representation of a 2-bit noisy OR with a noise level $r \in [0, 1]$ as shown in Eq. (3.6).

methods in testing dependence, i.e., $X \not\perp\!\!\!\perp Y$ in a fork structure (see the upper half of Tab. 3.5) and $X \not\perp\!\!\!\perp Y | Z$ in a collider structure (see the right half of Tab. 3.6). But the results indicate that the fluctuation of the kernel-based approach within different models is significantly smaller than that of the other two methods.

3.3.4. Numerical comparison of independence tests on discrete domain

The kernel independence test can be straightforwardly applied to both continuous and discrete variables. In order to give numerical evidence of the performance, we conduct experiments with toy data on discrete domains. The data are sampled from logically linked models, namely noisy OR gates. Such Boolean functions are simplified models for many intuitive causal relations in real life. Note that one can easily get an AND gate by inverting inputs and outputs from an OR gate, therefore the results of OR can be easily re-interpreted with reference to AND.

In general, an n -bit $X_1, \dots, X_n \in \{0, 1\}$ noisy OR gate (see Henrion [88]) can be characterized by the conditional probabilities

$$P(X_{n+1}=1 | x_1, \dots, x_n) = (1 - r_2) (1 - r_1^{x_1 + \dots + x_n}) + r_2$$

with parameters $r_1, r_2 \in [0, 1]$. r_1 can be interpreted as the probability of suppressing the input 1; r_2 can be interpreted as the probability for a spontaneous inversion of the output. If r_1 and r_2 vanish, the OR gate is deterministic. For the sake of notational simplicity, we chose $r_1 = r_2 =: r$ in our experiments, i.e.,

$$P(X_{n+1}=1 | x_1, \dots, x_n) = (1 - r) (1 - r^{x_1 + \dots + x_n}) + r. \quad (3.6)$$

We use the shorthand $\text{OR}_r\{X_1, \dots, X_n\}$ to depict a noisy OR gate with noise level $r \in [0, 1]$.

We sampled data from a 2-bit noisy OR (Fig. 3.16) with X, Y as inputs and Z as output. The underlying model implies $X \perp\!\!\!\perp Y$ and $X \not\perp\!\!\!\perp Y | Z$. We sampled 1000 datasets for each of the noise levels $r = 0, 0.1, 0.2, 0.3$ and sample sizes 20, 50, 100, 150, 200. Fig. 3.17 shows the noise statistics in the term of percentage of erroneous outputs in 1000 data points.

3.3. Simulated experiments with kernel independence test

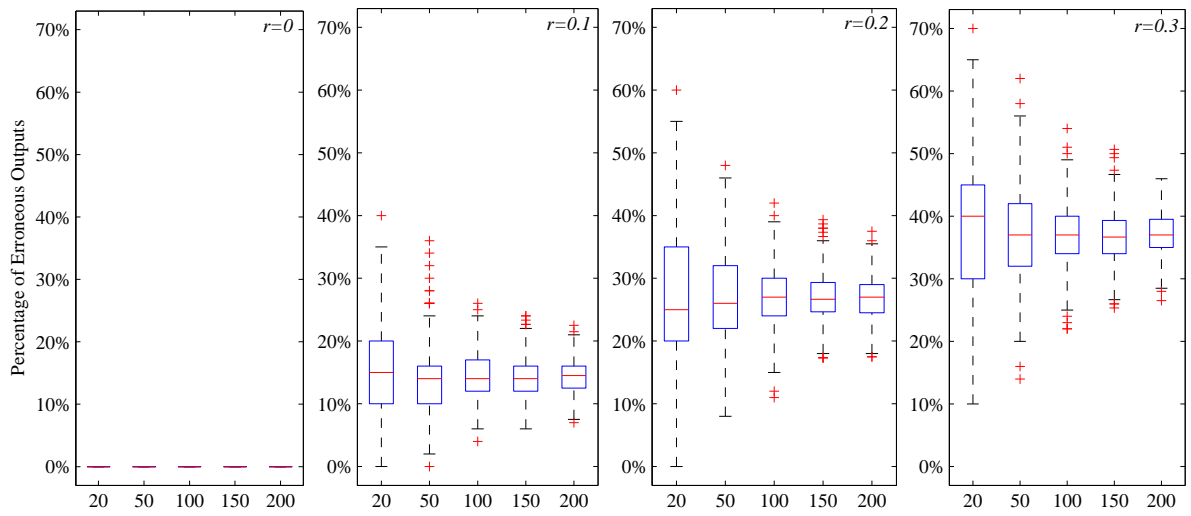


Figure 3.17.: Noise statistics in the term of percentage of erroneous outputs in 1000 data points sampled by the 2-bit noisy OR gate as shown in Fig. 3.16. The plots illustrate 4 different noise levels $r = 0, 0.1, 0.2, 0.3$ as shown in Eq. (3.6). Each box has lines at the lower quartile, median, and upper quartile values of the percentage of erroneous outputs. The whiskers are lines extending from each end of the box to show the extent of the rest of the percentage. Outliers are the percentage beyond the ends of the whiskers.

3. Kernel Statistical Test of Independence

	Accepting $X \perp\!\!\!\perp Y$														
Sample Size	20			50			100			150			200		
Noisy OR	χ^2	MI	KD	χ^2	MI	KD	χ^2	MI	KD	χ^2	MI	KD	χ^2	MI	KD
$r = 0$	94.0	97.4	88.0	94.7	96.3	90.8	95.6	95.9	92.5	94.3	93.7	92.4	94.1	94.3	91.8
$r = 0.1$	93.1	96.4	86.5	94.6	96.0	90.7	94.2	94.0	91.1	95.8	96.3	94.4	94.2	94.8	92.3
$r = 0.2$	93.6	96.9	86.9	94.9	96.1	91.3	96.3	96.1	93.1	95.7	95.7	93.5	93.6	94.0	91.4
$r = 0.3$	94.5	97.1	87.3	95.9	97.0	93.0	93.5	93.6	90.7	93.6	94.1	91.6	94.4	94.8	93.2
Noisy OR	Rejecting $X \perp\!\!\!\perp Y Z$														
$r = 0$	24.8	54.8	23.5	94.5	97.7	91.9	100	100	98.0	100	100	100	100	100	100
$r = 0.1$	23.5	33.7	16.9	57.6	57.0	54.8	85.9	84.7	89.2	97.6	97.3	98.0	99.3	99.2	99.7
$r = 0.2$	14.9	18.9	8.9	25.1	22.7	22.5	39.8	40.7	40.5	56.2	57.2	60.0	71.6	72.5	74.8
$r = 0.3$	9.9	10.9	7.0	10.4	10.3	9.6	16.3	16.5	16.3	19.1	21.5	19.2	23.1	23.9	23.9

Table 3.7.: Numerical comparison of three different independence tests, i.e., likelihood ratio χ^2 test, permutation test via mutual information (MI), and permutation test via kernel dependence (KD) measure. The generating models are noisy OR gates with 4 different noise levels $r = 0, 0.1, 0.2, 0.3$ as shown in Fig. 3.16 and Eq. (3.6). The experiments are conducted with 1000 replications. The entries show how often (in percentage) the constraint $X \perp\!\!\!\perp Y$ or $X \not\perp\!\!\!\perp Y | Z$ is verified.

We perform three independence tests i.e., likelihood ratio χ^2 test, permutation test via mutual information (MI) and permutation test via kernel dependence (KD) measure. A significance level of 5% are used for all tests. In permutation tests via MI and KD, we used a repetition factor of 100. As seen from Tab. 3.7, their performance are very similar, in the sense that the levels of type I and II errors are almost the same. The larger the sample size and the less noisy the model, the better the performance. The kernel-based method is slightly worse than the other two in the case of 20 data points. Taking the computational efficiency into account, the likelihood ratio χ^2 test is clearly the winner in this example. The actual benefit of the kernel test does not lie in the tests on discrete domains, but in tests on continuous or hybrid domains. Nonetheless, the kernel independence test provides an alternative to the popular tests on discrete domains.

3.4. Real-world experiments with kernel independence test

To demonstrate the effectiveness of statistical test of independence by means of kernel measures, we demonstrate some real-world applications in this section.

3.4. Real-world experiments with kernel independence test

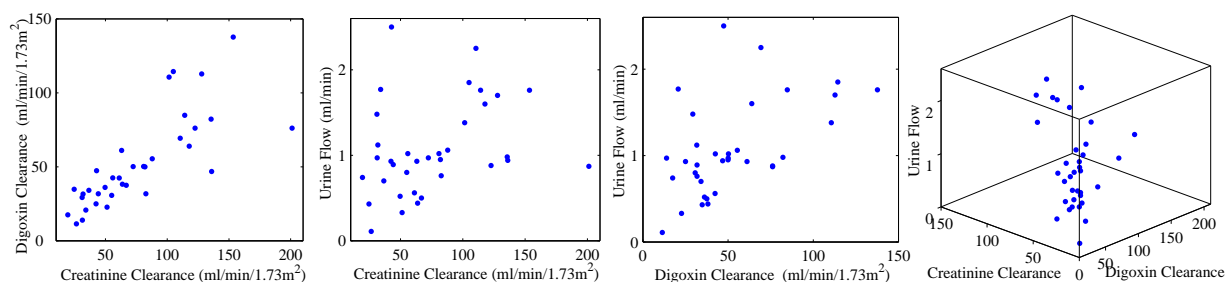


Figure 3.18.: Data on 35 consecutive patients under treatment for heart failure with the drug digoxin. Clearances are given in ml/min/1.73m², urine flow in ml/min.

3.4.1. Digoxin clearance

The study of the passage of drugs through the body is important in medical science. The right-most 3d-plot of Fig. 3.18 shows a real-world dataset on 35 consecutive patients under treatment for heart failure with the drug digoxin [82] (see also [5] p. 323 and [52] p. 42 for the same dataset). The renal clearances of digoxin, creatinine, and urine flow were determined simultaneously in each of the patients receiving digoxin, in most of whom there was prerenal azotemia. The digoxin clearance is the amount of blood that in a given interval is cleared of digoxin. The creatinine clearance is defined similarly and used as a measure of kidney function. Of medical interest is the hypothesis that digoxin clearance is independent of urine flow conditioning on creatinine clearance. Halkin et al. [82] and Edwards [52] based their analysis on the (partial) correlation coefficient. Recall that a partial correlation coefficient is calculated by the usual correlation coefficients as defined in Eq. (2.1):

$$\rho_{YX|Z} = \frac{\rho_{YX} - \rho_{ZY}\rho_{ZX}}{\sqrt{(1 - \rho_{ZY}^2)(1 - \rho_{ZX}^2)}}. \quad (3.7)$$

Tab. 3.8 shows the results of permutation test via kernel dependence measure in comparison with correlation analysis. A visual inspection of the data, as shown in the first plot of Fig. 3.18, indicates that the linearity assumption appears to be reasonable for the dependence between the creatinine and digoxin clearances (Fig. 3.18, leftmost). A linear relation between them was first suggested by Jelliffe et al. [95] and later confirmed by various clearance studies, which revealed a close relationship between creatinine and digoxin clearance in many patients. The ready explanation is that both creatinine and digoxin are mainly eliminated by the kidneys. In agreement with this explanation, both correlation analysis and kernel test indeed found the unconditional and conditional dependence (first and second row of Tab. 3.8).

As one can see from Fig. 3.18, the relations between creatinine clearance and urine flow (second plot) and between digoxin clearances and urine flow (third plot) are less linear than the relation between creatinine and digoxin clearance (first plot). The correlation analysis (see also [52] p. 43) did not find the dependence between creatinine clearance and urine flow, while kernel test did (third row of Tab. 3.8). Both partial correlation technique and test via the kernel measure

3. Kernel Statistical Test of Independence

Independence Hypothesis	Correlation Analysis			Kernel Dependence		
	Measure	p-Value	Test	Measure	p-Value	Test
Creatinine Clearance \perp Digoxin Clearance	0.7754	0.00	Reject	0.0625	0.00	Reject
Creatinine Clearance \perp Digoxin Clearance Urine Flow	0.7584	0.00	Reject	0.0134	0.00	Reject
Creatinine Clearance \perp Urine Flow	0.3092	0.07	Accept	0.0212	0.01	Reject
Creatinine Clearance \perp Urine Flow Digoxin Clearance	0.1914	0.40	Accept	0.0025	0.58	Accept
Digoxin Clearance \perp Urine Flow	0.5309	0.00	Reject	0.0254	0.00	Reject
Digoxin Clearance \perp Urine Flow Creatinine Clearance	0.4847	0.02	Reject	0.0040	0.17	Accept

Table 3.8.: Correlation analysis and kernel independence test on digoxin clearance data. The significance level $\alpha = 0.05$ is chosen.

found that, given digoxin clearance, creatinine clearance was not significantly related to urine flow rate (fourth row of Tab. 3.8).

Moreover, both methods found that in these patients digoxin clearance was significantly related to urine flow rate (fifth row of Tab. 3.8). This finding is consistent with the opinion of Halkin et al. [82], who suspected that the elimination of digoxin might be subject to reabsorption, which might give rise to a correlation with urine flow.

However, if the linear dependence model is wrong, a biased estimate of the partial correlation and a biased test for independence via linear model may result. Test via kernel dependence measure accepted the hypothesis that, given creatinine clearance, digoxin clearance is independent of urine flow, whereas the partial correlation did not confirm this hypothesis (sixth row of Tab. 3.8). The finding that digoxin clearance is independent of urine flow controlling for creatinine clearance is particularly of medical interest.

In summary, the results revealed that the test via kernel dependence measure is superior to correlation analysis. This example makes it clear that, in practice, independence by kernel measures does not necessarily require the independence by correlation analysis, although it is theoretically apparent that non-vanishing of correlation implies non-vanishing of dependence by the kernel measure.

3.4.2. Rats' weights

A dataset of rats' weights is studied first by Morrison [116], then by Mardia et al. [106] and by Edwards [52]. The data stem from a drug trial, in which the weight losses of male and female rats under 3 drug treatments are studied. 4 different kinds of rats of each sex are assigned at random to each drug. Weight losses are observed after one and two weeks. There are thus 24 observations ($= 4 \text{ rat} \times 2 \text{ gender} \times 3 \text{ drug}$) on variables: sex, drug, and weight loss after one and two weeks. The data, which are visualized in Fig. 3.19 can be found in [52] p. 76. Both "sex" and "drug" have categorical domains. The domain of variable "weight loss" is 2-dimensional, since weight losses are characterized by distinct values after one week and after two weeks.

Models on hybrid domains (mixture of categorical and continuous variables of different di-

3.4. Real-world experiments with kernel independence test

Independence Hypothesis	Omitting Drug C			Including Drug C		
	Kernel Measure	p-Value	Test	Kernel Measure	p-Value	Test
Sex \perp Drug	0.0000	0.33	Accept	0.0856	0.80	Accept
Sex \perp Drug Weight Loss	0.0049	0.46	Accept	0.0045	0.83	Accept
Sex \perp Weight Loss	0.3545	0.14	Accept	0.3690	0.00	Reject
Sex \perp Weight Loss Drug	0.0049	0.99	Accept	0.0030	0.91	Accept
Drug \perp Weight Loss	0.3545	0.08	Accept	0.2800	0.02	Reject
Drug \perp Weight Loss Sex	0.0049	0.91	Accept	0.0073	0.77	Accept

Table 3.9.: Kernel independence test on rats’ weight data. The right half of the table shows the test results on the whole dataset, while the left half of the table shows the test results on the dataset omitting drug C. This way, the effect of drug C is apparent.

mensions) can be treated in a sophisticated way by conventional methods (see e.g., [52] p. 76). In contrast, the kernel measure can deal with categorical or high-dimensional variables in a straightforward way. More precisely, in conventional methods, variable “sex” has the value set $\{1, 2\}$ for $\{\text{male, female}\}$, and variable “drug” has the value set $\{1, 2, 3\}$ for $\{\text{drug A, drug B, drug C}\}$ in [52]. The kernel methods use the assignment $\{(1, 0), (0, 1)\}$ for the value set of sex and $\{(1, 0, 0), (0, 1, 0), (0, 0, 1)\}$ for the value set of drug. In the binary case, the vectorial assignment makes no difference to the scalar assignment. In the case of a ternary value set, the vectorial assignment is more suitable, since the scalar assignment makes a restrictive assumption about the differences between the 3 drugs. Note that testing independence between vectorial variables can not be treated by conventional methods in a straightforward way like kernel methods.

Tab. 3.9 summarizes the results of kernel tests of non-trivial independence relations between “sex”, “drug” and “weight loss” in the case that drug C is omitted (12 data points) or included (24 data points). Edwards (see [52] p. 78) suspected that there is no difference between drug A and B with regard to weight loss, whereas drug C differs widely from them. One may expect this finding intuitively from the plot of data as shown in Fig. 3.19. The result of kernel independence test is consistent with this finding.

3.4.3. Doctor visits and age/gender

In some social and medical studies, an ensemble of associated hypotheses need to be tested. As an example, we study the behavior of doctor visits, more precisely, we analyze the relation between the age or gender of a person and the number of his/her doctor visits in Germany. Such studies are useful for countries with large public health sector where the incentive structures may not promote efficient use of resources. By means of this example, we will show that the kernel independence test provides more power than linear analysis in the so-called multiple testing.

The typical real-world data of this context provides a large number of subgroups (male over 50 who are poor-earning, etc.). The independence hypothesis can be tested on each of the subgroups. In particular, the underlying distributions of subgroups may differ from each other. Since

3. Kernel Statistical Test of Independence

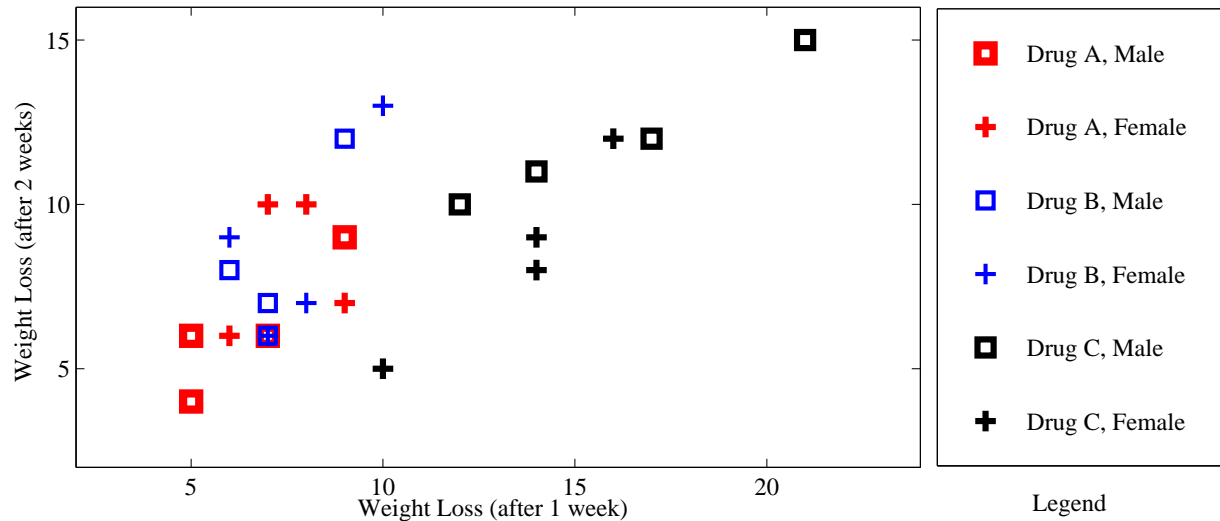


Figure 3.19.: Data from drug trial on rats. The weight losses after one week and after two weeks of 24 male and female rats under 3 drug (drug A, B and C) treatments are studied. Drug A, B and C are randomly assigned to the rats.

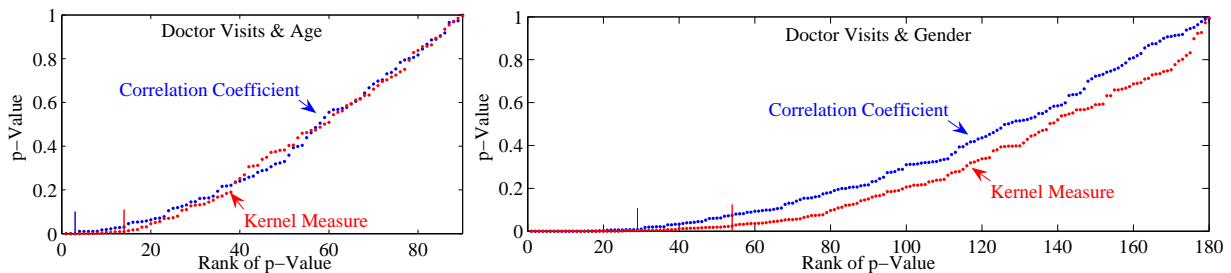


Figure 3.20.: Q-Q plot of p-values on doctor visit data: p-values of hypothesis tests on the relation between doctor visits and age (left plot), and between doctor visits and gender (right plot). FDR is controlled at a level of 5%. The vertical lines depict the cut-off of rejecting independence. The multiple testing via kernel measure rejected more independence hypotheses, i.e., detected more dependence, than the multiple testing via correlation analysis.

3.4. Real-world experiments with kernel independence test

a series of tests is conducted, each with a pre-specified significance level α , the appropriate threshold to declare a set of p-values significant becomes much more complex. This is known as the “multiple testing” (or “multiple comparison”) problem in statistical tests. In the absence of dependence, each test has a chance of α to yield a significant result, and the chance of drawing at least one false conclusion increases rapidly with the number of tests performed.

An elegant way to deal with this problem, which was first advocated for ecological studies by Garcia [64, 65], is to control the proportion of significant results that are in fact type I errors (“false discoveries”), the so-called false discovery rate (FDR), instead of controlling the chance of making even a single type I error. A comprehensive overview of various versions of FDR control can be found in [20, 168]. Storey presented in [160] a Bayesian interpretation of the FDR. Based on the FDR control, Benjamini et al. [18, 19, 21] developed the so-called multiple testing method.

The dataset that we studied originally come from the German Socio-Economic Panel, 1995-1999 [78], and are extracted by Winkelmann [177]. The observations include persons aged 20 – 60 associated with non-guest worker households in west Germany. Privately insured individuals (about 6% of the entries) as well as observations with missing values are excluded from the analysis. The final sample comprises 32, 837 observations. The 6 variables, which we are interested in, are DOCTOR VISITS (number of doctor visits in last three months), YEAR (calendar year of the observation, i.e., 1995, . . . , 1999), AGE (in years, the interval 20 – 60 is discretized into 2, . . . , 5 for 20 – 30, . . . , 50 – 60), GENDER (“0” for female and “1” for male), HEALTH (self-assessment, “-1” for bad, “1” for good, otherwise “0”), and INCOME (logarithm of monthly gross income, “-1” for low, i.e. smaller than 7, “0” for middle, i.e., between 7 and 8, “1” for high, i.e., larger than 8).

After partitioning the dataset subject to YEAR, GENDER, HEALTH, and INCOME, we obtained 90 subgroups for the study of age difference in doctor visits. After partitioning the dataset subject to YEAR, AGE, HEALTH, and INCOME, we obtained 180 subgroups for the study of gender difference in doctor visits. Fig. 3.20 visualizes the set of resulting p-values $p_{\pi(1)} \leq \dots \leq p_{\pi(i)} \leq \dots \leq p_{\pi(m)}$ of testing the set of 90 or 180 independence hypotheses via correlation and kernel measure. π is the permutation that sorts the p-values in an increasing order. The plot of $p_{\pi(i)}$ versus $\pi(i)$ is called Q-Q (“Q” stands for quantile) plot of p-values (see [139, 90] for details).

Using the Q-Q plot of p-values, Benjamini et al. (see [19] p. 71) presented a so-called adaptive procedure to control the FDR in multiple testing with independent test statistics. A graphical implementation and a detailed computational example of this procedure can be found in [19]. Conducting the Benjamini’s procedure, a cut-off of p-values for decision of dependence can be found when FDR is controlled at a level of 5%. As one can see from Tab. 3.10, the test via kernel measures provides more power than correlation analysis in the framework of multiple testing, in the sense that significantly more dependences can be detected via the kernel dependence measure than via the correlation coefficient.

In order to take a close look at the dependences detected by kernel methods, we illustrate the Q-Q plots with different colors for different subgroups. A dependence can be observed over the whole time period (1995-1999) studied (see first row of Fig. 3.21). Since a major health care reform took place in 1997 in Germany, this finding supports the conjecture that an age/gender difference in the number of doctor visits was insensitive to the system reform. The dependence

3. Kernel Statistical Test of Independence

Independence Hypothesis	Number of All Hypotheses	Number of Rejected Hypotheses \Rightarrow Dependence	
		Correlation Coefficient	Kernel Measure
Doctor Visits \perp Age	90	3	14
Doctor Visits \perp Gender	180	29	54

Table 3.10.: Multiple hypothesis testing by means of correlation and kernel measure on doctor visit data. The level of FDR is controlled at 5%. The kernel method has more power in the sense that testing via the kernel measure detected significantly more dependence than testing via correlation.

corresponds to the middle income group in most cases (see last row of Fig. 3.21).

The resulting subgroups, for which a gender difference in the behavior of doctor visits is verified, indicate that men less often visit doctor than women. This tendency is, in particular, present, if the person actually feels good or not so bad (see plot in row 3, column 2 of Fig. 3.21), and is relatively young, i.e., 20–40 (see plot in row 2, column 2 of Fig. 3.21).

The resulting subgroups, for which an age difference in the number of doctor visits is verified, indicate a positive correlation in men and negative correlation in women (see plot in row 2, column 1 of Fig. 3.21). That means the older the men, the more often his doctor visits. And the older the women, the less often her doctor visits. This tendency was observed, in particular, in subgroups of men of bad health or women of good health (see plot in row 3, column 1 of Fig. 3.21). We conjecture that this finding might be due to gynecologist visits of women in younger years and the relatively bad health status of men in older years. The dependence corresponds to a middle income group in most cases (see plot in last row of Fig. 3.21). We conjecture that an extremely high or low income substantially influences the behavior of doctor visits.

3.4. Real-world experiments with kernel independence test

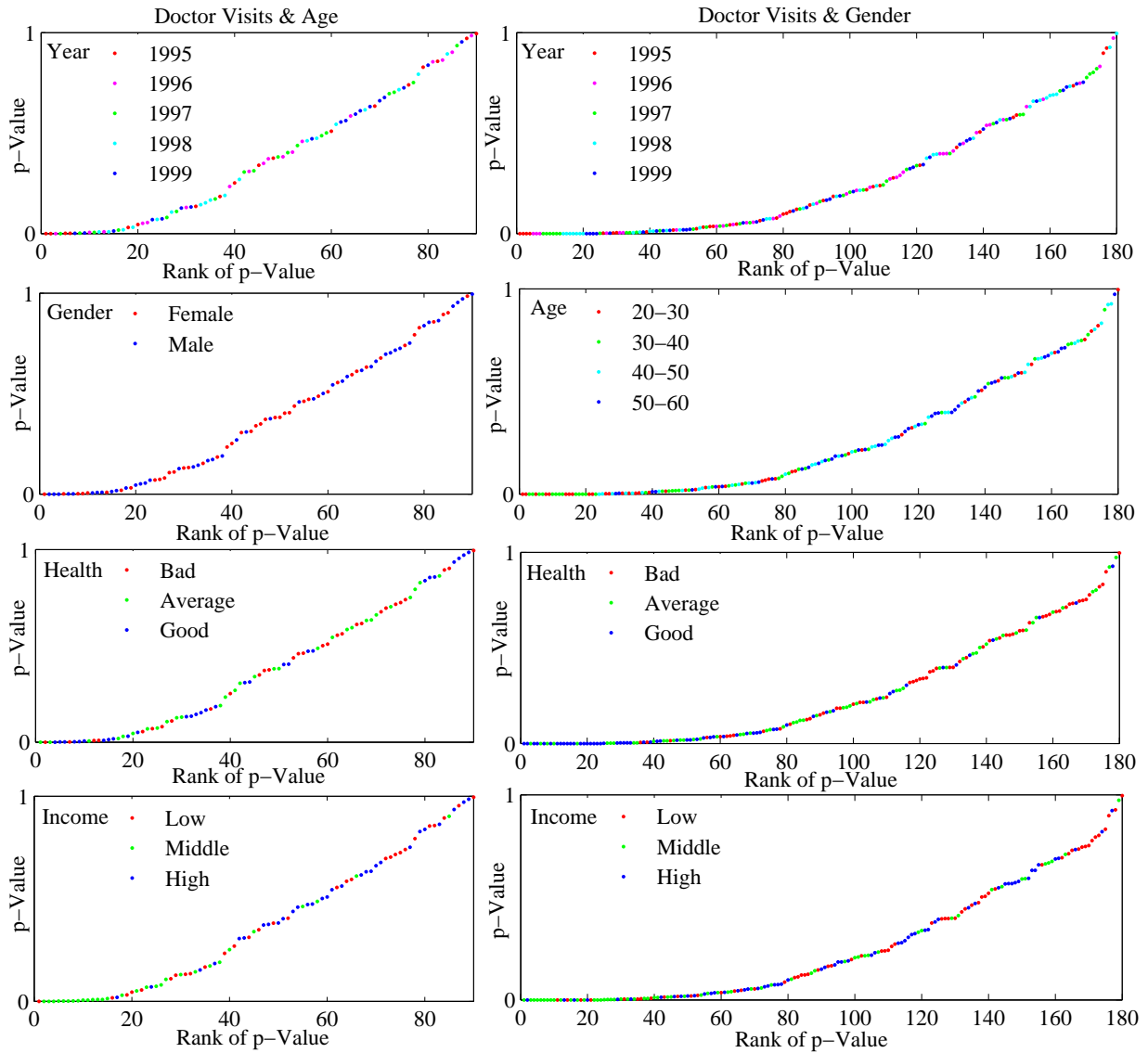


Figure 3.21.: Q-Q plot of p-values on doctor visit data: p-values of hypothesis tests on the relation between doctor visits and age (plots in left column), and between doctor visits and gender (plots in right column). Different colors indicate different subgroups.

4. From Independence Relations to Causal Structure

Having introduced the kernel-based tool of independence test, we move to the task of learning the causal structure based on obtained independence relations. At a first glance, it would be straightforward to incorporate the “oracle” (not necessarily kernel-based) which tells us of the independence into the schema of the IC algorithm. In practice, the oracle does not always work correctly, since we do not have direct access to the true population distribution and can only do inference based on finite data points. In this chapter, we will elaborate on the question how to get structural information by independence constraints, which might exhibit conflicts.

4.1. Logic of independence relations in DAG

Ideally, a good causal model should perfectly represent the underlying probability distribution of observed data. The term “perfectly” means that every independence relation induced by the causal structure, i.e., a DAG \mathcal{G} , is true in the underlying distribution P , and every independence relation in the distribution P is induced by the topological property, i.e., d-separation, of \mathcal{G} . In other words, we search for a faithful Bayesian network (\mathcal{G}, P) (see Section 1.4), which satisfies both Markov and faithfulness conditions.

In practice, the set of all possible independence relations obtained by some test could be incompatible, in the sense that there is no faithful Bayesian network, whose corresponding DAG represents all independence relations. To find a principled way of handling this problem, we take a closer look at the link between independence relations and DAGs.

Some logical rules as shown in Fig. 4.1 are exploited explicitly in a probabilistic graphical model, where independence is captured by d-separation in DAGs [154, 124, 67]. Note that these rules are merely necessary conditions for a faithful representation of independence relations by DAGs. Note that, to the best of our knowledge, how to complete these rules to a set of sufficient conditions is unknown. The rules (A1)-(A4) characterize all independence assertions that logically follow from a so-called semi-graphoid [126, 124, 162]. Those relations satisfying (A1)-(A5) are called graphoids. The logic rule (A5) does not hold universally [14, 15], but only under additional conditions, e.g., the strict positivity of the probability distribution P , in the sense that $P(X) = 0$ only for $X = \emptyset$,¹ is required by Spohn (see [155], Theorem 4). In addition, if the faithfulness is fulfilled, (A6) and (A7) hold.

¹We consider the general case where every node corresponds to a set of variables (instead of only a single variable), which can be straightforwardly identified with a vectorial variable. X should be generally understood as a set of variables.

(A1) Symmetry:	$(X \perp\!\!\!\perp Y \mid S) \Rightarrow (Y \perp\!\!\!\perp X \mid S).$
(A2) Decomposition:	$(X \perp\!\!\!\perp (Y_1, Y_2) \mid S) \Rightarrow (X \perp\!\!\!\perp Y_1 \mid S).$
(A3) Weak Union:	$(X \perp\!\!\!\perp (Y_1, Y_2) \mid S) \Rightarrow (X \perp\!\!\!\perp Y_1 \mid (S, Y_2)).$
(A4) Contraction:	$(X \perp\!\!\!\perp Y_1 \mid S) \wedge (X \perp\!\!\!\perp Y_2 \mid (S, Y_1)) \Rightarrow (X \perp\!\!\!\perp (Y_1, Y_2) \mid S).$
(A5) Intersection:	$(X \perp\!\!\!\perp Y_1 \mid (S, Y_2)) \wedge (X \perp\!\!\!\perp Y_2 \mid (S, Y_1)) \Rightarrow (X \perp\!\!\!\perp (Y_1, Y_2) \mid S).$
(A6) Weak Transitivity:	$(X \perp\!\!\!\perp Y \mid S) \wedge (X \perp\!\!\!\perp Y \mid (S, Z)) \Rightarrow (X \perp\!\!\!\perp Z \mid S) \vee (Y \perp\!\!\!\perp Z \mid S).$
(A7) Chordality:	$(X \perp\!\!\!\perp Y \mid (Z_1, Z_2)) \wedge (Z_1 \perp\!\!\!\perp Z_2 \mid (X, Y)) \Rightarrow (X \perp\!\!\!\perp Y \mid Z_1) \vee (X \perp\!\!\!\perp Y \mid Z_2).$

Figure 4.1.: Rules that characterize independence assertions that logically follow from the Markov and faithful conditions.

4.2. Conflicts of representing independence relations

As mentioned previously, the independence relations observed from real data are not always logically compatible. For one thing, the oracle which tells us of the conditional independence from finite data does not always work correctly. For the other thing, the assumptions we made, i.e., Markov, faithfulness, acyclicity and no-hidden-common-causes, etc., could be violated in real-world data.

4.2.1. Relevant Independence constraints

An independence constraint (or just constraint) is an independence relation $X \perp\!\!\!\perp Y \mid S$ or a dependence relation $X \not\perp\!\!\!\perp Y \mid S$ with disjoint subsets $X, Y, S \subseteq \mathcal{V}$. The number of all possible constraints is exponential in the number of variables in \mathcal{V} . In practice, we have to restrict ourselves to a subset of constraints. Let us first specify the relevant constraints with respect to an undirected graph (adjacency structure) \mathcal{G} .

Definition 16 (Relevant Constraints with respect to Undirected Graph) *A constraint $X \perp\!\!\!\perp Y \mid S$ or $X \not\perp\!\!\!\perp Y \mid S$ is relevant with respect to an undirected graph \mathcal{G} over \mathcal{V} , if the following conditions are satisfied:*

- (1) $X, Y \subset \mathcal{V}$ are two distinct nodes and $S \subseteq \mathcal{V} \setminus \{X \cup Y\}$ is a set of nodes in \mathcal{G} .
- (2) The conditioning set S satisfies the “necessary path condition”, which states that every node in S occurs on an undirected path between X and Y in \mathcal{G} [158].

As a set of variables, X or Y could be empty. But, as a node, X or Y is non-empty, since nodes representing empty sets are not allowed. As a set of nodes, S could be empty. We call the cardinality of S , i.e., the number of nodes in S , the order of the constraint. The shorthand \mathcal{C}_k depicts the class of constraints of order $k \in \mathbb{N}$. If \mathcal{G} is fully connected, all non-trivial constraints are relevant. But, if some edges can be excluded, e.g., due to marginal independence, the necessary path condition can reduce the number of constraints which need to be considered in exploring structure. The other key point of the relevant constraints is that we intend to restrict ourselves to the constraints of entire nodes, not parts of nodes, as the construction of nodes plays an important role in our method. We will elaborate on this later.

4. From Independence Relations to Causal Structure

<p>(R1) $(X \perp\!\!\!\perp Z S) \wedge (Y \perp\!\!\!\perp Z S) \wedge (X \perp\!\!\!\perp Y S) \Rightarrow (X \perp\!\!\!\perp Y (S, Z)).$ (R2) $(X \perp\!\!\!\perp Z S) \wedge (Y \perp\!\!\!\perp Z S) \wedge (X \perp\!\!\!\perp Y S) \Rightarrow (X \perp\!\!\!\perp Z (S, Y)) \vee (Y \perp\!\!\!\perp Z (S, X)).$ (R3) $(X \perp\!\!\!\perp Y Z_1) \wedge (X \perp\!\!\!\perp Y Z_2) \Rightarrow (X \perp\!\!\!\perp Y (Z_1, Z_2)) \vee (Z_1 \perp\!\!\!\perp Z_2 (X, Y)).$</p>
--

Figure 4.2.: Implications from constraints of a lower order to constraints of a higher order that are induced by the rules in Fig. 4.1.

Speaking of relevant constraints with respect to a DAG \mathcal{G} requires a pre-specified definition or construction of nodes in \mathcal{G} . Each node in \mathcal{G} can correspond to a single variable, but in real-world applications, we have also the situation that one node corresponds to a subset of measured variables \mathcal{V} . How to construct meaningful nodes is a non-trivial problem. Since the construction of nodes in \mathcal{G} exhibits a clustering of \mathcal{V} , the task can also be understood as a kind of causally meaningful clustering of \mathcal{V} . We will propose a constraint-based approach to exploring such a clustering in association with structural learning in Section 4.3.

Now, we elaborate on the implications among constraints of different orders induced by a faithful Bayesian network. The semi-graphoid rules (A1)-(A4) in Fig. 4.1 are satisfied by every probability distribution, although the results of independence tests in practice are not necessarily consistent with them. We consider only rules (A5)-(A7). Keeping the goal of structural learning in mind, we rephrase them into three corresponding rules (R1)-(R3) as shown in Fig. 4.2 to clarify the implications from constraints of a lower order to constraints of a higher order. More precisely, rule (R1) or (R2) states a general way to get constraints of class \mathcal{C}_{k+1} from constraints of class \mathcal{C}_k , where k denotes the cardinality of S . Rule (R3) describes a logical implication from constraint class \mathcal{C}_1 to constraint class \mathcal{C}_2 . The derivation of (R1)-(R3) from (A5)-(A7) will be apparent in the following sections.

Corresponding to (R1)-(R3), three conflicting situations can occur in real-world data i.e., the constraints obtained from the empirical independence tests do not follow the logical rules. The next sections will elaborate on these conflicts and propose ways of handling them.

4.2.2. Non-transitivity conflicts

The first conflicting situation, where the weak transitivity property (A6) is violated, can be exemplified by the rats' weight data introduced in Section 3.4.2. The independence constraints between X , Y and Z obtained by the kernel test of independence are

$$(X \perp\!\!\!\perp Z) \wedge (Y \perp\!\!\!\perp Z) \wedge (X \perp\!\!\!\perp Y) \wedge (X \perp\!\!\!\perp Y | Z), \quad (4.1)$$

where X : SEX, Y : DRUG (including drug C), Z : WEIGHT LOSS (see the right half of Tab. 3.9). If X and Y are indeed unconditionally and conditionally independent, rule (A6) implies

$$(X \perp\!\!\!\perp Y) \wedge (X \perp\!\!\!\perp Y | Z) \Rightarrow (X \perp\!\!\!\perp Z) \vee (Y \perp\!\!\!\perp Z).$$

4.2. Conflicts of representing independence relations

The marginal independence between X and Z or between Y and Z displays a conflict with the constraints in Eq. (4.1) obtained by tests.

From another point of view, applying the equivalence $(a \Rightarrow b) \equiv (\neg b \Rightarrow \neg a)$ to the weak transitivity property, the unconditional dependence between X and Z and the marginal dependence between Y and Z imply an unconditional or conditional dependence between X and Y , because

$$(X \not\perp\!\!\!\perp Z) \wedge (Y \not\perp\!\!\!\perp Z) \Rightarrow (X \not\perp\!\!\!\perp Y) \vee (X \not\perp\!\!\!\perp Y | Z).$$

The unconditional or conditional dependence between X and Y contradicts the given constraints in Eq. (4.1) as well.

Definition 17 (Non-transitivity Conflict) *If constraints*

$$(X \not\perp\!\!\!\perp Z | S) \wedge (Y \not\perp\!\!\!\perp Z | S) \wedge (X \perp\!\!\!\perp Y | S) \wedge (X \perp\!\!\!\perp Y | (S, Z)) \quad (4.2)$$

for a triple of distinct nodes $X, Y, Z \subset \mathcal{V}$ and a set of nodes $S \subseteq \mathcal{V} \setminus \{X \cup Y \cup Z\}$ in \mathcal{G} is obtained, then a so-called non-transitivity conflict is present.

In the presence of a non-transitivity, (A6) or (R1) does not hold. Consequently, there are no faithful Bayesian networks representing all these four constraints. The derivation above shows also that the non-transitivity conflict can be resolved if we assume that one of the four constraints is wrong. Now, the question is which one is more likely to be false.

In general, a constraint of low order is more reliable, because testing independence relations with a large conditioning set is more difficult, given a certain number of data points. For this reason, the following assumption of reliable constraints can be reasonably made.

Assumption 4 *Let $(X_1, Y_1), (X_2, Y_2)$ be two pairs of distinct nodes in a DAG \mathcal{G} , $Z \subseteq \mathcal{V} \setminus \{X_1 \cup X_2 \cup Y_1 \cup Y_2\}$ a node in \mathcal{G} , and $S \subseteq \mathcal{V} \setminus \{X_1 \cup X_2 \cup Y_1 \cup Y_2 \cup Z\}$ a set of k nodes in \mathcal{G} . The identification of constraint $X_1 \perp\!\!\!\perp Y_1 | S$ or $X_1 \not\perp\!\!\!\perp Y_1 | S$ is more reliable than the identification of constraint $X_2 \perp\!\!\!\perp Y_2 | (S, Z)$ or $X_2 \not\perp\!\!\!\perp Y_2 | (S, Z)$.*

Note that $X_1 \neq Y_1$ and $X_2 \neq Y_2$, but X_1 and X_2 (or Y_1 and Y_2) could denote the same node. Effectively, we assume that the constraints (of order k) on the left side of (R2) can be more reliably tested than those (of order $k+1$) on the right side, without checking the weak transitivity property explicitly.

Assumption 4 introduces a partial order on constraints and prefers $(X_1 \perp\!\!\!\perp Y_1 | S) \in \mathcal{C}_k$ to $(X_2 \perp\!\!\!\perp Y_2 | (S, Z)) \in \mathcal{C}_{k+1}$, which can be incorporated to resolve the non-transitivity conflict as shown in Eq. (4.2). Consequently, we conclude

$$(X \perp\!\!\!\perp Y | S) \wedge (X \not\perp\!\!\!\perp Y | (S, Z)), \quad (4.3)$$

which indicates a v -structure (Fig. 1.4), i.e., an unshielded collider on Z , according to Step 2 of IC (Fig. 1.5).

In other words, accepting Assumption 4 in addition to Markov and faithfulness assumptions, an unshielded collider on Z can be identified by three constraints of order k (k : number of nodes

4. From Independence Relations to Causal Structure

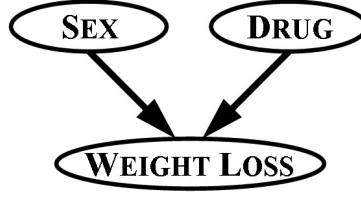


Figure 4.3.: Rats' weight data represented by a DAG, if Assumption 4 is made.

in S)

$$(X \not\perp\!\!\!\perp Z | S) \wedge (Y \not\perp\!\!\!\perp Z | S) \wedge (X \perp\!\!\!\perp Y | S). \quad (4.4)$$

If S is empty, we need only marginal constraints (without any conditional constraints)

$$(X \not\perp\!\!\!\perp Z) \wedge (Y \not\perp\!\!\!\perp Z) \wedge (X \perp\!\!\!\perp Y) \quad (4.5)$$

to identify v -structures in the DAG. Based on this consideration, under Assumption 4 the constraints

$$(\text{SEX} \perp\!\!\!\perp \text{DRUG}) \wedge ((\text{SEX} \not\perp\!\!\!\perp \text{WEIGHT LOSS}) \wedge (\text{DRUG} \not\perp\!\!\!\perp \text{WEIGHT LOSS})),$$

obtained by kernel independence test on the afore-mentioned rats' weight data lead to a v -structure as shown in Fig. 4.3, saying that SEX and DRUG influence the WEIGHT LOSS of rats.

It should be mentioned that a non-transitivity conflict could also be traced back to the fact that the true distribution underlying the real-world data is indeed not faithful. Consider the simplest example with an underlying causal structure $X \rightarrow Z \rightarrow Y$. If causation fails to be transitive, we would then observe the exactly same constraints as shown in Eq. (4.1).

4.2.3. Non-intersection conflicts

Another uncertain situation, which often occurs in real-world data, can be exemplified by the digoxin clearance data already discussed in Section 3.4.1. The independence constraints obtained by kernel tests are

$$(X \not\perp\!\!\!\perp Y) \wedge (X \not\perp\!\!\!\perp Z), \quad (4.6)$$

and

$$(X \perp\!\!\!\perp Y | Z) \wedge (X \perp\!\!\!\perp Z | Y), \quad (4.7)$$

where X : URINE FLOW, Y : DIGOXIN CLEARANCE, Z : CREATININE CLEARANCE (see Tab. 3.8). If we make the Assumption 4 stating that the unconditional constraints can be tested more reliably and thus are true, a non- v -structure (Fig. 1.7) would be reasonable for the set of constraints obtained by tests. The question is how to represent the two conditional independence constraints by a DAG, i.e., remove edge $X - Y$ or remove edge $X - Z$. The two conditional constraints in Eq. (4.7) are indeed not compatible. If we apply the rules of intersection and

4.2. Conflicts of representing independence relations

decomposition to Eq. (4.7), we obtain two marginal independence relations:

$$(X \perp\!\!\!\perp Y | Z) \wedge (X \perp\!\!\!\perp Z | Y) \Rightarrow (X \perp\!\!\!\perp (Y, Z)) \Rightarrow (X \perp\!\!\!\perp Y) \wedge (X \perp\!\!\!\perp Z).$$

The two marginal independence relations induced by the logic rules contradict the constraints in Eq. (4.6). As a note aside, in the real data example of digoxin clearance, the kernel independence test confirmed $X \not\perp\!\!\!\perp (Y, Z)$ with a p-value of 0.007, i.e., urine flow is dependent of clearances.

Consequently, the intersection property (A5) is violated, if $(X \perp\!\!\!\perp Y | Z) \wedge (X \perp\!\!\!\perp Z | Y)$ and one of the marginal dependences, i.e., $X \not\perp\!\!\!\perp Y$ or $X \not\perp\!\!\!\perp Z$, is true. Nonetheless, we need the all marginal dependences between X, Y, Z to be true, i.e., $(X \not\perp\!\!\!\perp Z) \wedge (Y \not\perp\!\!\!\perp Z) \wedge (X \not\perp\!\!\!\perp Y)$, otherwise the necessary path condition for the constraints $(X \perp\!\!\!\perp Y | Z)$ and $(X \perp\!\!\!\perp Z | Y)$ would be not fulfilled according to Definition 16.

Definition 18 (Non-intersection Conflict) *If the constraints*

$$(X \not\perp\!\!\!\perp Z | S) \wedge (Y \not\perp\!\!\!\perp Z | S) \wedge (X \not\perp\!\!\!\perp Y | S) \wedge (X \perp\!\!\!\perp Y | (S, Z)) \wedge (X \perp\!\!\!\perp Z | (S, Y)) \quad (4.8)$$

hold for distinct nodes $X, Y, Z \subset \mathcal{V}$ and a set of nodes $S \subseteq \mathcal{V} \setminus \{X \cup Y \cup Z\}$ in \mathcal{G} are obtained, then a so-called non-intersection conflict is present.

As mentioned previously, the intersection property (A5) does not hold in general. Martín [110] pointed out that the assumption of strict positivity of the joint density, under which the intersection property (A5) is valid [155], is actually too strong and not necessary. He showed that the intersection property only holds, when Y and Z are measurably separated conditionally on S . The so-called “measurable separability” concept is introduced by Florens et al. [57] and provides a sufficient assumption to make the intersection property valid [110].

A trivial example for violation of the intersection property is that Y and Z are related deterministically with each other (see [141, 66, 103] for more theoretical discussions about deterministic relations between nodes), i.e., Y and Z contain entire information about each other. The uncertainty of Y (or Z) vanishes due to the knowledge of Z (or Y), then any node in the graph is independent of Z given Y and independent of Y given Z . Then, testing conditional dependences between X and Y given Z and between X and Z given Y cannot provide any evaluable information about the dependence between X and (Y, Z) . Note that we observed in many real-world applications that the empirical kernel dependence measure $\hat{\mathbb{H}}_{YZ|S}$ is indeed large when non-intersection is present, which indicates, in general, a high degree of dependence between Y and Z given S .

It should be stressed that the situation that Y and Z are deterministically related is a very specific case of non-intersection conflicts defined above. The condition of Eq. (4.8) is substantially weaker, since it needs to hold for merely one node X in the graph. Essentially, it reveals some symmetry of constraints between Y and Z with respect to only one node X .

It is obvious that Assumption 4 cannot help us to prefer one of the constraints

$$(X \perp\!\!\!\perp Y | (S, Z)), (X \perp\!\!\!\perp Z | (S, Y)) \in \mathcal{C}_{k+1},$$

where k is the number of nodes in S . To avoid speculating on which constraint might be more

4. From Independence Relations to Causal Structure



Figure 4.4.: Graphical representation of Digoxin clearance data.



Figure 4.5.: Graphical representation of Rats' weight data.

reliable under some additional restrictive assumptions, we propose to group the nodes Y and Z to a new node representing the vectorial variable (Y, Z) in the model. The intuition behind the grouping strategy is that the new node (Y, Z) shall represent some joint feature of Y and Z , since Y and Z contain some equivalent information with respect to X . All constraints that involve Y or Z , as shown in Eq. (4.8), will not be considered in exploring the structure. Therefore, the grouping strategy is in fact very conservative, in the sense that we do not speculate on the reliability of any of the incompatible constraints.

In the digoxin clearance data, the graphical output would be an undirected graph as shown in Fig. 4.4. The output represents only the fact that urine flow and clearances are dependent. Note that a constraint-based approach, in principle, cannot further specify the orientation of edges between two dependent nodes. Since we intend to interpret the resulting graph causally, the subsequent question is whether the resulting clusters of variables are meaningful. In this real data example, grouping variable “digoxin clearance” and variable “creatinine clearance” is intuitively more meaningful than grouping one of clearances with “urine flow”. The marginal dependence measures, regardless of linear or kernel-based, between variable “digoxin clearance” and variable “creatinine clearance” witnessed the highest degree of mutual dependence in the sample (see Tab. 3.8).

Another example of non-intersection can be found in rats' weight data introduced in Section 3.4.2. The kernel independence tests between variables SEX , $DRUG$ (including drug C), $WEIGHT LOSS$ showed that (right half of Tab. 3.9)

$$(SEX \not\perp\!\!\!\perp WEIGHT LOSS) \wedge (DRUG \not\perp\!\!\!\perp WEIGHT LOSS)$$

and

$$(SEX \perp\!\!\!\perp WEIGHT LOSS \mid DRUG) \wedge (DRUG \perp\!\!\!\perp WEIGHT LOSS \mid SEX).$$

To resolve this non-intersection conflict, Nodes SEX and $DRUG$ should be merged to one node “ $(SEX, DRUG)$ ”, which represents a five-dimensional variable. The resulting graphical representation of rats' weight data is shown in Fig. 4.5.

However, clustering of variables based on the symmetry of independence relations does not necessarily group the variables with the highest degree of mutual dependence. As an example, we

4.2. Conflicts of representing independence relations

MOM5	Maternal	Regulates polarity of the EMS blastomere.
MEX3	Maternal	Specifies the identities of the anterior AB blastomere and its descendants.
POP1	Maternal	Blocks END-1, END-3 activation in mesoderm precursor cells.
PAL1	Maternal	Homeodomain protein, Caudal ortholog.
HLH1	Mesoderm	bHLH transcription factor. Required for proper bodywall muscle development and function.
HND1	Mesoderm	Hand bHLH transcription factor required for normal viability. Expressed in embryonic mesodermal precursor cells generating (mostly) body wall muscles.
PHA4	Mesoderm	FoxA transcription factor. Regulation of pharynx/foregut development.
TBX38	Mesoderm	T box transcription factor. Notch-mediated mesoderm induction in descendants of the ABa blastomere.
HLH25	Mesoderm	Unknown.
END-1	Endoderm	GATA transcription factor. Initiates endoderm differentiation.
END-3	Endoderm	Paralogous to END-1.
ELT-2	Endoderm	GATA transcription factor. Differentiation of the intestine.
ELT-7	Endoderm	Paralogous to ELT-2.

Table 4.1.: Genes and groups involved in *C. elegans* and their function.

consider a small gene regulatory network of endoderm of *Caenorhabditis elegans* (short: *C. elegans*) [16]. The time-lapse gene expression data of the early embryogenesis of *C. elegans* consist of 42 measurements for 13 genes. Genes were manually selected and prior knowledge was used to group the genes into maternally inherited, mesoderm related and endoderm related. Tab. 4.1 summarizes the genes used in the analysis. The dataset consisted of multiple measurements, taken at pc+6min (3), pc+36min (4), fc+0min (3), fc+23min (4), fc+44min (3), fc+53min (3), fc+66min (4), fc+83min (4), fc+101min (3), fc+122min (4), fc+143min (3) and fc+186min (3). The term “pc” indicates pseudo-cleavage and “fc” the four cell stage. The number of measurements at each time point is shown in parenthesis. Each time point coincides with a cell division. The “fc+186min” is approximately at the 200 cell stage, the middle of the gastrulation. For an extensive review of early *C. Elegans* development we refer to [104].

Symmetries of independence relations, like non-intersection conflicts, might be often expected in a gene regulatory network, since different genes could have very similar behaviors within a structure. The *C. elegans* data are real-valued and have a sample size of 42 (without missing values). The measurements of genes involved in endoderm is shown in Fig. 4.6. All non-trivial independence constraints between END-1, END-3, ELT-2, and ELT-7 are tested. The only independence relations obtained by kernel test are the following 4 constraints:

$$(\text{ELT-2} \perp\!\!\!\perp \text{END-1} \mid \text{END-3}), (\text{ELT-2} \perp\!\!\!\perp \text{ELT-7} \mid \text{END-3}) \quad (4.9)$$

4. From Independence Relations to Causal Structure

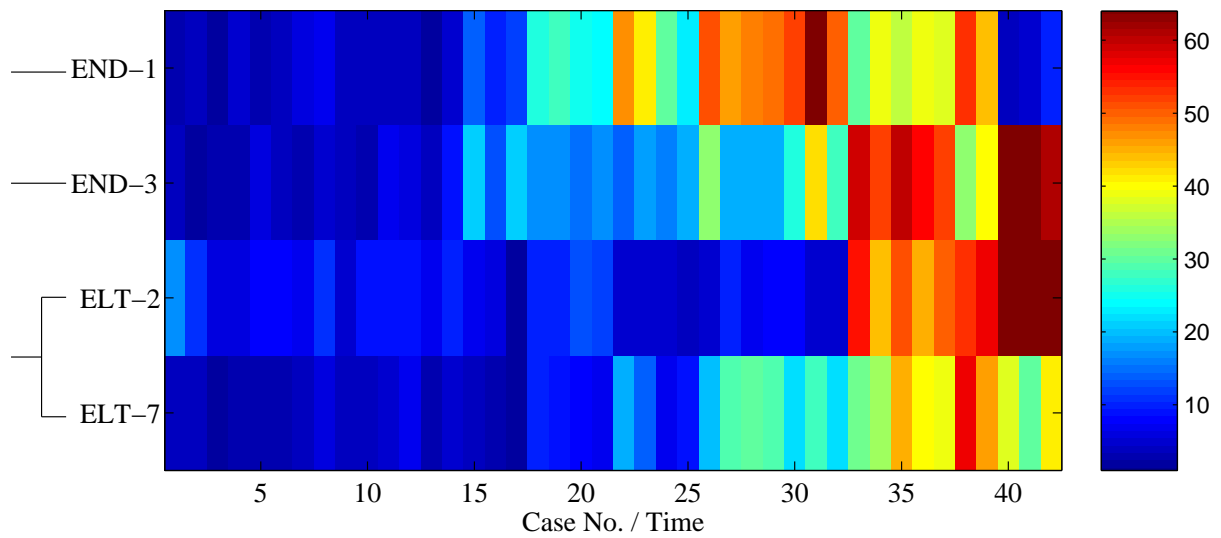


Figure 4.6.: Heatmap of endodermal data of *C. elegans* data with genes END-1, END-3, ELT-2, and ELT-7. The gene names and the clustering results due to resolving a non-intersection conflict (see text) are described on the left side of the plot.

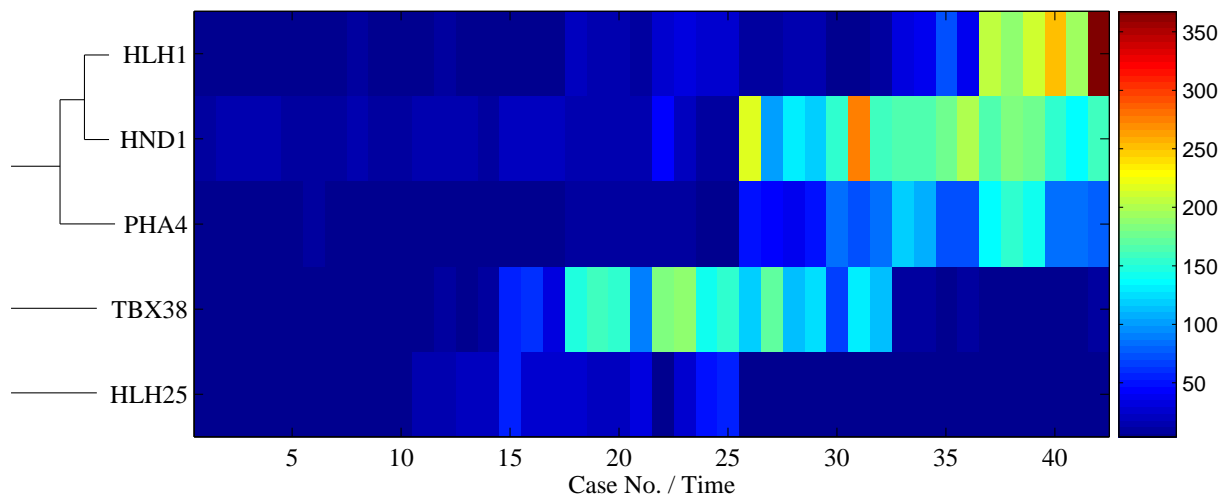


Figure 4.7.: Heatmap of mesodermal data of *C. elegans* with genes HLH1, HND1, PHA4, TBX38, and HLH25. The gene names and the clustering results due to resolving first a non-chordality conflict and then a non-intersection conflict (see text) are described on the left side of the plot.

4.2. Conflicts of representing independence relations

$\hat{\mathbb{H}}_{XY}$	END-3	ELT-2	ELT-7	$\hat{\mathbb{H}}_{XY END-1}$	ELT-2	ELT-7
END-1	0.0414	0.0149	0.0618	END-1	–	–
END-3	–	0.0460	0.0689	END-3	0.0107	0.0103
ELT-2	–	–	0.0470	ELT-2	–	0.0141

Table 4.2.: Empirical kernel dependence measure of genes in endoderm of *C. elegans*.

$\hat{\mathbb{H}}_{YX}$	HND1	HLH25	PHA4	TBX38
HLH1	0.0297	0.0062	0.0280	0.0101
HND1	–	0.0412	0.1305	0.0059
HLH25	–	–	0.0314	0.0158
PHA4	–	–	–	0.0113

Table 4.3.: Empirical kernel dependence measures of genes involved in mesoderm of *C. elegans*.

and

$$(\text{END-3} \perp\!\!\!\perp \text{ELT-2} \mid (\text{END-1}, \text{ELT-7})), (\text{END-3} \perp\!\!\!\perp \text{ELT-7} \mid (\text{END-1}, \text{ELT-2})). \quad (4.10)$$

Other 20 non-trivial constraints exhibit (conditional) dependence between genes.

The constraints in Eq. (4.10) suggest the grouping of genes ELT-2 and ELT-7 to a new node “(ELT-2,ELT-7)” in the final output. The marginal kernel dependence measures (see Tab. 4.2) witnessed a similar degree of the mutual dependence between ELT-2 and ELT-7 and the dependence between END-1 and END-3. However, only ELT-2 and ELT-7 share the symmetric feature with respect to END-3, given END-1. The constraints propose to consider END-1 and END-3 as separate nodes in the final output. Grouping of ELT-2 and ELT-7 is meaningful from the biological viewpoint, as some biologists have already done in their studies [105]. Note that, given END-1, the conditional kernel dependence measure between ELT-2 and ELT-7 is slightly larger than other conditional measures.

4.2.4. Non-chordality conflicts

Rule (A7) in Fig. 4.1 or rule (R3) in Fig. 4.2 describes a case of four separate nodes in a faithful Bayesian network. Situations that are in conflict with this rule can be occasionally found in real data. As an example, we use the data of mesoderm of *C. elegans* [16]. Five genes HLH1, HND1, HLH25, PHA4, and TBX38 are studied in mesoderm. The data are real-valued and have a sample size of 42 (without missing values). The measurements of genes involved in mesoderm is shown in Fig. 4.7. The marginal kernel dependence measures between these five genes are listed in Tab. 4.3.

After testing all possible subsets of four variables, a violation of the rule (A7) or (R3) is found for the set of genes HLH1, HND1, HLH25, and PHA4. We obtained the following dependences

4. From Independence Relations to Causal Structure

by means of the kernel independence test:

$$(\text{HLH25} \not\perp\!\!\!\perp \text{PHA4} \mid \text{HLH1}) \wedge (\text{HLH25} \not\perp\!\!\!\perp \text{PHA4} \mid \text{HND1}).$$

At the same time the following independence were accepted by the kernel test:

$$(\text{HLH25} \perp\!\!\!\perp \text{PHA4} \mid (\text{HLH1}, \text{HND1})) \wedge (\text{HLH1} \perp\!\!\!\perp \text{HND1} \mid (\text{HLH25}, \text{PHA4})).$$

These four constraints exhibit a conflicting situation, because they contradict rule (A7) or rule (R3). If we take a closer look at the constraints, we will see that they reveal a symmetry in HLH1 and HND1 with respect to the dependence between HLH25 and PHA4. Note that it does not necessarily imply that the constraints are also symmetric in HLH25 and PHA4 with respect to the dependence between HLH1 and HND1, as the constraints

$$(\text{HLH1} \not\perp\!\!\!\perp \text{HND1} \mid \text{HLH25}) \wedge (\text{HLH1} \perp\!\!\!\perp \text{HND1} \mid \text{PHA4})$$

were obtained from real data.

Definition 19 (Non-chordality Conflict) *If the constraints*

$$(X \not\perp\!\!\!\perp Y \mid Z_1) \wedge (X \not\perp\!\!\!\perp Y \mid Z_2) \wedge (X \perp\!\!\!\perp Y \mid (Z_1, Z_2)) \wedge (Z_1 \perp\!\!\!\perp Z_2 \mid (X, Y)) \quad (4.11)$$

are obtained, a so-called non-chordality is present.

This kind of conflicts can actually be further divided into three distinct cases, which can be treated in different ways. The first case is that the following constraints are present in addition to the constraints in Eq. (4.11):

$$\text{Non-chordality 1: } (Z_1 \not\perp\!\!\!\perp Z_2 \mid X) \wedge (Z_1 \perp\!\!\!\perp Z_2 \mid Y). \quad (4.12)$$

That means the constraints are only symmetric in Z_1, Z_2 , and not symmetric in X, Y . To resolve this conflicting situation, we propose to group Z_1 and Z_2 to a new node representing the vectorial variable (Z_1, Z_2) . The constraint $X \perp\!\!\!\perp Y \mid (Z_1, Z_2)$ survives after grouping.

Our data example above corresponds exactly to this case. The genes HLH1 and HND1 are suggested to be merged to a new node in the DAG. A ready explanation from biological viewpoint is that genes HLH1 and HND1 are both involved in the specification of muscle cell fates, as opposed to the genes HLH25, PHA4, TBX38.

Other situations are that the constraints containing Z_1 and Z_2 are indeed symmetric in X and Y . One case is

$$\text{Non-chordality 2: } (Z_1 \not\perp\!\!\!\perp Z_2 \mid X) \wedge (Z_1 \not\perp\!\!\!\perp Z_2 \mid Y), \quad (4.13)$$

and the other case is

$$\text{Non-chordality 3: } (Z_1 \perp\!\!\!\perp Z_2 \mid X) \wedge (Z_1 \perp\!\!\!\perp Z_2 \mid Y), \quad (4.14)$$

4.2. Conflicts of representing independence relations

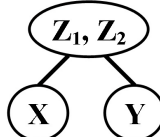
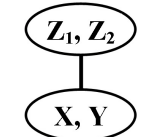
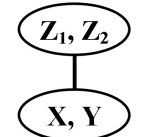
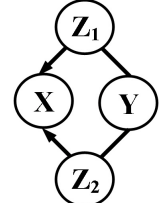
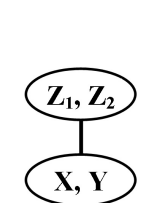
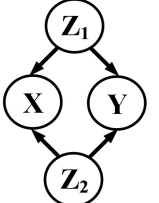
Basic Constraints	$X \not\perp Y Z_1 \quad X \perp Y (Z_1, Z_2)$ $X \not\perp Y Z_2 \quad Z_1 \perp Z_2 (X, Y)$		
Additional Constraints	$(Z_1 \not\perp Z_2 X)$ $(Z_1 \perp Z_2 Y)$	$(Z_1 \not\perp Z_2 X)$ $(Z_1 \not\perp Z_2 Y)$	$(Z_1 \perp Z_2 X)$ $(Z_1 \perp Z_2 Y)$
Non-chordality	Case 1	Case 2	Case 3
No Assumption			
Assumption 4			

Table 4.4.: Handling non-chordality conflicts by different strategies.

in addition to the constraints in Eq. (4.11). In both cases, we propose to merge nodes X and Y to a new node (X, Y) and merge Z_1 and Z_2 to (Z_1, Z_2) . Tab. 4.4 summarizes the strategy for these three cases of non-chordality conflicts. The main advantage of the grouping strategy is that it makes no restrictive assumptions. However, as mentioned previously, the grouping strategy is very conservative, since it does not make any statements about the constraints involved in the conflict.

If we make Assumption 4, the first and the last case of the non-chordality conflicts can be resolved by preferring the constraints $(Z_1 \perp Z_2 | X), (Z_1 \perp Z_2 | Y) \in \mathcal{C}_1$ to $(Z_1 \perp Z_2 | (X, Y)) \in \mathcal{C}_2$, because we actually have a non-transitivity conflict in the first case due to

$$(Z_1 \perp Z_2 | Y) \wedge (Z_1 \perp Z_2 | (X, Y)),$$

and two non-transitivity conflicts in the last case due to

$$(Z_1 \perp Z_2 | X) \wedge (Z_1 \perp Z_2 | Y) \wedge (Z_1 \perp Z_2 | (X, Y)).$$

X, Y and Z_1, Z_2 are symmetric only in the second case. Without giving preference to either of the constraints of the same order 2

$$(X \perp Y | (Z_1, Z_2)) \quad \text{and} \quad (Z_1 \perp Z_2 | (X, Y)),$$

we can only group X, Y and Z_1, Z_2 together to resolve the conflicts (middle column in Tab. 4.4) as proposed above. But in the other two cases, under Assumption 4, we achieve more specific structure (first and last column in Tab. 4.4).

4.3. Constraint-based clustering algorithm

For one thing, we have to cluster variables due to conflicting situations of constraints. For the other thing, the kernel test of independence is able to treat high-dimensional variables in a straightforward way, which makes nodes representing vectorial variables in the DAG possible. Regardless of the possibility of a meaningful interpretation of the node construction in the graph, we first introduce the possible partitions of the set of all measured variables \mathcal{V} , if a causal structure over \mathcal{V} is known.

Definition 20 (Manipulation-consistent Partition) *Let \mathcal{G}_1 be a DAG with m_1 nodes $\mathcal{L}_1 := \{Y_1, \dots, Y_{m_1}\}$, i.e., a partition of the set of measured variables $\mathcal{V} = \{X_1, \dots, X_N\}$. A (coarse-grained) partition $\mathcal{L}_2 := \{Z_1, \dots, Z_{m_2}\}$ of \mathcal{V} with $Z_i \subseteq \{Y_1, \dots, Y_{m_2}\}$ is manipulation-consistent with \mathcal{L}_1 , if there exists a DAG \mathcal{G}_2 with m_2 nodes representing $\{Z_1, \dots, Z_{m_2}\}$, in which $Z_i \rightarrow Z_j$ ($i \neq j$) are present in \mathcal{G}_1 for all arrows $Y_k \rightarrow Y_l$ in \mathcal{G}_1 with $Y_k \in Z_i$ and $Y_l \in Z_j$.*

The coarse-grained ($m_2 \leq m_1$) structure \mathcal{G}_2 is obtained by grouping nodes in \mathcal{G}_1 . If all arrows in the original DAG \mathcal{G}_1 indeed describe the potential effects of manipulation (see Eq. (1.3) for manipulation criterion) between variables, all arrows in the coarse-grained structure \mathcal{G}_2 satisfy the manipulation criterion and could also be interpreted causally.

A node in the causal structure is generally understood as a factor that causally explains the associations measured over a single variable or a group of variables. The motivation of introducing factors represented by multiple variables is that models of complex phenomena often consist of hypothetical entities called “unmeasurable factors”, which cannot be directly measured by some single variable, but might be identifiable by a group of variables that describe different aspects of this unmeasurable factor. Such factors measured indirectly can play an important role in understanding and predicting the dynamics of those phenomena. For instance, in social science, questionnaires are designed to target specific indirect measurements, such as “stress”, “job satisfaction”, and so on.

Silva et al. [144] proposed a formal framework: the so-called generalized measurement models, to represent the unmeasurable factors. They call them “latent factors”. Subsequently, Silva et al. [145] presented a principled way to discover latent factors in linear models. The construction of latent factors provide also a partition of the set of all measured variables. Unlike their approach, we propose to cluster the measured variables from the viewpoint of structural learning. A manipulation-consistent partition of measured variables serves as an appropriate construction of nodes in causal structure. The key point here is that it can occur that a certain partition of variables makes the construction of faithful Bayesian networks possible, while the other partition does not.

To make this apparent, we consider a coarse-grained structure \mathcal{G}_1 as shown in the right plot of Fig. 4.8. It is obtained by grouping distinct nodes Z_1 and Z_2 in a “fine-grained” structure \mathcal{G}_0 (left plot) into a new node representing vectorial variable (Z_1, Z_2) . A faithful Bayesian network with respect to the fine-grained structure implies the relevant constraint $X \perp\!\!\!\perp Y \mid (Z_1, Z_2)$, while a faithful Bayesian network with respect to the coarse-grained structure requires the relevant constraint $X \not\perp\!\!\!\perp Y \mid (Z_1, Z_2)$. Given a probability distribution P , it could occur that the Bayesian network (\mathcal{G}_0, P) is faithful, but (\mathcal{G}_1, P) not, or the other way around. Handling violations of the



Figure 4.8.: A coarse-grained structure \mathcal{G}_1 (right plot) is obtained by merging distinct nodes Z_1 and Z_2 of a fine-grained structure \mathcal{G}_0 (left plot) to a two-dimensional variable $Z := (Z_1, Z_2)$. Both describe consistent manipulation potentials (see Eq. (1.3) for manipulation criterion), however, do not share the same feature with respect to the independence relations between X and Y conditional on (Z_1, Z_2) .

properties implied by faithful Bayesian networks can help us to find an appropriate partition of measured variables.

Having the structural learning in mind, we formulate the variable partitioning problem: given a set of variables, partition the variables into clusters (nodes) to make a representation of data by a faithful Bayesian network possible. In many real-world applications, variables are reasonably pre-specified through experimental design, so that the trivial clustering of variables, i.e., each distinct cluster corresponds to a single variable, is already a meaningful initial construction of nodes. Sometimes, the prior knowledge, i.e., the meaning of variables, can help us to construct nodes. Nevertheless, we come up with conflicts in exploring a faithful Bayesian network representing the distribution. As showed above, we group variables by using some symmetric properties of independence constraints without making restrictive assumptions. Fig. 4.9 summarizes the so-called constraint-based clustering procedure. Our method utilizes the property of independence relations of triples, instead of using dependence measures between pairs of variables. This differs from the approach taken by standard clustering algorithms and especially from the recent work by Song et al. [149]. The procedure can be initially started with the trivial clustering of variables, i.e., each node corresponds to a single variable. In real-world applications, prior knowledge, e.g., the meaning of measured variables, can also be helpful to determine the initial clustering of variables. It is obvious that the clustering algorithm converges, because after each iteration the cardinality of S is increased by 1 or the number of nodes is decreased by 1. The algorithm converges fast, if the graph is sparse, since less fully connected triples need to be checked. The output of the procedure is an appropriate clustering of variables depicting nodes, in the sense that there are no non-intersection and no non-chordality conflicts between constraints of orders up to some pre-specified integer k .

4.4. Constraint-based orientation algorithm

The identification of v -structures is the essential strategy of constraint-based approaches. Under Assumption 4, we use the constraints in Eq. (4.4) to identify v -structures. This inference rule

4. From Independence Relations to Causal Structure

Input: A set of N nodes and an integer k .

Step 1: If a set of four distinct nodes X, Y, Z_1, Z_2 can be found that the condition in Eq. (4.11) is satisfied, group Z_1 and Z_2 to a new variable (Z_1, Z_2) and set $N := N - 1$. If the condition in Eq. (4.13) or Eq. (4.14) is additionally satisfied, group X and Y to a new variable (X, Y) and set $N := N - 1$. Repeat step 1 so long as the set of nodes does not change.

Step 2: For $i=0$ to $\min\{k, N-3\}$, if a triple of distinct nodes X, Y, Z and a set of i nodes S (not including X, Y, Z) can be found that the condition in Eq. (4.8) is satisfied, group Y and Z to a new node (Y, Z) and set $N := N - 1$. If the set of nodes changes, goto Step 1, otherwise continue.

Output: A set of $N' \in [2, N]$ nodes.

Figure 4.9.: Constraint-based clustering procedure.



Figure 4.10.: Orientation using only marginal independence relations as shown in Eq. (4.15). The bi-directed edge in the left plot is traced back to a collider on X_2 and a collider on Y_2 and represents the conflicting information of orientation obtained by v -structure identification. If acyclicity is assumed, a structure with latent common cause L is explanation for the dependence between X_2 and Y_2 (right plot).

for v -structures has consequences for the learning of the whole structure. For instance, if the marginal independence constraints

$$X_1 \perp\!\!\!\perp Y_2 \wedge X_2 \perp\!\!\!\perp Y_1 \wedge X_2 \perp\!\!\!\perp Y_2 \wedge X_1 \not\perp\!\!\!\perp Y_1 \wedge X_1 \not\perp\!\!\!\perp X_2 \wedge Y_1 \not\perp\!\!\!\perp Y_2 \quad (4.15)$$

are obtained, we infer two v -structures $X_1 \rightarrow X_2 \leftarrow Y_2$ and $X_2 \rightarrow Y_2 \leftarrow Y_1$ due to Eq. (4.5). Graphically, a collider on X_2 and a collider on Y_2 as exemplified in the left plot of Fig. 4.10. The bi-directed edge between X_2 and Y_2 represents the conflicting information of orientation obtained by v -structure identification. The resulting structure violates the assumption of acyclicity. Making the assumption of acyclicity, both colliders leave the existence of a latent common cause as the only explanation for the observed dependence between X_2 and Y_2 (Fig. 4.10, right).

Combining the orientation information obtained by the inference rule as in Eq. (4.4) for all possible triples (X, Y, Z) and all conditioning sets S , we can learn the orientation of the whole structure. We implement this idea by a voting procedure, i.e., an identified v -structure $X \rightarrow Z \leftarrow$

Number of all non-trivial Marginal Independence Hypotheses accepted by Tests												
0	1	2	3	0	1	2	3	4	5	6	...	
												...
Patterns of 3 nodes				Patterns of 4 nodes							...	

Table 4.5.: Possible patterns of 3 or 4 nodes when only marginal independence relations are known. The extension for structures of more than 4 nodes is straightforward.

Y gives a respective vote to $X \rightarrow Z$ and $Z \leftarrow Y$. Inconsistent voting results will be represented by a bi-directed edges. Using this voting procedure, a patten can be found to represent all marginal dependence relations. Tab. 4.5 illustrates the resulting patterns of 3 and 4 nodes. A 3-node structure has 6 non-trivial constraints, and 3 of them are marginal. A 4-node structure has 24 constraints, and 6 of them are marginal. Surprisingly, in many cases, most edges can be already oriented by just using marginal independence constraints, even though we used only 50% (3-node structures) and 25% (4-node structures) of all non-trivial constraints between variables. Actually, we do not need many constraints to infer structures, if the constraints are consistent. However, the highly redundant set of all possible constraints could lead to many conflicting situations, which makes structural learning unreliable. A reasonable assumption like Assumption 4 is desirable. That means the constraints of small order should be preferred.

We propose a constraint-based orientation procedure as shown in Fig. 4.11. This procedure follows the strategy that the marginal constraints, i.e., the constraint class \mathcal{C}_0 , should first be considered and represented by the DAG. If no independence is obtained within the marginal constraints, we obtain the fully connected and undirected structure. In this situation, constraint class \mathcal{C}_1 would be taken into account to further detect orientation. If it fails, $\mathcal{C}_2, \mathcal{C}_3, \dots, \mathcal{C}_{N-1}$ could be successively considered, where N the number of nodes involved in the fully connected and undirected structure. Typically, we choose the fully connected undirected graph \mathcal{G} of N nodes as the initial structure. The parameter k can be typically set to $N - 3$.

4.5. Robust causal learning algorithm (RCL)

Combining the clustering and orientation procedures, we propose the so-called robust causal learning (RCL) algorithm as shown in Fig. 4.12 to find a faithful Bayesian network representing observed data. The term “robust” refers to the strategy that we start with constraints of lower orders and construct a structure representing as many reliable and compatible constraints as possible. The result of this strategy is expected to be robust with respect to statistical fluctuations of small samples. The user can directly bound the order of constraints that to be considered from

4. From Independence Relations to Causal Structure

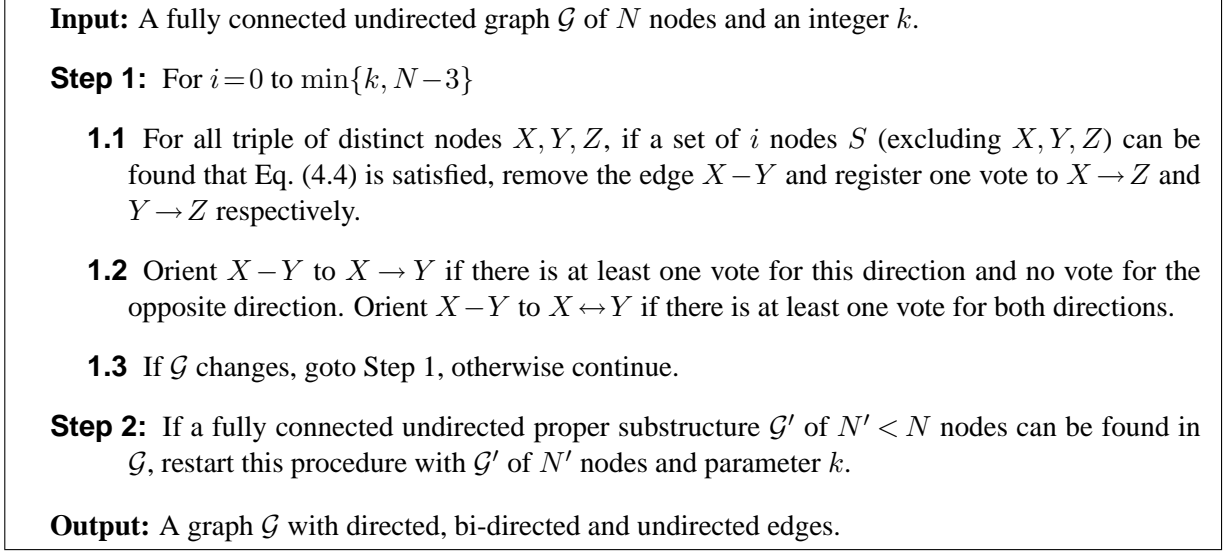


Figure 4.11.: Constraint-based orientation procedure.

above by some pre-specified integer $k \leq N-3$. The most general choice of k is $N-3$.

RCL starts with a fully connected graph. Based on the information from all constraints of class \mathcal{C}_0 , Step 1.1 learns a partially directed graph through v -structure identification. After that, Step 1.2 searches for non-intersection conflicts with respect to the underlying adjacency structure. If two nodes are merged to a new node, RCL will be restarted with the new set of fewer nodes, otherwise, the constraints of class \mathcal{C}_1 will be considered to infer orientation for the remaining undirected substructures. Having taken all constraints of order up to k into account or having oriented all edges involved in the graph, Step 2 of RCL checks for non-chordality conflicts. Since Assumption 4 is made anyway by RCL, only the special case as specified by Eq. (4.13) should be treated. Step 3 removes unnecessary edges with respect to the topology of the graph learned by previous steps. For this purpose, we introduce the so-called relevant constraints with respect to a given directed Graph with uni-, and bi-directed edges.

Definition 21 (Relevant Constraints with respect to Directed Graph) A constraint $X \perp\!\!\!\perp Y \mid S$ is relevant with respect to a directed graph \mathcal{G} over \mathcal{V} , if the following conditions are satisfied:

- (1) $X, Y \subset \mathcal{V}$ are two distinct nodes and $S \subseteq \mathcal{V} \setminus \{X \cup Y\}$ is a set of nodes in \mathcal{G} .
- (2) The conditioning set S satisfies the “potential ancestor condition”, which states that every node Z in S is potential ancestor of X or Y in \mathcal{G} , i.e., there exists at least one directed path from Z to X or Y .

The motivation is that, if X and Y are connected in a directed graph, only conditioning on potential ancestors of X or Y can make them independent. An edge $X-Y$ in \mathcal{G} is removed by a constraint $X \perp\!\!\!\perp Y \mid S$, if the constraint satisfies the potential ancestor condition. In comparison to the necessary path condition, the potential ancestor condition takes additionally the orientation of \mathcal{G} into account. The number of queries to the independence oracle can be reduced. Since our

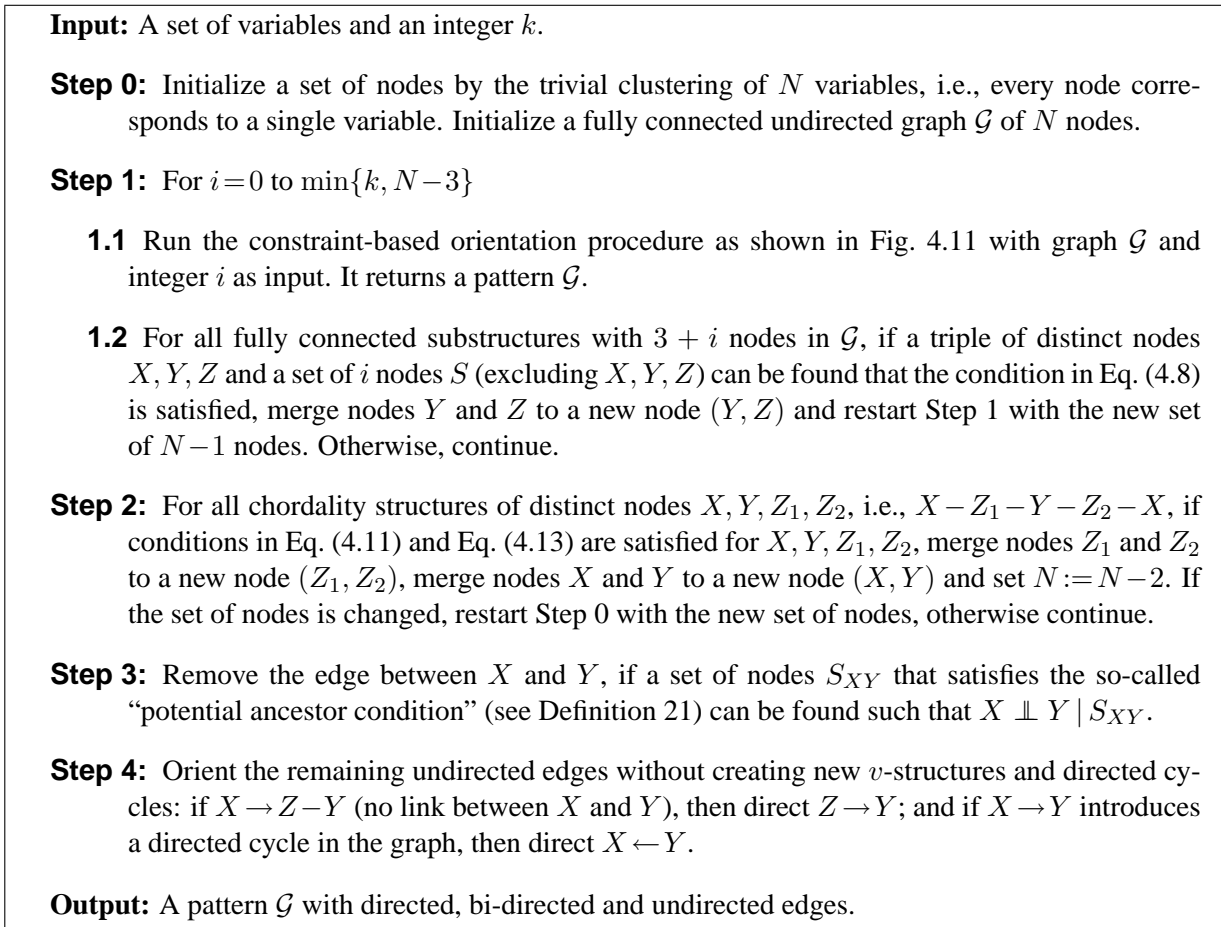


Figure 4.12.: Robust causal learning algorithm (RCL).

4. From Independence Relations to Causal Structure

mixed graph \mathcal{G} could contain undirected edges we consider all undirected edges as bi-directed ones to identify potential ancestors. Step 4 orients the remaining undirected edges under the assumption that all v -structures are identified by previous steps and the underlying structure is acyclic. The output of RCL is a pattern with uni-, bi- and un-directed edges. The bi-directed edges could be traced back to a violation of the assumption of no hidden-common-causes or acyclicity.

RCL explores the non-intersection conflicts after resolving non-transitivity conflicts. That means, if a triple that satisfies the condition of both non-transitivity and non-intersection, RCL will orient the triple to a v -structure under Assumption 4. The rats' weight data is an example (see discussions in Section 4.2.2 and Section 4.2.3). The output of RCL is then the structure as shown in Fig. 4.3, and not that in Fig. 4.5.

Obviously, the computational complexity of RCL depends on the number of constraints that are tested. RCL has to test all non-trivial constraints, if no conditional independence can be obtained (worst-case scenario). The number of independence constraints increases exponentially with respect to the number of nodes. Therefore, RCL is only computationally feasible if there exists a sparse structure representing the data. The more conditional independence relations occur in the low-order constraints, the faster RCL converges.

If there exists a faithful Bayesian network with the trivial clustering of variables representing the data, i.e., very node corresponds to a single variable, RCL coincides with the IC algorithm and will find it. However, if the constraints obtained from data are highly incompatible, the resulting structure could be less informative due to the strategy of merging nodes. Note that the construction of nodes in the final output, i.e., the clustering of variables, is not always unique, due to different orders of merging nodes (see Section 4.6.4 for an example).

4.6. Real-world Experiments with RCL

We demonstrate some experiments of real-world data with RCL from different scientific fields. In our experiments, if not explicitly stated otherwise, the kernel independence test is used due to its general applicability. The real-world data are challenging, because the assumptions we made, e.g., acyclicity, faithfulness, etc., are not necessarily fulfilled. There might exist no faithful Bayesian network at all to represent the observed data. However, RCL often generates faithful structure on an appropriate clusters of variables.

4.6.1. College plans

Sewell et al. [140] investigated factors that influence the intention of high school students to attend college. They measured five variables for 10,318 Wisconsin high school seniors: (SEX): male, female; Socio-economic Status (SES): low, lower middle, upper middle, high; Intelligence Quotient (IQ): low lower middle, upper middle, high; Parental Encouragement (PE): low, high; and College Plans (CP): yes, no. This dataset is already discussed by Spirtes et al. [153] (with constraint-based PC algorithm) and by Heckerman et al. [86] (with the Bayesian score-based method).

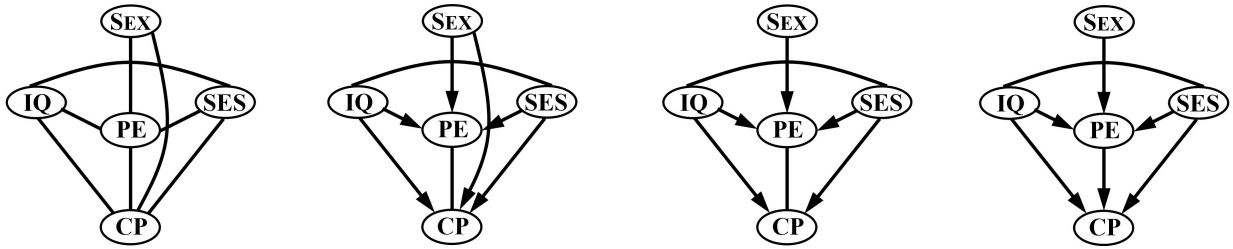


Figure 4.13.: Stepwise results of RCL (using χ^2 test) applied to college plan data.

We ran RCL on this dataset. Because of the large sample size, we used likelihood ratio χ^2 tests to check the independence constraints. Fig. 4.13 illustrates the stepwise results of RCL. We started RCL with testing marginal independence relations and obtained two marginal independence constraints, namely $SEX \perp IQ$ and $SEX \perp SES$. This leads to an adjacency structure as shown in the leftmost plot. The variables will not be merged, since no non-intersection or non-chordality conflicts were detected. The second plot shows the resulting directed graph according to the orientation procedure described in Fig. 4.11. After that, due to $SEX \perp CP \mid PE$, the edge between SEX and CP is unnecessary and thus removed, as shown in the third plot.

The remaining undirected edge between PE and CP can be oriented in $PE \rightarrow CP$ by Step 4 of RCL, otherwise a new v -structure $SEX \rightarrow PE \leftarrow CP$ would be created. The edge between IQ and SES remains undirected. The final output of RCL (rightmost plot) coincides with the result of the constraint-based PC algorithm (see [153] for discussions), but slightly differs from the result of the score-based Bayesian approach (see [86] for discussions). Additional χ^2 tests showed that all detected unshielded colliders on PE and CP display no non-transitivity conflicts. Thus, this final output is a faithful Bayesian network that perfectly represents the data. This example shows that if there indeed exists a faithful Bayesian network with the trivial clustering of variables, i.e., very node in graph corresponds to a single variable, RCL works like IC and find the faithful representation.

4.6.2. Egyptian skulls

This dataset [164] consists of four measurements of male Egyptian skulls from five different historical periods ranging from 4000 B.C. to 150 A.D. 30 skulls are measured from each time period, i.e., 150 cases in total. The data are analyzed to determine if there are any differences in the skull sizes between the time periods and if they show any changes with time. The researchers theorize that a change in skull size over time is evidence for the interbreeding of the Egyptians with immigrant populations over the years. The measurements of skulls are MB (maximal breadth), BH (basibregmatic height), BL (basialveolar length), NH (nasal height). The predictor variable is APPROXIMATE YEAR (approximate year of skull formation).

RCL converges after testing the marginal constraints, i.e., those in \mathcal{C}_0 . The output as shown in Fig. 4.14 represents all marginal constraints, in particular the marginal independence between some of the measurements, e.g., between MB and BL and between BH and NH.

4. From Independence Relations to Causal Structure

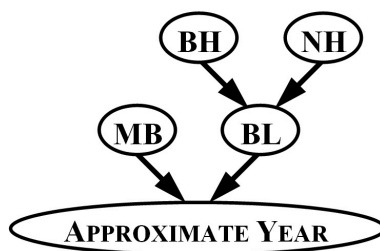


Figure 4.14.: Output of RCL on Egyptian skull data.

The two v -structures, which are identified by marginal constraints, is not confirmed by conditional dependences. That means, two non-transitivity conflicts are present and resolved by Assumption 4. One possible reason is that the underlying distribution is indeed not faithful. This dataset is listed as an example for the software project TETRAD (containing the PC algorithm) on its webpage http://www.phil.cmu.edu/projects/tetrad_examples. The output of RCL is consistent with the output of PC. We used default parameters of TETRAD 4.3.8 and set significance level $\alpha = 0.05$.

Note that the causal ground truth is actually not quite clear in this example. If we distinguish between the real year (yet unknown) and the estimated year (based on the size variables: MB, BH, BL and NH) the former should be considered as a cause of the size variables (given that the skull size has indeed changed over the years) and the latter as an effect of them. Due to the fact that the kernel independence test detected marginal independence between MB, BH, and between MB and NH, it is not very plausible to assert, based on this dataset, a really significant change in skull size over time.

4.6.3. Montana outlook poll

The data contain the outcomes in the Montana Economic Outlook Poll conducted in May 1992, with accompanying demographics for 209 out of 418 poll respondents. After removing records with missing values, the dataset has 163 entries. More information about data can be found at <http://lib.stat.cmu.edu/DASL/Stories/montana.html>. The Montana poll asked a random sample of Montana residents whether the respondent feels his/her personal financial status is worse, the same, or better than a year ago, and whether they view the state economic outlook as better over the next year. Respondents are classified by age, income, political orientation, and area of residence in the state.

The dataset contains the following 7 discrete variables: AGE = 1 meaning under 35, 2 meaning 35 to 45, 3 meaning 55 and over; SEX = 1 meaning male, 2 meaning female; yearly INCOME = 1 meaning under \$20K, 2 meaning \$20 – 35K, 3 meaning over \$35K; POLITICAL = 1 meaning Democrat, 2 meaning Independent, 3 meaning Republican; AREA = 1 meaning Western, 2 meaning Northeastern, 3 meaning Southeastern Montana; FINANCIAL status = 1 meaning worse, 2 meaning same, 3 meaning better than a year ago; state economic OUTLOOK = 1 meaning better, 2 meaning not better than a year ago. We interpret the values numerically, since the difference of

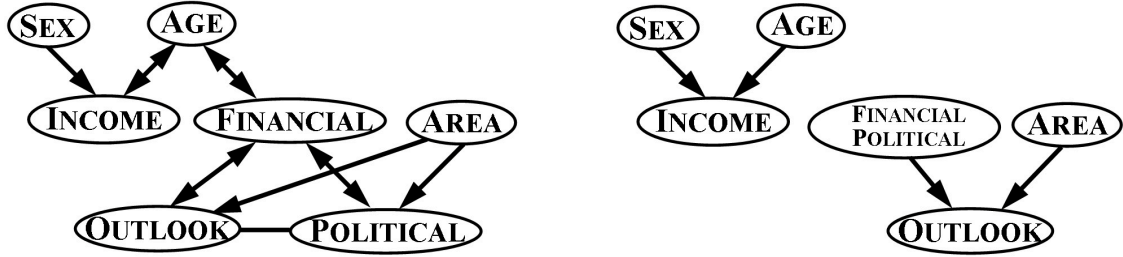


Figure 4.15.: Stepwise results of RCL on Montana data. The left plot illustrates the structure representing marginal constraints with a non-intersection conflict. The right plot illustrates the final output of RCL. The output is a faithful Bayesian network representing the independence relations obtained by the kernel independence test.

values are somewhat meaningful.

We ran kernel independence tests. The left plot of Fig. 4.15 shows the adjacency structure based on marginal constraints. The fully connected triples $(\text{FINANCIAL}, \text{POLITICAL}, \text{OUTLOOK})$ and $(\text{AREA}, \text{POLITICAL}, \text{OUTLOOK})$ are checked for non-intersection conflicts. The following conditional independence is obtained for the former triple:

$$(\text{OUTLOOK} \perp\!\!\!\perp \text{FINANCIAL} \mid \text{POLITICAL}) \wedge (\text{OUTLOOK} \perp\!\!\!\perp \text{POLITICAL} \mid \text{FINANCIAL}).$$

Therefore, Step 1.2 of RCL merged FINANCIAL and POLITICAL together to a new node containing both. Due to

$$(\text{SEX} \perp\!\!\!\perp \text{AGE}) \wedge (\text{SEX} \not\perp\!\!\!\perp \text{INCOME}) \wedge (\text{AGE} \not\perp\!\!\!\perp \text{INCOME}),$$

and

$$((\text{FINANCIAL}, \text{POLITICAL}) \perp\!\!\!\perp \text{AREA}) \wedge ((\text{FINANCIAL}, \text{POLITICAL}) \not\perp\!\!\!\perp \text{OUTLOOK}) \wedge (\text{AREA} \not\perp\!\!\!\perp \text{OUTLOOK}),$$

we infer two v -structures (see Fig. 4.15, right). Both v -structures can be confirmed by the conditional dependences via the kernel independence test

$$(\text{SEX} \not\perp\!\!\!\perp \text{AGE} \mid \text{INCOME}) \wedge ((\text{POLITICAL}, \text{FINANCIAL}) \not\perp\!\!\!\perp \text{AREA} \mid \text{OUTLOOK}).$$

Thus, the output of RCL as shown in the right plot of Fig. 4.15 is indeed a perfect map of data, i.e., a faithful Bayesian network.

In this example, we obtained the following constraints by means of the kernel independence test:

$$\text{FINANCIAL} \perp\!\!\!\perp \text{AREA} \quad (\text{with a p-value of } 0.013) \quad (4.16)$$

and

$$\text{POLITICAL} \not\perp\!\!\!\perp \text{AREA} \quad (\text{with a p-value of } 0.504) \quad (4.17)$$

4. From Independence Relations to Causal Structure

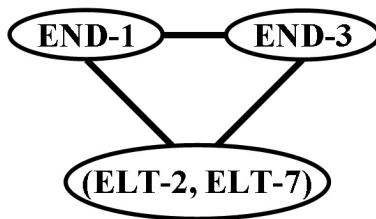


Figure 4.16.: Output of RCL on endodermal data of *C. elegans*. The fully connected undirected graph represents 6 DAGs ($3! = 6$ different orderings of 3 nodes).

but

$$(\text{POLITICAL, FINANCIAL}) \perp\!\!\!\perp \text{AREA} \quad (\text{with a p-value of } 0.057).$$

It is obvious that the decomposition property (A2) as defined in Fig. 4.1 is violated, when the significance level is chosen to be 0.05. Nonetheless, the constraints as shown in Eq. (4.16) and Eq. (4.17) are not required by the resulting structure due to Definition 16, since only independence relations between entire nodes, not parts of a node, are considered. That is why (A2) or (A3) need not be considered in this thesis. To modify our kernel measures so that the properties of decomposition (A2) and weak union (A3) are inherently fulfilled is an interesting line of further research. Note that constraints via our kernel independence test always satisfy the property of symmetry (A1) due to the design of the kernel measures described in Section 2.3 and Section 2.4.

4.6.4. *Caenorhabditis elegans*

Biological regulatory networks appear to be composed of small, function-centered regulatory sub-networks in which most of the regulation is exhibited between a small number of highly interactive genes, with only limited input from the rest of the network. Therefore, it is interesting to explore the relationships between a small number of genes. However, discovering biological regulatory networks is challenging, because such applications concern small sample sizes and noisy data.

In this experiment, we study the small gene regulatory networks of *C. elegans* again (see Section 4.2.3 for data). First, we consider the endoderm in *C. elegans*. Having resolved the non-intersection conflict by merging the genes ELT-2 and ELT-7 due to Eq. (4.10) to one node, we obtained a structure of three nodes without non-trivial (conditional) independence relations. The final output of RCL is then the fully connected undirected graph as shown in Fig. 4.16 representing the Markov equivalence class of DAGs ($3! = 6$ different orderings of 3 nodes).

The resulting structure generated for the endodermal data by RCL is not very informative, but is consistent with our current understanding of the roles of these genes. END-1 and END-3 both belong to the GATA family of transcription factors and are the earliest endoderm specific genes expressed [184]. Evidence points to END-3 being activated first with END-1 following shortly after. Both subsequently trigger the expression of ELT-2 and ELT-7, GATA factors them-

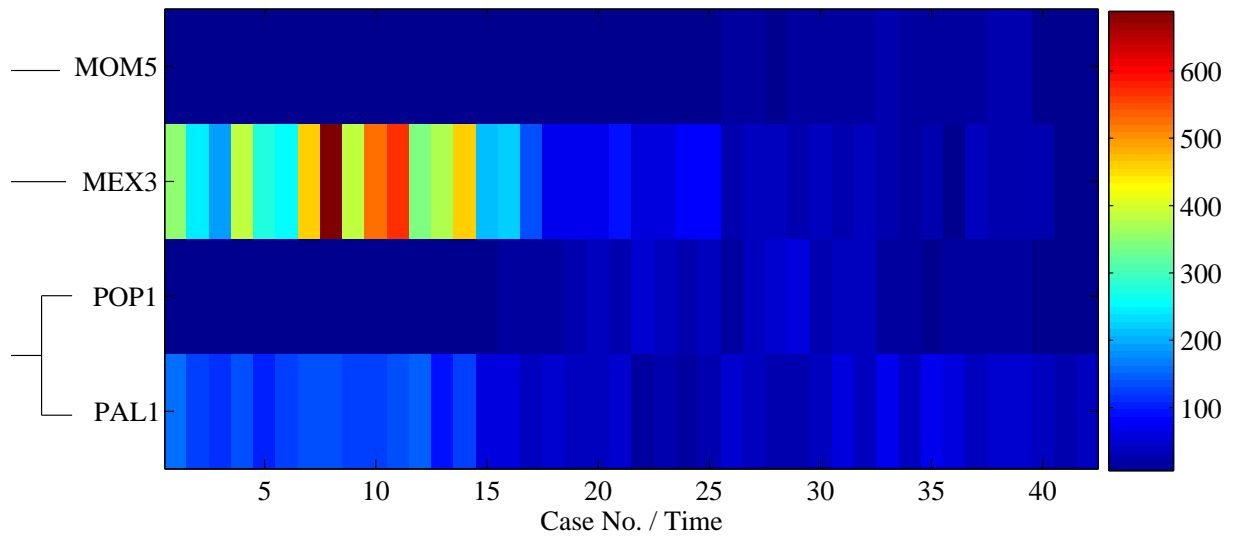


Figure 4.17.: Heatmap of maternal data of *C. elegans* with genes MOM5, MEX3, POP1, and PAL1. The gene names and the clustering results due to resolving a non-intersection conflict (see text) are described on the left side of the plot.

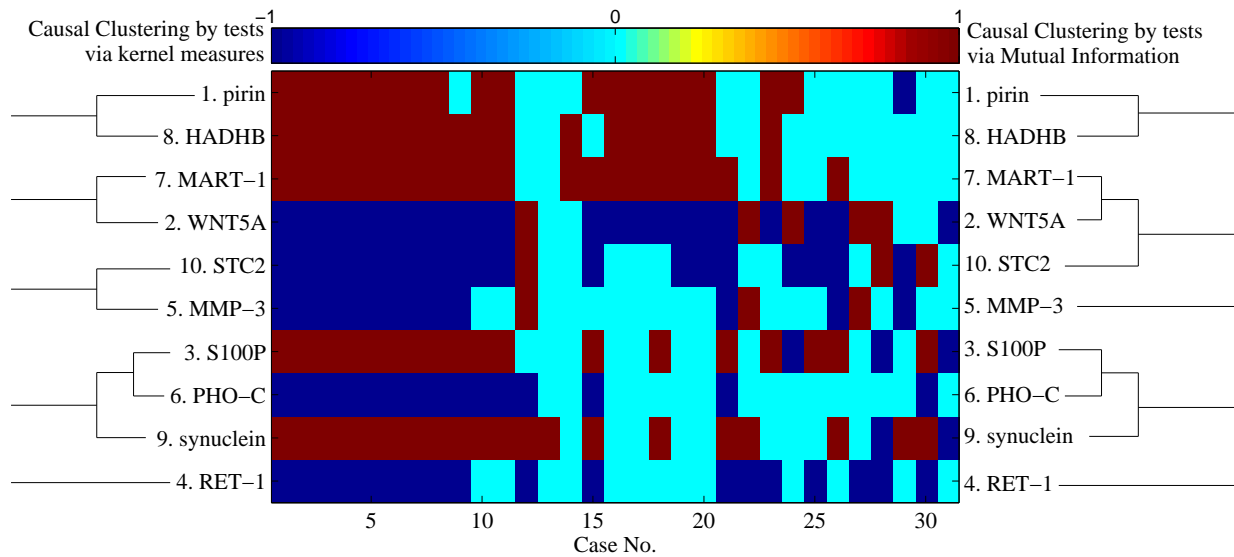


Figure 4.18.: Heatmap of 10 genes of metastatic melanoma data. The gene names and the clustering results by means of independence tests via kernel measures and mutual information (see text) are showed on the left and the right side of the plot respectively.

4. From Independence Relations to Causal Structure

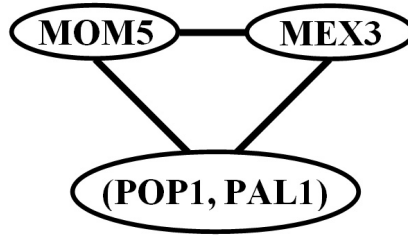


Figure 4.19.: Output of RCL on maternal data of *C. elegans*.

selves [105].

Second, we consider another small gene regulatory network of *C. elegans*, namely maternal. The measurements of genes are illustrated in Fig. 4.17. Due to the non-intersection conflict

$$(\text{MOM5} \perp\!\!\!\perp \text{POP1} \mid \text{PAL1}), (\text{MOM5} \perp\!\!\!\perp \text{PAL1} \mid \text{POP1}),$$

genes POP1 and PAL1 are merged to a new node in the structure. After that, no non-trivial independence relations can be detected between nodes (POP1,PAL1), MOM5 and MEX3. The output of RCL is a fully connected undirected graph as shown in Fig. 4.19.

Causal analysis of maternally inherited transcripts is difficult. Cause-and-effect relationships are hard to identify, since a significant amount of transcripts has been placed in the egg maternally. The developing embryo eventually starts expressing its own transcripts, but the initial amount supplied maternally skews the data so that causal analysis becomes difficult. One possibility would be restricting the analysis to the later stages, where the ratio of maternal transcripts becomes negligible compared to the ones of embryo. Unfortunately, selection of data points does not improve the performance of RCL in this data sample. We conjecture that it is due to the relatively small sample size, since it would remove at least half the measurements from the dataset. Although the output of RCL as shown in Fig. 4.19 lacks directionality in the edges, it resembles the factual knowledge on the genes [104, 49]: MEX-3 regulates levels of PAL-1 and MOM5 acts downstream of POP1 and PAL1.

Through these two examples, we can see that, if no independence relations can be accepted, RCL has to test all non-trivial independence constraints between nodes and is not able to infer any direction of edges in the structure. Consequently, RCL will be computationally infeasible if the number of nodes is large.

The last small gene regulatory network of *C. elegans* is the network of mesoderm, which has already been discussed in Section 4.2.4. If we first search for non-intersection conflicts, we have to merge genes HND1 and PHA4 to a new node (HND1,PHA4), because

$$(\text{HLH1} \perp\!\!\!\perp \text{HND1} \mid \text{PHA4}) \wedge (\text{HLH1} \perp\!\!\!\perp \text{PHA4} \mid \text{HND1}).$$

Within the new set of 4 nodes, no non-intersection and no non-chordality conflicts can be found.

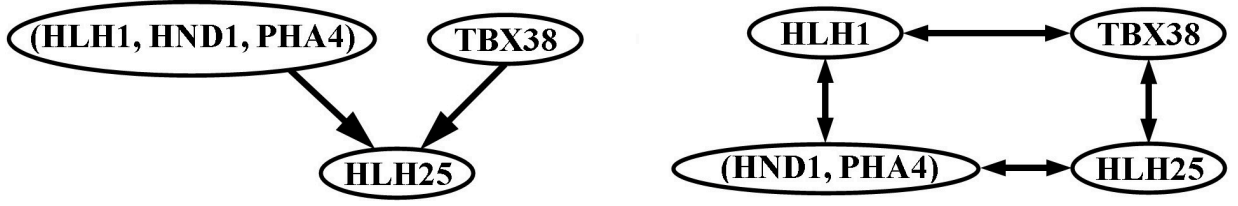


Figure 4.20.: Graphical representations of mesodermal data of *C. elegans*. If we first search for non-chordality conflicts and then non-intersection conflicts, we will have the left plot as the output. If we first search for non-intersection conflicts and resolve them, no non-chordality conflicts can be further detected, thus we obtain the output as shown in the right plot. Different orders of resolving conflicts can, in practice, lead to different clustering results of variables.

Two marginal independence relations

$$(\text{HLH1} \perp\!\!\!\perp \text{HLH25}) \wedge ((\text{HND1}, \text{PHA4}) \perp\!\!\!\perp \text{TBX38})$$

and other 4 marginal dependences between nodes are obtained. The right plot of Fig. 4.20 is the graphical representation of data.

On the other side, if we first search for non-chordality conflicts (see Section 4.2.4), the genes HLH1 and HND1 will be merged to (HLH1,HND1). After that, a non-intersection conflict can be further identified within the fully connected nodes (HLH1,HND1), PHA4 and HLH25, since

$$(\text{HLH25} \perp\!\!\!\perp (\text{HLH1}, \text{HND1}) \mid \text{PHA4}) \wedge (\text{HLH25} \perp\!\!\!\perp \text{PHA4} \mid (\text{HLH1}, \text{HND1})).$$

Therefore, we have to merge (HLH1,HND1) and PHA4 to a new node (HLH1,HND1,PHA4). Within the new set of 3 nodes, we obtained the following constraints by means of the kernel independence test

$$((\text{HLH1}, \text{HND1}, \text{PHA4}) \perp\!\!\!\perp \text{TBX38}) \wedge ((\text{HLH1}, \text{HND1}, \text{PHA4}) \not\perp\!\!\!\perp \text{HLH25}) \wedge (\text{TBX38} \not\perp\!\!\!\perp \text{HLH25}),$$

which indicates a v -structure, i.e., the unshielded collider on HLH25. This v -structure is confirmed by the conditional dependence, i.e.,

$$(\text{HLH1}, \text{HND1}, \text{PHA4}) \not\perp\!\!\!\perp \text{TBX38} \mid \text{HLH25}.$$

Therefore, the DAG as shown in the left plot of Fig. 4.20 is indeed faithful with respect to data by means of kernel independence test.

Summing up, the output in the left plot of Fig. 4.20 is obtained by first searching for non-chordality and then non-intersection conflicts, while the result in the right plot is obtained by first searching for non-intersection conflicts. This example makes clear that, in practice, different orders of resolving conflicts could lead to different clustering of variables, and different

4. From Independence Relations to Causal Structure

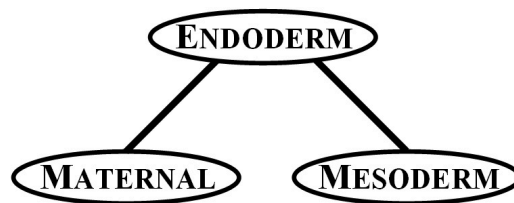


Figure 4.21.: Output of RCL on data of *C. elegans*.

construction of nodes could lead to different structures. A structure that contains more information, in the sense that more edges are directed, is desirable.

The mesodermal data proves to be the most complicated to analyze. This is due to different cell lineages separating early and showing lineage specific gene expression patterns that become overlaid in the microarray data. We recommend at this point to separate expression profiles of genes according to their spatial distribution. Nevertheless, it is possible to observe fragments of the pathways in the resulting structure. The structure still retains the coupling of HND1 and PHA4, the connection between HLH1 and HND1 and a bi-directed connection between HLH25 and HND1 which indicates a probable hidden common cause, in our case most likely MED-1,2 [104].

For the sake of completeness, we ran RCL on the whole dataset of *C. elegans* containing all 13 genes as shown in Tab. 4.1. Prior knowledge was used to group the genes into maternally inherited, mesoderm related and endoderm related. The nodes in the structure correspond to the groups of genes, namely MATERNAL, MESODERM, and ENDODERM. Only one non-trivial independence relation $\text{MATERNAL} \perp\!\!\!\perp \text{MESODERM} \mid \text{ENDODERM}$ is detected. The final output of RCL is shown in Fig. 4.21, which excludes ENDODERM being the common effect of MESODERM and MATERNAL.

From a biological point of view, the result of RCL did not capture the essential relationships between the groups of genes. The prior knowledge states that MATERNAL influences ENDODERM and MESODERM. The endodermal and mesodermal factors interact with each other. Actually, this fact can most likely be attributed to the temporal nature of the data. In the beginning the maternal transcripts are the driving force of the development and are the primary causative element for endodermal and mesodermal factors but later on the system switches to a more networked state where the gene groups are starting to influence each other [104]. Further research concerning the change of causal relationships over time is needed in order to properly deal with this sort of data, i.e., time series.

The other point is that the factors MATERNAL, ENDODERM and MESODERM are represented by a four- or five-dimensional variable. Given the same sample size, the independence constraints represented by the graph in Fig. 4.21 are expected to be less reliable than those in Fig. 4.17, Fig. 4.7 and Fig. 4.6. For this reason, a structure with nodes represented by low-dimensional variables can be tested more reliably.

The resulting causal structure should never be seen as something definitive. Especially in biological systems it is often the case that functionally unrelated components show a high degree

of correlation in their activity, which might induce a connection in the causal structure. Also the feedback-driven nature of biological systems does not coincide with the acyclicity assumption and thus tends to result in complications in the construction of the causal structure. The RCL algorithm itself does not include information in the resulting structure to differentiate between connections that cause or inhibit an event. In the case of two nodes only representing single events, this can easily be extracted by a correlation analysis of the dataset. For nodes containing multiple variables it is not clear how to recover the type of relationship between groups of variables.

4.6.5. Metastatic melanoma

From the practical viewpoint, if more than 4 variables are measured, we need an order of fully connected substructures that are considered for exploring the non-intersection conflicts by Step 1.2 of RCL (Fig. 4.12).

Assumption 5 *Let \mathcal{G}_1 and \mathcal{G}_2 be two fully connected substructures in DAG \mathcal{G} . If the weakest marginal dependence between any two distinct variables X_1 and Y_1 involved in \mathcal{G}_1 is larger than the weakest marginal dependence between any two distinct variables X_2 and Y_2 involved in \mathcal{G}_2 , then the adjacency structure corresponding to \mathcal{G}_1 is more reliable than the adjacency structure corresponding to \mathcal{G}_2 .*

The intuition behind this assumption is that, given some appropriate measure of dependences, the stronger the dependence between two variables can be measured, the more reliable a connection between the nodes representing the variables can be inferred. We propose to first consider the more reliable adjacency structure corresponding \mathcal{G}_1 for exploring the conflicts among the constraints, then \mathcal{G}_2 .

Now, we consider real data from biology, namely metastatic melanoma. Even though only 4% of observed skin cancer incidences are melanoma, it is responsible for almost 80% of all deaths attributed to this type of cancer. Only 14% of patients with metastatic melanoma survive for 5 years [114]. It is widely accepted that major risk factors of melanoma are genetic predisposition and exposure to UV light.

We applied the RCL algorithm (with additional Assumption 5) to the 31 gene expression profiles generated in the study of metastatic melanoma [23]. To restrict the number of genes we concentrated on a small set likely connecting to a local regulatory network. In the expression profiling study of Bittner et al. [23], WNT5A has been identified as a gene of interest involved in melanoma. It was experimentally proved that increasing the level of WNTA5 (2) protein can influence the cell's metastatic potential [173]. Due to its implication in the metastatic spread of melanoma cells, gene WNT5A was chosen in the regulatory network.

Methods for choosing the subset of 10 genes involved in a small local network that includes the activity of WNT5A is described in [97]. The network contains the 10 most significant genes which are narrowed down from 587 genes: pirin (1), WNT5A (2), S100P (3), RET-1 (4), MMP-3 (5), PHO-C (6), MART-1 (7), HADHB (8), synuclein (9), and STC2 (10). Tab. 4.6 summarizes the genes and their function.

4. From Independence Relations to Causal Structure

1.	pirin	Implied in transcription activation and apoptosis.
2.	WNT5A	Secreted signaling protein.
3.	S100P	S100 calcium binding protein P.
4.	RET-1	Reticulon-1 (RTN-1). Predominantly expressed in brain tissue.
5.	MMP-3	Proteins of the matrix metalloproteinase (MMP) family are involved in the breakdown of extracellular matrix.
6.	PHO-C	Phospholipase C, Gamma 1 (PLCG1).
7.	MART-1	Antigen that is specific to the melanocyte lineage, found in normal skin, the retina, and melanocytes (Melan-A).
8.	HADHB	Subunit of the mitochondrial trifunctional protein.
9.	synuclein	May be involved in the regulation of dopamine release and transport.
10.	STC2	Anti-hypocalcemic action on calcium and phosphate homeostasis.

Table 4.6.: Genes involved in metastatic melanoma data and their function.

The expression data was quantized to a ternary state $\{-1, 0, 1\}$ indicating reduced, normal and enhanced expression levels. Quantization smoothes errors introduced by noise and other factors indirectly influencing measured expression levels. Fig. 4.18 visualizes the data of these 10 genes with three expression levels.

In the first run, we interpret the variables as continuous ones and use the empirical kernel dependence measures (as defined in Definition 13 with Gaussian kernels) to quantify the degree of dependence. Fig. 4.22 illustrates the stepwise results of RCL on the melanoma data. The leftmost plot is the underlying adjacency structure induced by marginal constraints. We explore non-intersection conflicts within fully connected substructures of this adjacency structure. To resolve conflicts, following reconstruction of nodes are necessary:

$$\begin{aligned}
(2 \perp\!\!\!\perp 1 \mid 8) \wedge (2 \perp\!\!\!\perp 8 \mid 1) &\Rightarrow \text{merge 1 and 8 to } (1,8), \\
(5 \perp\!\!\!\perp 3 \mid 6) \wedge (5 \perp\!\!\!\perp 6 \mid 3) &\Rightarrow \text{merge 3 and 6 to } (3,6), \\
(10 \perp\!\!\!\perp 2 \mid 7) \wedge (10 \perp\!\!\!\perp 7 \mid 2) &\Rightarrow \text{merge 2 and 7 to } (2,7), \\
((3,6) \perp\!\!\!\perp 5 \mid 10) \wedge ((3,6) \perp\!\!\!\perp 10 \mid 5) &\Rightarrow \text{merge 5 and 10 to } (5,10), \\
((5,10) \perp\!\!\!\perp 9 \mid (3,6)) \wedge ((5,10) \perp\!\!\!\perp (3,6) \mid 9) &\Rightarrow \text{merge 9 and } (3,6) \text{ to } (3,6,9).
\end{aligned}$$

We obtain a new set of 5 nodes without non-intersection conflicts as shown in the second plot (from left) in Fig. 4.22.

Based on this clustering of variables, we first test marginal independence between distinct clusters of variables, i.e., the nodes in the graph. The resulting adjacency structure of these nodes is shown in the second plot of Fig. 4.12. The third plot shows the result of inferring v -structures after Step 1.1 of RCL. No non-chordality conflicts can be found by Step 2. Step 3 removes the unnecessary edge between $(3, 6, 9)$ and $(5, 10)$ due to $((3, 6, 9) \perp\!\!\!\perp (5, 10) \mid (2, 7), 4)$ with respect to the potential ancestor condition. All edges are then directed. The final result is shown in the rightmost plot. The bi-directed edges between $(2, 7)$ and $(3, 6, 9)$ and between

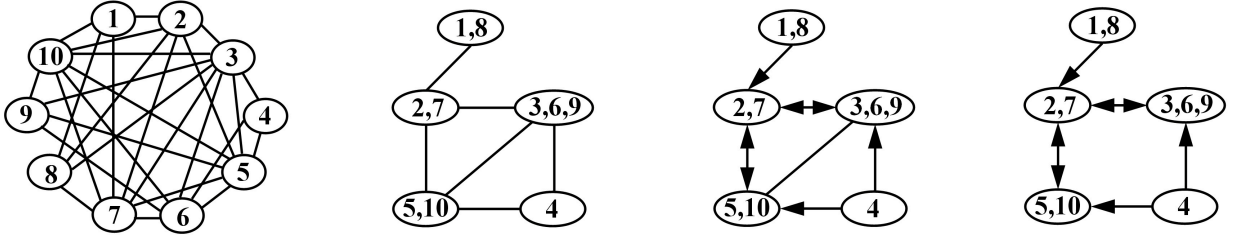


Figure 4.22.: Stepwise results of RCL (by means of independence tests via kernel measures) applied to metastatic melanoma data.

(2, 7) and (5, 10) could be traced back to latent common causes. Another explanation is that the underlying model is indeed cyclic.

In the second run, we interpret the variables as categorical ones and use the mutual information to measure dependences between variables. The magnitude of dependences measured by mutual information differs from the magnitude measured by kernel methods. In particular, under Assumption 5, the order on the fully connected substructures that are considered for exploring non-intersection conflicts by Step 1.2 of RCL is different.

The leftmost plot of Fig. 4.23 is the underlying adjacency structure induced by marginal constraints (via mutual information). In comparison to the leftmost plot of Fig. 4.22 (with kernel measures), the only difference is that tests via mutual information accept the constraint $(3 \perp\!\!\!\perp 8)$ with a p-value of 0.056, while the test via kernel measure reject the constraint $(3 \perp\!\!\!\perp 8)$ with a p-value of 0.044.

We explore non-intersection conflicts within fully connected substructures of this adjacency structure. To resolve conflicts, following reconstruction of nodes are necessary:

$$\begin{aligned}
 (3 \perp\!\!\!\perp 2 \mid 7) \wedge (3 \perp\!\!\!\perp 7 \mid 2) &\Rightarrow \text{merge 2 and 7 to } (2,7), \\
 (9 \perp\!\!\!\perp (2,7) \mid 10) \wedge (9 \perp\!\!\!\perp 10 \mid (2,7)) &\Rightarrow \text{merge } (2,7) \text{ and } 10 \text{ to } (2,7,10), \\
 ((2,7,10) \perp\!\!\!\perp 1 \mid 8) \wedge ((2,7,10) \perp\!\!\!\perp 8 \mid 1) &\Rightarrow \text{merge 1 and 8 to } (1,8), \\
 (5 \perp\!\!\!\perp 3 \mid 6) \wedge (5 \perp\!\!\!\perp 6 \mid 3) &\Rightarrow \text{merge 3 and 6 to } (3,6), \\
 (5 \perp\!\!\!\perp 9 \mid (3,6)) \wedge (5 \perp\!\!\!\perp (3,6) \mid 9) &\Rightarrow \text{merge 9 and } (3,6) \text{ to } (3,6,9).
 \end{aligned}$$

We obtained a set of 5 nodes as shown in the second plot (from left) of Fig. 4.23, which only slightly differs from the clustering as shown in the second plot of Fig. 4.22 using kernel measures. The only difference is that the gene STC2 (10) belongs to different clusters.

Step 2 of RCL detects no non-chordality conflicts. Step 3 removes the unnecessary edge between (3, 6, 9) and (5) due to $((3, 6, 9) \perp\!\!\!\perp 5 \mid (2, 7, 10), 4)$ with respect to the potential ancestor condition. All edges are then directed. The final output of RCL by means of mutual information is shown in the rightmost plot of Fig. 4.23.

Interestingly, the clustering results and final graphical outputs by means of the permutation tests using mutual information and kernel measures as dependence measures are quite similar, although mutual information interpreted the variables as categorical ones while kernel measures

4. From Independence Relations to Causal Structure

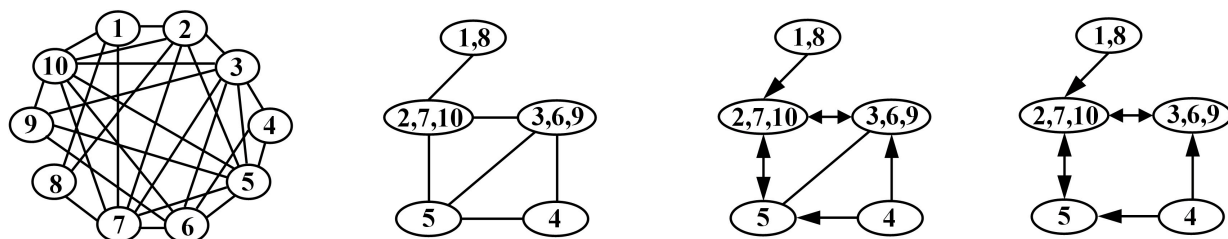


Figure 4.23.: Stepwise results of RCL algorithm (by means of independence tests via mutual information) applied to metastatic melanoma data.

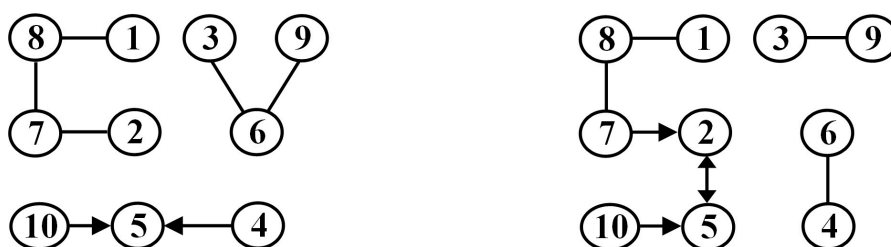


Figure 4.24.: Outputs of PC applied to metastatic melanoma data. The left plot is the result, if we consider the underlying domains categorical and use likelihood-ratio χ^2 test. The right plot is the result, if we consider the underlying domains continuous and use Fisher's Z test. The significance level is chosen to be 0.05

interpreted them as continuous ones. We conjecture that it is due to the ternary domains. More precisely, 4 genes, i.e., RET-1 (4), PHO-C (6), MART-1 (7), HADHB (8), have only two expression levels in the 31 cases observed (see Fig. 4.18). The similar results achieved by different measures showed that the constraint-based clustering procedure (Fig. 4.9) is reasonably robust with respect to the order of checking conflicts.

Note that the conventional constraint-based PC algorithm does not handle probable violations of the faithfulness assumption at all. In conflicting situations, the output of PC depends on the order of checking independence. For comparison, we performed PC on the metastatic melanoma data with likelihood-ratio χ^2 and Fisher's Z test (see Fig. 4.24 for outputs). Interestingly, we can observe that the outputs of PC contains the edges between variables which are clustered to a node by RCL and the edges between sets of variables which are represented by nodes in the outputs of RCL are mostly absent.

It is indeed difficult to evaluate the performance of RCL on such a biological dataset, because the ground truth is not completely known. Nonetheless, the causal interpretation gained from the resulting structure can partly be confirmed by prior knowledge of biologists or are consistent with other studies.

Genes pirin (1) and HADHB (8) are identified as the start point of causal chain by RCL (see Fig. 4.22 and Fig. 4.23). The gene pirin (1) is a transcription factor and believed to have influence

on apoptosis [2, 121]. Datta et al. [45] suggested in their study to control the level of WNT5A (2) directly or through pirin (1), since they believe that controlling the influence of WNT5A (2) in the regulation can reduce the chance of melanoma metastasizing. For pirin (1) it is possible to image a role as a regulatory element of the other nodes in the structure. Gene HADHB (8) is part of a mitochondrial protein complex responsible for oxidation of fatty acids [120]. The regulatory role of HADHB (8) is unknown and thus it is possible that its connections are due to non-functional dependences.

Gene WNT5A (2) is a secreted signaling protein whose deregulation plays a central role in cancer progression [128]. RNAi evidence points towards a connection of WNT5A (2) and MART-1 (7) [150].

Gene MMP-3 (5) belongs to a family of secreted proteins that breakdown the protein components of the extra cellular matrix. This detachment from the matrix allows cancer cells to migrate and develop metastasis distant from the primary tumor. It is well known that multiple members of the MMP family are involved in this process [42, 181]. Gene STC2 (10) plays a role in the maintenance of the calcium homeostasis. Deregulation of calcium levels is believed to help cancer cells achieve their anti-apoptotic property [127].

The remaining nodes of the structure contain genes expressed in brain tissue, i.e., RET-1 (4), dealing with energy metabolism, i.e., S100P (3), PHO-C (6), and a dopamine release synuclein (9). Both RET-1 (4) and synuclein (9) are primarily expressed in neural tissue, which makes the connection between them likely. However, the functional role of RET-1 (4) is yet unknown. It could be that the causal connections to those genes are due to non-functional correlations in activity. Alternatively, it is imaginable that the connection between (2, 7)/(2, 7, 10) and (3, 6, 9) is actually traced back to a hidden common cause (maybe a transcription factor) controlling both nodes.

In summary, the structure serves as a good example for the discriminative power of the RCL algorithm. Even on data of small sample size, it is possible to extract meaningful causal relationships which are kept separate from genes not likely to participate in the functional network.

5. From Magnitude of Dependences to Causal Structure

The main shortcoming of learning causal structure from independence relations is that it cannot learn anything, if no (conditional) independence can be verified. Further, a reliable test on independence constraints is of utmost importance. It is, however, not guaranteed when the sample size is small or the conditioning set is large. In this chapter, we will propose to make use of the magnitude of dependences measured by kernels to get hints about the causal structure, even when no conditional independence is present.

5.1. Problems of learning structure via independence tests

If we had direct access to the true distribution, we would always make the correct decision about the independence in the population. In practice, the decision is made based on sample, the observed data may not be very representative of the population and therefore leads us to an incorrect decision. As mentioned previously (Tab. 3.1), The errors made by independence tests can be classified as type I and II error. The common way of controlling errors made by a single hypothesis test is using significance level α (usually 5%) to control the type I error. Under a fixed level of type I error, one tries to keep the type II error level as low as possible. Therefore, it could happen that type II error cannot be kept to a low level, when type I error is controlled to a pre-specified level α . It is very difficult to handle the trade-off between the level of type I and II error, which is utmostly important for learning structure from independence constraints.

In particular, if the sample size is small, statistical tests will be unreliable. Note that the term “small” is relative and depends on the size of the model, because data, even when considered as “large”, might often be small with respect to the number of joint states of variables with a large domain.

As an example, we describe a real dataset, which was used to study food products for palatability by Street et al. [161] (see <http://lib.stat.cmu.edu/DASL/Datafiles/tastedat.html> for data). The experiment involved the effects on palatability of a coarse versus fine screen (large “pieces” versus small “pieces”) and of a low versus high concentration of a liquid component. The dataset consists of 16 cases and three variables, i.e., SCORE: total palatability score for 50 consumers: general Foods employed a 7-point scale from -3 (terrible) to $+3$ (excellent) with 0 representing “average”; LIQUID: liquid level (0: “low”, 1: “high”); and SCREEN: screen type (0: “coarse”, 1: “fine”). The sample size of data is 16, which is quite small, since we have 28 possible states of

5.1. Problems of learning structure via independence tests

variables (7 points \times 2 liquid level \times 2 screen type).

We ran hypothesis tests of non-trivial independence relations between three variables. Tab. 5.1 summarizes the results of tests based on correlation analysis and kernel dependence measures. As seen from the table, tests provided the same set of independence relations.

Independence Hypothesis	Correlation Analysis			Kernel Dependence			Resampling-based Multiple Test
	Measure	p-Value	Test	Measure	p-Value	Test	
Liquid \perp Screen	0.0000	0.596	Accept	0.0000	0.601	Accept	Accept
Liquid \perp Screen Score	0.0710	0.336	Accept	0.0019	0.328	Accept	Reject
Liquid \perp Score	-0.2475	0.365	Accept	0.0156	0.500	Accept	Reject
Liquid \perp Score Screen	-0.4093	0.133	Accept	0.0144	0.438	Accept	Reject
Screen \perp Score	0.7965	0.000	Reject	0.1145	0.000	Reject	Reject
Screen \perp Score Liquid	0.8221	0.001	Reject	0.0639	0.004	Reject	Reject

Table 5.1.: Correlation analysis and kernel independence test on taste score data.

In order to see whether the observed sample may be representative of the population or not, we amplify the original sample of size 16 by subsamples of size 17, 24, 32, 48, 80, 144. The subsamples are resampled with replacement from the original data. For each of the 6 subsample sizes, we sampled 100 subsamples and calculated the p-value for each of the 100 subsamples by means of the kernel measure. Thus, we obtained a set of 100 p-values for every independence hypothesis. The set of 100 p-values is reordered from small to large. Fig. 5.1 shows the so-called Q-Q plots (“Q” stands for quantile) of the set of reordered 100 p-values with difference colors for different subsample sizes.

In the case of obvious independence, i.e., LIQUID \perp SCREEN (top left plot of Fig. 5.1), the Q-Q plot of p-values is close to the diagonal line, since the p-values are somewhat uniformly distributed in $[0, 1]$. The form of the Q-Q plot of p-values does not significantly change, as the size of subsamples increases. In the case of an obvious dependence, e.g., SCREEN $\not\perp$ SCORE (top right plot), the Q-Q plot is very close to the lower line, which means that almost all of the hypotheses on 100 subsamples would be rejected. The larger the size of subsamples, the closer the Q-Q plot to the lower line, the more likely the dependence.

The ambiguous case is more interesting, i.e., the non-significant dependence between LIQUID and SCORE (top middle plot). As the subsample size increases, more and more tests will reject the independence hypothesis, i.e., the Q-Q plot of p-values varies from the diagonal position to the lower line. Consequently, LIQUID and SCORE would be dependent, if we could observe more data points, although the independence test on the original sample did not detect significant dependence. Resampling-based hypothesis tests, e.g., with a resampling size 48, would revise the set of constraints. The revised constraints in the last column of Tab. 5.1 lead to a v -structure between LIQUID, SCREEN and SCORE, as shown in Fig. 5.2. Note as aside, a resampling-based hypothesis test via correlation achieved the same results as shown in the last column of Tab. 5.1.

It is clear that a weak dependence can always be verified as a significant one, if the resampling

5. From Magnitude of Dependences to Causal Structure

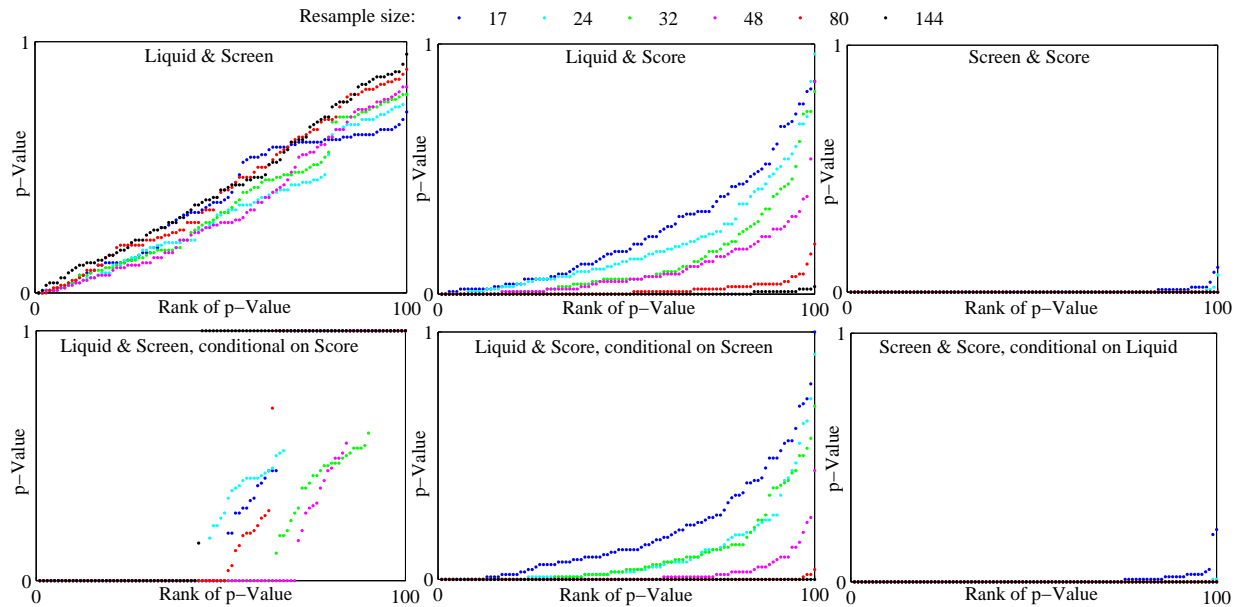


Figure 5.1.: Q-Q plot of p-values of a resampling-based kernel independence test on taste score data. The upper row shows the tests of unconditional constraints using different sizes of resampling (different colors). The original sample size is 16. The lower row shows the tests of conditional constraints.

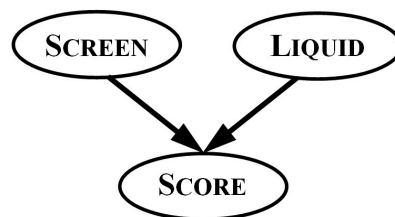


Figure 5.2.: Taste score data represented by a DAG, given the independence constraints obtained by a resampling-based kernel independence test (last column of Tab. 5.1).

5.2. Identifying colliders via magnitude of dependences

size is large enough. The choice of resampling size actually handles the trade-off between type I and II error implicitly. Our simulated experiments gave numerical evidence of power increase by such resampling-based hypothesis test (see Section C.1 in Appendix C for the procedure and some experiments). We believe that the resampling-based hypothesis test is a better way to balance the type I and II error than a direct choice of the level of merely type I error, in particular, if the sample size is extremely small. However, there is yet no principled way to choose the resampling size. Further, such multiple testing is obviously extremely time-consuming.

An alternative way to avoid rejecting too many dependences is a direct use of the magnitude of dependences measured by kernels for learning causal structure. In the example of taste score data, we have

$$\widehat{\mathbb{H}}_{\text{Liquid, Screen}} < 10^{-30} \approx 0 \quad \text{and} \quad \widehat{\mathbb{H}}_{\text{Liquid, Screen}|\text{Score}} = 0.0019 \gg 0,$$

which could be interpreted as indicator for an unshielded collider on SCORE, in the spirit of the criterion as in Eq. 4.3. Or, analogous to the condition as in Eq. 4.4 or Eq. 4.5, the magnitude of dependences

$$\widehat{\mathbb{H}}_{\text{Liquid, Score}} = 0.0156 \gg 0, \widehat{\mathbb{H}}_{\text{Screen, Score}} = 0.1145 \gg 0, \quad \text{and} \quad \widehat{\mathbb{H}}_{\text{Liquid, Screen}} < 10^{-30} \approx 0$$

can also serve as an indicator for an unshielded collider on SCORE. Following sections will systematically elaborate on the question how to use the magnitude of dependences to infer the causal structure.

5.2. Identifying colliders via magnitude of dependences

We first describe some criteria that may give evidence of a collider in the structure. As a start, we consider an unshielded triple $X - Z - Y$ (X and Y nonadjacent). The identification of Z as an unshielded collider in the structure establish an essential basis part of a constraint-based approach. Under faithfulness assumption, conditioning on Z should induce dependence between X and Y , i.e., Z activates the path between them. Only the empty set blocks the path between X and Y . By means of kernel dependence measures, $\widehat{\mathbb{H}}_{YX} \approx 0$ and $\widehat{\mathbb{H}}_{YX|Z} \gg 0$ strongly indicate that Z is the common effect of probably independent X, Y . This leads to the following criterion.

Criterion 1 *Given variables X, Y, Z , if the ratio $\frac{\widehat{\mathbb{H}}_{YX|Z}}{\widehat{\mathbb{H}}_{YX}}$ is very large, Z is a strong candidate for being a collider on the path between X and Y .*

Graphically, $X \rightarrow Z$ and $Z \leftarrow Y$ can be inferred. In this case, variables X and Y are probably independent, i.e., an unshielded collider on Z .

Since we believe that $\widehat{\mathbb{H}}_{YX}$ and $\widehat{\mathbb{H}}_{YX|Z}$ quantify the magnitude of dependences in a reasonable sense (with an appropriate choice of kernels), we dare to go a step further and extract hints on direction in a shielded triple $X - Y - Z - X$, i.e., a fully connected adjacency structure.

5. From Magnitude of Dependences to Causal Structure

Criterion 2 Given variables X, Y, Z , if

$$\frac{\widehat{\mathbb{H}}_{YX|Z}}{\widehat{\mathbb{H}}_{YX}} > \frac{\widehat{\mathbb{H}}_{ZX|Y}}{\widehat{\mathbb{H}}_{ZX}} \quad \text{and} \quad \frac{\widehat{\mathbb{H}}_{YX|Z}}{\widehat{\mathbb{H}}_{YX}} > \frac{\widehat{\mathbb{H}}_{ZY|X}}{\widehat{\mathbb{H}}_{ZY}},$$

then Z is a weak candidate for being a collider on the path between X and Y .

Graphically, $X \rightarrow Z$ and $Z \leftarrow Y$ can be inferred. In this case, variables X and Y are probably dependent, i.e., a shielded collider on Z .

We unify Criterion 1 for unshielded collider identification and Criterion 2 for shielded collider identification into a so-called “ λ -collider condition”.

Definition 22 (λ -Collider Condition) For any triple (X, Y, Z) with the substructure $X-Z-Y$, where X and Y may be adjacent or nonadjacent, variable Z is a candidate for being a collider between X and Y , if and only if

$$\widehat{\mathbb{H}}_{YX|Z} > \lambda \widehat{\mathbb{H}}_{YX} \quad (5.1)$$

with appropriate $\lambda > 0$.

If the collider is indeed unshielded, i.e., $\widehat{\mathbb{H}}_{YX} \approx 0$, one would expect that the inequality holds for a very large λ , say larger than some pre-specified constant λ_1 . In the case of a shielded collider, λ is chosen to be, say λ_2 , based on Criterion 2

$$\lambda_2 := \rho \cdot \max \left\{ \frac{\widehat{\mathbb{H}}_{ZX|Y}}{\widehat{\mathbb{H}}_{ZX}}, \frac{\widehat{\mathbb{H}}_{ZY|X}}{\widehat{\mathbb{H}}_{ZY}} \right\} \quad \text{with } \rho \geq 1. \quad (5.2)$$

λ_1 should be chosen sufficiently large and it is clear that $\lambda_1 \gg \lambda_2$. In our implementation, we chose $\lambda_1 := 100$. Given observed sample, λ_2 can be calculated empirically. The parameter ρ is used to avoid the uncertainty of probable sample errors. In our experiments, we chose $\rho = 1.2$.

To ensure the numerical stability of the scores in Eq. (5.2), we add a very small regularization constant ϵ to the kernel matrices that are used to compute the score $\widehat{\mathbb{H}}_{YX}$ by

$$\frac{1}{(n-1)^2} \text{Tr}((\widehat{K}_Y + \epsilon I)(\widehat{K}_X + \epsilon I)),$$

when $\widehat{\mathbb{H}}_{YX}$ appears in the denominator. In our experiments, we set $\epsilon = 10^{-5}$ throughout the thesis.

It is crucial to use the ratio, not the difference, of conditional and unconditional measure for the criteria, because the fact that one of the unconditional dependences is close to zero, i.e., one of the three ratios is significantly larger than other two, is essential for the decision of a collider structure. Under the faithfulness assumption, one can actually identify an unshielded collider on Z , if marginal constraints $X \perp\!\!\!\perp Y \wedge X \not\perp\!\!\!\perp Z \wedge Y \not\perp\!\!\!\perp Z$ can be indeed verified, instead of all non-trivial constraints as shown in the first column of Tab. 1.1 (see Section 4.2.2 for more discussions). This means unconditional measures $\widehat{\mathbb{H}}_{ZX}, \widehat{\mathbb{H}}_{ZY} \gg \widehat{\mathbb{H}}_{YX} \approx 0$ would be sufficient to make the decision for a collider on Z .

5.3. Orientation heuristics via collider identification

The situation will be more apparent if mutual information, a popular dependence measure, is used. For a triple (X, Y, Z) , the following equation holds in general [179]:

$$\mathbb{I}(X, Y) - \mathbb{I}(X, Y|Z) = \mathbb{I}(X, Z) - \mathbb{I}(X, Z|Y) = \mathbb{I}(Y, Z) - \mathbb{I}(Y, Z|X) =: \mathbb{I}(X, Y, Z).$$

Note that the quantity $\mathbb{I}(X, Y, Z)$, the mutual information among triple (X, Y, Z) , could be positive or negative. The difference of unconditional and conditional mutual information, i.e., $\mathbb{I}(X, Y, Z)$, reflects a joint property of the triple (see also [132] for more details), which cannot provide any information about the structure among them. In contrast, the three ratios are not equal to each other, i.e.,

$$\frac{\mathbb{I}(X, Y|Z)}{\mathbb{I}(X, Y)} \neq \frac{\mathbb{I}(X, Z|Y)}{\mathbb{I}(X, Z)} \neq \frac{\mathbb{I}(Y, Z|X)}{\mathbb{I}(Y, Z)}.$$

This is due to

$$\frac{\mathbb{I}(X, Y, Z)}{\mathbb{I}(X, Y)} \neq \frac{\mathbb{I}(X, Y, Z)}{\mathbb{I}(X, Z)} \neq \frac{\mathbb{I}(X, Y, Z)}{\mathbb{I}(Y, Z)}.$$

In these ratios, the magnitude of unconditional dependences plays an essential role.

The reason why we use kernel dependence measures, not mutual information, is in part the practical implementation, in particular, on continuous domains. In the limit of infinite sampling, HS-norm of the conditional cross-covariance operator provides a general distribution-free tool to capture dependences. Having chosen a kernel such that functions being less smooth correspond to larger RKHS-norms, large dependence measures will then indicate correlations between smooth functions. A finite cut-off value for dependence measures corresponds to neglecting correlations if they are small or if they occur only on complex (not sufficiently smooth) functions (see Section 2.7 for more discussions), which is certainly a reasonable indicator for independence. Criterion 2 takes the quantitative information about dependence measured by kernels into account and makes, in fact, some implicit assumption, via the choice of kernels, on the prior probability distribution of the transition probabilities that occur in nature. Only extensive experiments with real-world data can really decide whether the assumption behind our criteria provide useful hints or not, because, if the true model is indeed fully connected, all joint probability distributions can, in principle, be generated. Then, it will be very hard to find a reliable principled way to prefer one of them.

5.3. Orientation heuristics via collider identification

The λ -collider condition (Definition 22) shows that Criterion 1 and Criterion 2 can be considered as two related conditions of increasing strength and correspond to different degrees of reliability. It is reasonable to expect that the weaker the assumption, i.e., the larger the value of λ , the fewer collider structures will be erroneously identified. This suggests that the collider structures identified by λ_1 have priority over those by λ_2 . Note that we do not intend to interpret the ratios $\widehat{\mathbb{H}}_{YX|Z}/\widehat{\mathbb{H}}_{YX}$, $\widehat{\mathbb{H}}_{ZX|Y}/\widehat{\mathbb{H}}_{ZX}$, and $\widehat{\mathbb{H}}_{ZY|X}/\widehat{\mathbb{H}}_{ZY}$ as scoring functions for the evidence of being a collider, since only the comparison of the ratios gives hints on being a collider, not the value of

5. From Magnitude of Dependences to Causal Structure

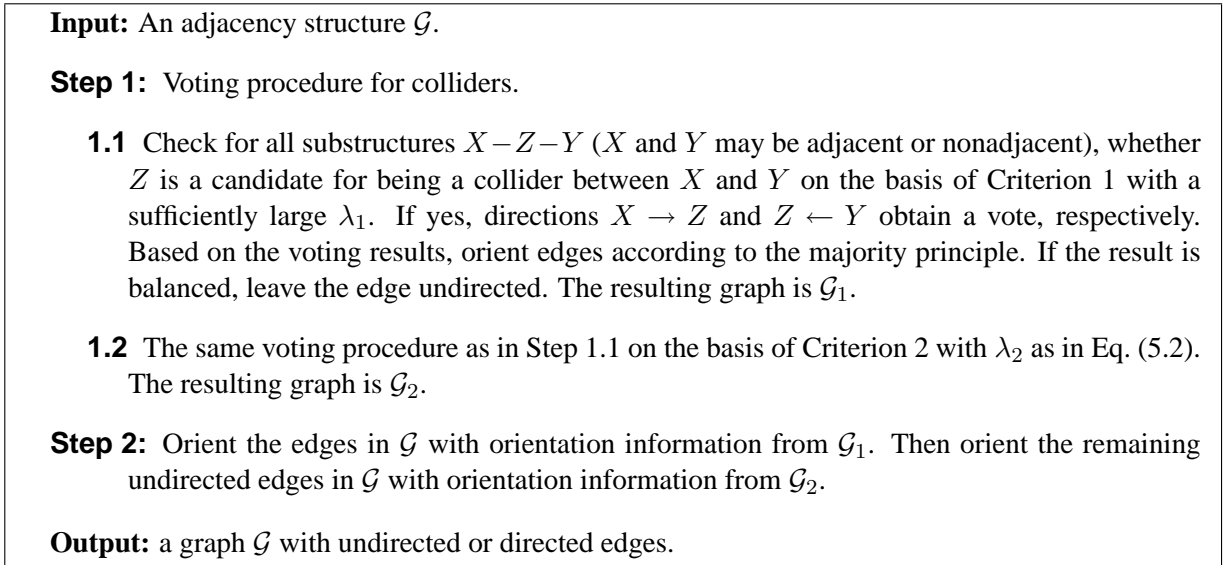


Figure 5.3.: Orientation procedure A (OPA) by a majority vote.

ratios itself. Such a “two- λ -scheme” guarantees that an unshielded collider or a collider with weakly dependent parents can be first identified by very reliable Criterion 1, and prevented from getting (probably) wrongly re-oriented through the less reliable Criterion 2.

If we consider a fully-connected adjacency structure of triple (X, Y, Z) isolated from the whole network, the λ -collider condition can only be justified by hand-waving arguments. However, if we consider a network with more than three variables, we can use criterion 1 or criterion 2 for all triples of measured variables. If Criterion 1 or Criterion 2 identifies Z as a collider between X and Y , we register this as a vote for orientations $X \rightarrow Z$ and $Y \rightarrow Z$, respectively. After having checked all possible triples, we infer the orientation of each edge by a majority vote. A similar voting procedure is proposed in Section 4.4. The difference is that we orient the edges here by the majority principle. Consequently, the resulting graph contains no bi-directed edges.

Combining the voting procedure with the two- λ -scheme, we present an orientation heuristics without testing independence, called orientation procedure A (short: OPA), as shown in Fig. 5.3. Actually, OPA assumes that a certain pattern of marginal and conditional dependences makes some of the orderings of a triple more likely. The voting procedure considers this preference for a particular structure significant only if the evidence provided by many distinct triples is consistent. The detailed pseudocode of OPA can be found in Appendix B.1. Note that the resulting graph of OPA is not necessarily completely directed, since the voting results can be balanced.

The main advantage of OPA is that it works even for a fully connected adjacency structure. OPA as a heuristics, however, has the main shortcoming that a small λ (e.g., λ_2 used in Step 1.2) sometimes leads to wrong votes. Hence, the voting majority could be unreliable.¹

¹The way that our method makes use of the quantitative information about the strength of dependence has some analogy to the “monotone faithfulness principle” and BN-PC algorithm proposed by Cheng et al. [28]. It states that blocking a previously active path that connects two nodes decreases the mutual information. Chickering et

5.4. Simulated experiments with orientation heuristics

To make the orientation more reliable, we propose to incorporate the information about probably absence of edges, i.e., if the HS-norm $\widehat{\mathbb{H}}_{YX}$ between X and Y is smaller than some threshold, say 10^{-4}). A slightly modified version of OPA, called orientation procedure B (short: OPB), is proposed in Fig. 5.4. OPB, instead of the majority principle, orients edges only by a unanimous vote, i.e., no dissenting votes. In the case of a mixed voting result, i.e., at least one vote for both directions, we mark the edge with a bi-directed edge. Undirected edge depicts no votes for both directions. The output of OPB is a mixed graph with un-, uni-, and bi-directed edges. Actually, the output of PC sometimes contains also bi-directed edges or even directed cycles (see Fig. 2 in [44] for example, a so-called “pinwheel” structure). They can be interpreted as an indicator for violation of assumptions. A possible interpretation of bi-directed edges obtained by our voting procedure will be discussed in Section 5.4.2.

To take the adjacency structure, i.e., absence of edges, into account, Step 1 of OPB (Fig. 5.4) infers structure by unshielded collider identification. Having identified all unshielded colliders, Step 2 of OPB orients as many of the remaining undirected edges as possible whenever their directions follow from the assumption that neither additional unshielded colliders nor directed cycles exist.² For this purpose, orientation rule 1, 2, 3 for obtaining a maximally oriented pattern (Fig. B.1 in Appendix B.3) can be applied. After that, it could happen that some edges remain undirected. Step 3 of OPB uses the orientation heuristics again to identify shielded colliders, with respect to the given partially directed graph. The detailed pseudocode of OPB can be found in Appendix B.2.

5.4. Simulated experiments with orientation heuristics

Some experiments are conducted on simulated data, which are sampled from functional or logically-linked models. The kernel dependence measures can not only be used to infer the orientation of edges, but also infer the absence or presence of edges by just thresholding the measure. The output will give hints about how reasonable is the magnitude of dependences measured by kernels. Apart from graphical representations, the detailed statistics of edges are useful to give numerical evidence how reliably can the kernel measures specify the set of necessary arrows.

5.4.1. Simulated data from noisy OR gates

We present experiments with six different OR gates as defined in Eq. (3.6). “2-Bit-IndDet” and “3-Bit-IndDet” are deterministic OR gates with 2 and 3 independent input bits, respectively; “2-Bit-IndPro” and “3-Bit-IndPro” are probabilistic OR gates with 2 and 3 independent inputs;

al. [34] showed, however, that this principle could not generally be valid. For networks with many nodes one will usually find several nodes that violate it. Nonetheless, BN-PC will be conducted for performance comparison.

²In a mixed graph, a pair of consecutive edges meeting at a vertex Z on a path form a collider if both edges have an arrowhead at Z , i.e., $\rightarrow Z \leftarrow$, $\leftrightarrow Z \leftrightarrow$, $\leftarrow Z \leftarrow$, $\rightarrow Z \leftrightarrow$. A directed cycle is a directed path $X \rightarrow \dots \rightarrow X$ on which every edge is of the form \rightarrow or \leftrightarrow and all the edges \rightarrow have the same orientation.

5. From Magnitude of Dependences to Causal Structure

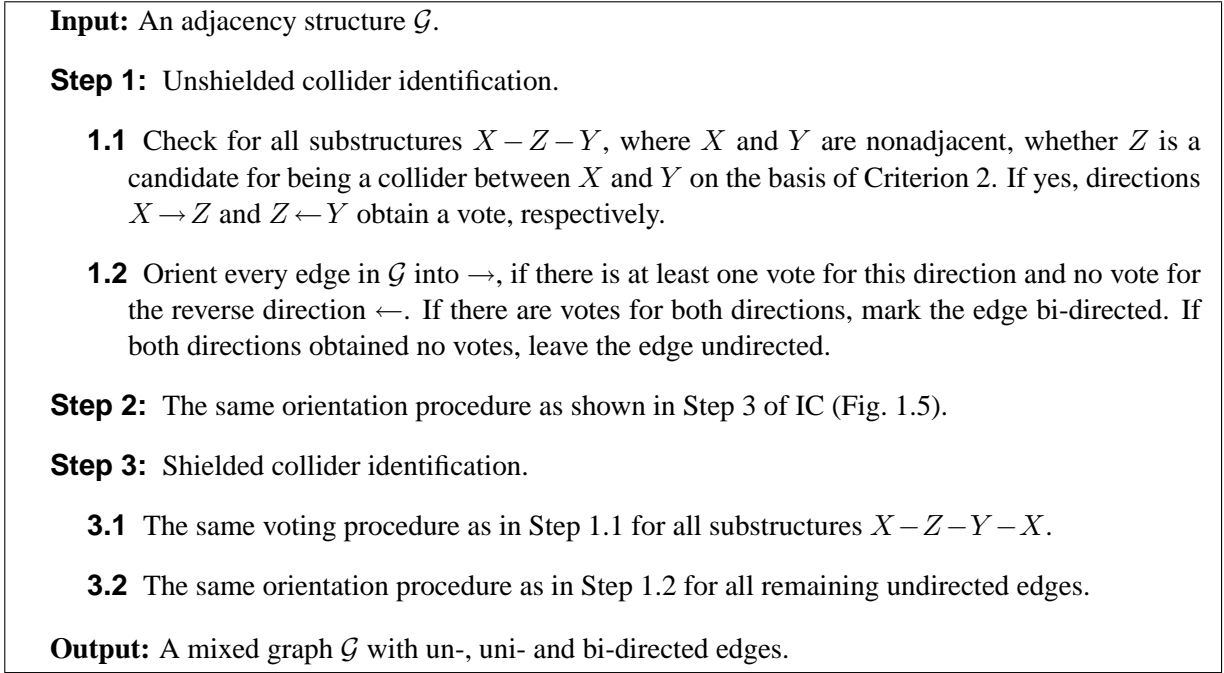


Figure 5.4.: Orientation procedure B (OPB) by a unanimous vote.

whereas the probabilistic OR gates “2-Bit-DepPro” and “3-Bit-DepPro” were fed with 2 and 3 dependent inputs, respectively. The parameters of these models are summarized in Tab. 5.2.

To give some intuition of the value of kernel dependence measures, we randomly picked out three samples of 200 data points, one for each of the three 2-bit OR gates. As seen from Tab. 5.3, the ratio $\hat{\mathbb{H}}_{X_1 X_2 | X_3} / \hat{\mathbb{H}}_{X_1 X_2}$ achieves always the maximum within rows and X_3 can be thus identified as a collider between X_1 and X_2 . The kernel measures describe exactly the fact that conditioning on the output, the inputs become dependent. The ratios for OR gates with independent inputs, i.e., 2-Bit-IndDet and 2-Bit-IndPro, are extremely large, which indicates an unshielded collider on X_3 . In 2-Bit-DepPro, no conditional independence is present, but Criterion 2 is still applicable and indicates a shielded collider on X_3 .

We compared OPB with PC, BN-PC and various score-based Bayesian methods (see Appendix C.2 for details about these methods), based on 1000 replications of the experiments with respective 200 data points sampled from these six OR gates. In the case of OPB, we use cut-off value 10^{-4} for thresholding kernel dependence measure $\hat{\mathbb{H}}_{YX}$ and remove the edge between X and Y . The detailed statistics of 1000 replications can be found in Tab. C.3 and Tab. C.4 in Appendix C.2. Tab. 5.4 summarizes the resulting graph of algorithms in the majority case.³

³The PC algorithm allows no latent variable, thus its output normally contains only directed “ \rightarrow ” and undirected “ $-$ ” edges. FCI, which allows latent common causes, has additionally the “ $X \circ \rightarrow Y$ ” arrow, meaning Y is not an ancestor of X , i.e., X potentially causes Y (common cause not ruled out). The undirected edge “ $X - Y$ ” in the output of FCI is often graphically represented by “ $X \circ - Y$ ” in many literatures. For the sake of simplicity, we use the notation of “ $X - Y$ ” throughout this thesis.

5.4. Simulated experiments with orientation heuristics

	2-Bit-IndDet	2-Bit-IndPro	2-Bit-DepPro	3-Bit-IndDet	3-Bit-IndPro	3-Bit-DepPro
X_1	$P(X_1) = 0.6$	$P(X_1) = 0.6$	$P(X_1) = 0.6$	$P(X_1) = 0.6$	$P(X_1) = 0.6$	$P(X_1) = 0.6$
X_2	$P(X_2) = 0.5$	$P(X_2) = 0.5$	$(1 - X_1)_{0.1}$	$P(X_2) = 0.5$	$P(X_2) = 0.5$	$(1 - X_1)_{0.1}$
X_3	$\text{OR}_0\{X_{1,2}\}$	$\text{OR}_{0.2}\{X_{1,2}\}$	$\text{OR}_{0.2}\{X_{1,2}\}$	$P(X_3) = 0.4$	$P(X_3) = 0.4$	$\text{OR}_{0.2}\{X_{1,2}\}$
X_4	–	–	–	$\text{OR}_0\{X_{1,2,3}\}$	$\text{OR}_{0.2}\{X_{1,2,3}\}$	$\text{OR}_{0.2}\{X_{1,2,3}\}$

Table 5.2.: Parameters of models linked by 2/3-bit deterministic and probabilistic OR gates. $P(X_i)$ is shorthand for $P(X_i = 1)$. $\text{OR}_0\{X_{1,\dots,i}\}$ denotes a deterministic OR gate with X_1, \dots, X_i as inputs; $\text{OR}_{0.2}\{X_{1,\dots,i}\}$ denotes the noisy OR gate as in Eq. (3.6) with $r = 0.2$. $(1 - X_1)_{0.1}$ depicts a variable whose value is with probability 0.1 given by an inverse of X_1 and with probability 0.9 by uniform noise.

	$\frac{\widehat{\mathbb{H}}_{X_2 X_3 X_1}}{\widehat{\mathbb{H}}_{X_2 X_3}}$	$\frac{\widehat{\mathbb{H}}_{X_1 X_3 X_2}}{\widehat{\mathbb{H}}_{X_1 X_3}}$	$\frac{\widehat{\mathbb{H}}_{X_1 X_2 X_3}}{\widehat{\mathbb{H}}_{X_1 X_2}}$
2-Bit-IndDet	$0.0709/0.0377 = 1.8790$	$0.1285/0.0651 = 1.9732$	$\frac{0.0321}{7.1182 \times 10^{-6}} = \mathbf{4502.8}$
2-Bit-IndPro	$0.0605/0.0454 = 1.3316$	$0.0409/0.0350 = 1.1665$	$\frac{0.0096}{1.3740 \times 10^{-5}} = \mathbf{698.49}$
2-Bit-DepPro	$0.0656/0.0311 = 2.1050$	$0.0756/0.0461 = 1.6411$	$0.0305/0.0015 = \mathbf{20.0435}$

Table 5.3.: Estimated kernel dependence measures of a random sample from 2-bit OR gates (see Tab. 5.2). In all cases, ratio $\frac{\widehat{\mathbb{H}}_{X_1 X_2 | X_3}}{\widehat{\mathbb{H}}_{X_1 X_2}}$ achieves the maximum, which is taken as a hint that X_3 is the output, X_1 and X_2 are inputs.

5. From Magnitude of Dependences to Causal Structure

	2-Bit-IndDet	2-Bit-IndPro	2-Bit-DepPro	3-Bit-IndDet	3-Bit-IndPro	3-Bit-DepPro
True Model						
OPB						
PC						
BN-PC						
Exhaustive Search						
Greedy Search						
MWST+Greedy Search						
MWST+K2						
MCMC						

Table 5.4.: The underlying true model and outputs generated by different algorithms (see Appendix C.2 for details about algorithms). The first row illustrates the generating models in graphical representation (see Tab. 5.2 for parameters). Rows 2 to 9 show graphical outputs of algorithms or combinations of algorithms. Each graph consists of at most 4 nodes, which are represented by circles: X_1 : top left, X_2 : top right, X_3 : bottom left, X_4 : bottom right.

5.4. Simulated experiments with orientation heuristics

As seen from Tab. 5.4, both constraint-based and score-based Bayesian methods achieved quite good results in learning 2-bit OR. In learning 3-bit OR, the constraint-based algorithms seem often to perform better than score-based Bayesian methods. In learning 2-Bit-DepPro and 3-Bit-DepPro, OPB detected the connection between X_1 and X_2 (Tab. 5.4, row OPB), whereas PC wrongly removed the edge in both cases (Tab. 5.4, row PC). Had PC detected the dependence between X_1 and X_2 correctly, it would not have been able to orient any edge. The result would be a fully connected and undirected graph. In contrast, although all dependences are correctly captured (actually, no conditional independence is verified) by thresholding kernel measures, OPB provides useful hints about orientation in the structure. Both PC and OPB have left edges undirected in 3-Bit-DepPro. OPB performs slightly better than PC in the sense that the former oriented as many edges as PC, but no edges are wrongly removed.

5.4.2. Simulated data from models with hidden common causes

The issue of hidden common causes is not the main concern of this thesis. Recall that we usually made the assumption of causal sufficiency, i.e., all the common causes of measured variables are measured. Nonetheless, we would like, by means of some simulated examples, to explore what happens when OPB is applied to situations when the causal sufficiency does not hold.

We study five generating causal structures as shown in the first row of Tab. 5.5. For some reason, variable L contained in each structure, which is a common cause of Y_1 and Y_2 or a common cause of Y_1 , Y_2 , and Y_3 , cannot be measured. The second row of Tab. 5.5 shows the voting procedure of OPB in the case that L is not observed, when the true adjacency structure is known. The expected outputs of OPB is shown in the third row of Tab. 5.5.

The latent variable L in the first two models as shown in column 1 and 2 in Tab. 5.5 cannot be indicated by OPB, since no conflicts will be generated in the voting procedure of the collider identification by leaving L out. In contrast, the latent variable L in models as shown in column 3 to 5 can be, in principle, identified, because the collider identification causes conflicting orientation information. Note that, if we group X_2 and X_3 in the model in column 4 to one variable, we would have the same model in column 3. For this reason, we omit the model in column 3 in our simulations and demonstrate experiments with models in column 4 and 5. The variables are linked by OR gates.

We define the first model, graphically presented in row 1 column 4 of Tab. 5.5, by a 2-Bit-IndDet OR gate (defined in Tab. 5.2) with X_1 and L as inputs and Y_1 as output, and a 3-Bit-IndDet OR gate (defined in Tab. 5.2) with X_1 , X_2 , and L as inputs and Y_2 as output. The second model, graphically presented in row 1 column 5 of Tab. 5.5, is defined by three 2-Bit-IndPro OR gate (defined in Tab. 5.2) with L and X_i ($i=1, 2, 3$) as inputs and Y_i as output. We generated 200 data points from both models and performed OPB on data without measuring variable L .

As the statistics of the resulting structures in Tab. 5.6 and Tab. 5.7 showed, OPB correctly detected the spurious associations between Y_i and Y_j ($i, j = 1, 2, 3$) in the majority cases and oriented the edges between them bi-directed. As expected, the result of the model without noise (76.5% in the first row of Tab. 5.6) is more reliable than that with noise (ca. 43% in the first three rows of Tab. 5.7). Note that, in the second model (Tab. 5.5, row 1 column 5), conditional independence between Y_i and Y_j is erroneously detected in 26 – 27% of the cases (see first three

5. From Magnitude of Dependences to Causal Structure

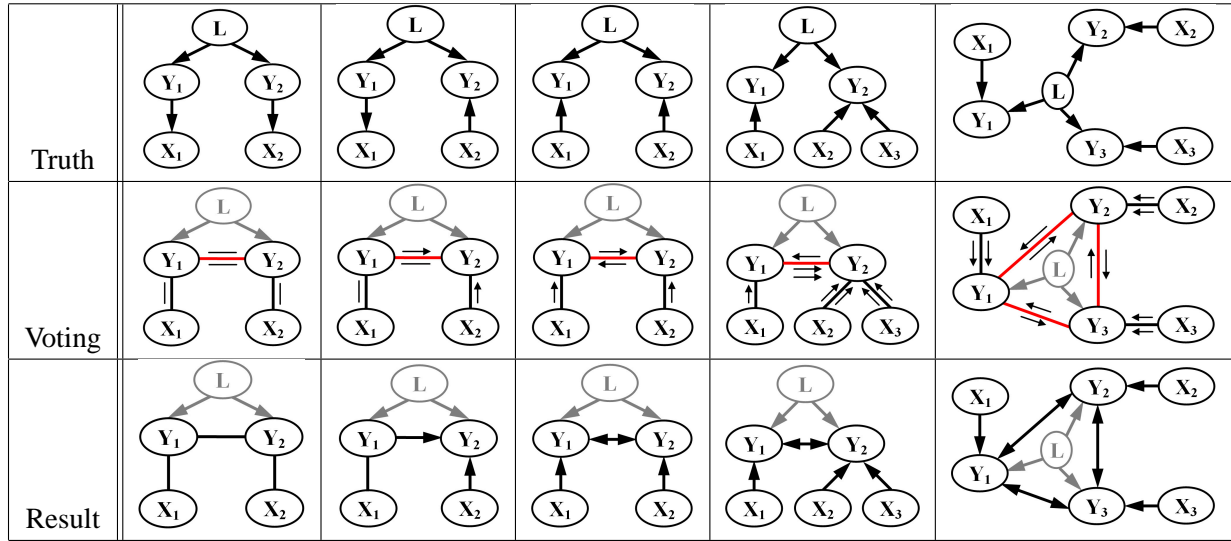


Table 5.5.: Five different generating models (first row) containing a common cause L , which is not measured. The corresponding skeleton of the observed variables is oriented by the voting procedure of OPB. The second row visualizes the number of votes on the basis of collider identification for each edge. The third row shows the final output of OPB.

rows of Tab. 5.7). That means, the most errors occur in thresholding independence measures, rather than in the orientation step. Although the examples suggest that bi-directed edges in the output of OPB could be traced back to hidden common causes, conflicting voting results do not automatically indicate hidden variables. For instance, it may happen that the underlying model is indeed cyclic.

5.4.3. Simulated data from Asia network

In this experiment, we apply our criteria to a larger network and focus on the “voting triples”. We use the Asia network, an expert-designed causal network with logical links, to sample data. This model was first introduced by Lauritzen et al. [100] who have specified reasonable transition properties for each variable given its parents. Due to deterministic relationships between variables, learning structure from independence constraints have various problems (see [180] for more details and discussions).

The underlying structure (Fig. 5.5) expresses the following known qualitative medical knowledge. DYSPTNOEA may be due to tuberculosis (TUB), LUNG cancer (together TUB/LUNG) or BRONCHITIS, or none of them, or more than one of them. A recent visit to ASIA increases the chances of tuberculosis, while SMOKING is known to be a risk factor for both lung cancer and bronchitis. The results of a single chest X-RAY do not discriminate between lung cancer and tuberculosis, and neither does the presence or absence of DYSPTNOEA.

We first consider the simpler situation: the true adjacency structure is known. We test OPB on

5.4. Simulated experiments with orientation heuristics

Correct Pattern	• •	• - •	• → •	• ← •	• ↔ •
$Y_1 \leftrightarrow Y_2$	3.5	0.6	12.6	6.8	76.5
$X_1 \rightarrow Y_1$	0.0	11.6	79.2	6.6	2.6
$X_2 \rightarrow Y_2$	0.0	2.5	87.8	7.4	2.3
$X_3 \rightarrow Y_2$	0.0	2.6	90.8	4.8	1.8
$X_1 \quad X_2$	95.3	1.7	1.7	0.9	0.4
$X_1 \quad X_3$	95.3	2.0	1.2	1.0	0.5
$X_1 \quad Y_2$	92.9	0.0	1.8	3.4	1.9
$X_2 \quad X_3$	94.5	3.9	1.0	0.6	0.0
$X_2 \quad Y_1$	94.7	0.1	0.7	3.9	0.6
$X_3 \quad Y_1$	96.4	0.0	0.7	2.5	0.4

Table 5.6.: OPB is applied to causal models with a hidden common cause L (Tab. 5.5, column 4). The variables are linked by a 2-Bit and a 3-Bit OR gate (see text). “•” is a placeholder for an observed variable. The entries are percentages of 1000 replications having the considered patterns as output.

Correct Pattern	• •	• - •	• → •	• ← •	• ↔ •
$Y_1 \leftrightarrow Y_2$	26.6	1.7	13.7	14.5	43.5
$Y_1 \leftrightarrow Y_3$	26.9	2.2	13.4	13.7	43.7
$Y_2 \leftrightarrow Y_3$	27.3	2.2	13.5	13.1	43.9
$X_1 \rightarrow Y_1$	0.0	17.4	72.5	6.8	3.3
$X_2 \rightarrow Y_2$	0.0	15.0	74.5	5.5	5.0
$X_3 \rightarrow Y_3$	0.0	16.4	74.3	5.4	3.9
$X_1 \quad X_2$	95.3	1.0	1.4	1.3	1.0
$X_1 \quad X_3$	94.9	1.6	1.2	1.2	1.1
$X_1 \quad Y_2$	97.4	0.0	1.2	0.9	0.5
$X_1 \quad Y_3$	97.2	0.2	0.6	1.1	0.9
$X_2 \quad X_3$	95.5	1.2	0.9	1.3	1.1
$X_2 \quad Y_1$	96.6	0.0	1.4	1.1	0.9
$X_2 \quad Y_3$	96.2	0.1	1.2	1.5	1.0
$X_3 \quad Y_1$	97.0	0.0	1.3	1.1	0.6
$X_3 \quad Y_2$	96.6	0.0	1.2	0.9	1.3

Table 5.7.: OPB is applied to causal models with a hidden common cause L (Tab. 5.5, column 5). The variables are linked by three 2-Bit noisy OR gates (see text). “•” is a placeholder for an observed variable. The entries are percentages of 1000 replications having the considered patterns as output.

5. From Magnitude of Dependences to Causal Structure

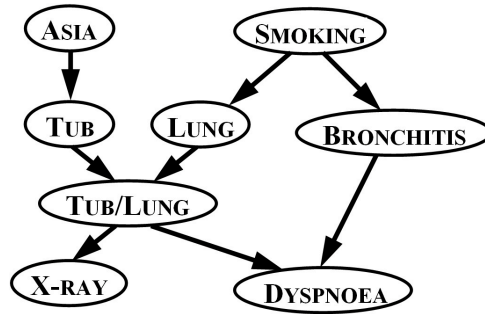


Figure 5.5.: Graphical representation of medical knowledge by Asia network. Each node has two possible states representing responses “yes” and “no”. In total, the underlying domain contains $2^8 = 256$ possible states.

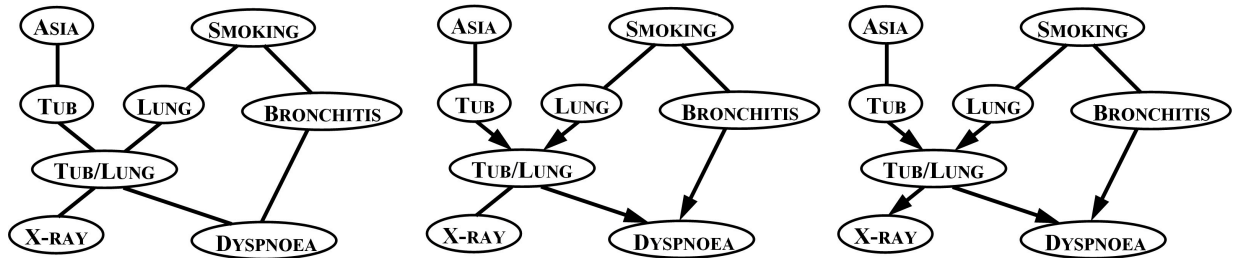


Figure 5.6.: Stepwise results of OPB (Fig. 5.4) with the prior knowledge of the true adjacency structure (leftmost plot). The middle plot illustrates the result after Step 1 of OPB. The rightmost graph illustrates the result after Step 2 of OPB. Step 3 of OPB cannot further orient the remaining undirected edges.

the true adjacency structure (Fig. 5.6, leftmost). Tab. 5.8 shows the statistics after 1000 replications for the 11 involved “voting triples”, which are required for recovering the orientation of the 8 arrows in Fig. 5.5. To test the sensitivity of the empirical dependence measures to changes in sample size, we conducted the experiments for datasets of a sample size of 200 or 400. As seen from Tab. 5.8, the frequency with which one of the three ratios achieves the maximum is quite robust with respect to the sample size. Extensive statistics of the orientation of the 8 edges by OPB can be found in Tab. C.5 in Appendix C.3. Taking ratios and λ of different levels into account, Fig. 5.6 shows the stepwise results of OPB. Based on the correct corresponding skeleton (Fig. 5.6, leftmost), step 1 of OPB (Fig. 5.4) detected two unshielded colliders $TUB \rightarrow TUB/LUNG \leftarrow LUNG$ and $TUB/LUNG \rightarrow DYSPNOEA \leftarrow BRONCHITIS$ (see middle plot of Fig. 5.6). The undirected edge $TUB/LUNG - X-RAY$ can be further directed by Step 2, since it is implied by the first detected collider (see rule 1 in Fig. B.1 and Fig. B.2 in Appendix B.3). The three remaining undirected edges (Fig. 5.6, rightmost) are due to the limitations of methods are based on collider identification. The rightmost plot is what such methods can maximally achieve.

5.4. Simulated experiments with orientation heuristics

(X, Y, Z)	Sample Size	(m_X, m_Y, m_Z)	(r_X, r_Y, r_Z)	Voting Result
(ASIA, TUB, TUB/LUNG)	200	(1.0188, 0.7812, 1.2038)	(39.5, 0.1, 60.4)	no voting
	400	(1.0143, 0.7446, 1.1928)	(42.2, 0.3, 57.5)	
(TUB, LUNG, TUB/LUNG)	200	(1.0409, 1.2553, 206.4190)	(0.3, 1.7, 98.0)	TUB \rightarrow TUB/LUNG
	400	(1.0305, 1.2254, 250.4862)	(0.1, 0.4, 99.5)	LUNG \rightarrow TUB/LUNG
(TUB, TUB/LUNG, X-RAY)	200	(0.7453, 0.0002, 0.2982)	(84.4 , 3.3, 12.3)	no voting
	400	(0.7414, 0.0005, 0.2584)	(97.2 , 1.6, 1.2)	
(TUB, TUB/LUNG, DYSPTNOEA)	200	(0.8132, 0.2252, 1.6511)	(0.5, 26.7, 72.8)	no voting
	400	(0.7512, 0.1276, 1.2778)	(0.1, 18.9, 81.0)	
(SMOKING, LUNG, BRONCHITIS)	200	(0.8555, 0.9747, 0.8748)	(38.1, 58.8 , 3.1)	no voting
	400	(0.5286, 0.9918, 0.8560)	(36.3, 62.4 , 1.3)	
(SMOKING, LUNG, TUB/LUNG)	200	(1.4851, 0.0117, 0.0335)	(98.2 , 1.0, 0.8)	no voting
	400	(1.5176, 0.0054, 0.0273)	(100 , 0, 0)	
(SMOKING, BRONCHITIS, DYSPTNOEA)	200	(0.8134, 0.0908, 0.2650)	(97.3 , 2.7, 0)	no voting
	400	(0.8091, 0.0536, 0.2500)	(100 , 0, 0)	
(LUNG, TUB/LUNG, X-RAY)	200	(0.0241, 4.3×10^{-6} , 0.2595)	(2.2, 0, 97.8)	no voting
	400	(0.0267, 7.3×10^{-6} , 0.2492)	(0.2, 0, 99.8)	
(LUNG, TUB/LUNG, DYSPTNOEA)	200	(0.0175, 0.0078, 1.2523)	(0.4, 1.0, 98.6)	no voting
	400	(0.0248, 0.0039, 1.2235)	(0, 0.2, 99.8)	
(BRONCHITIS, TUB/LUNG, DYSPTNOEA)	200	(0.9260, 1.0527, 4.4805)	(1.8, 24.5, 73.7)	BRONCHITIS \rightarrow DYSPTNOEA
	400	(0.9781, 1.0576, 6.6893)	(0.2, 19.4, 80.4)	TUB/LUNG \rightarrow DYSPTNOEA
(TUB/LUNG, X-RAY, DYSPTNOEA)	200	(0.0640, 0.2373, 1.2188)	(8.2, 0.5, 91.3)	no voting
	400	(0.0270, 0.2477, 1.2165)	(3.6, 0, 96.4)	

Table 5.8.: Empirical kernel dependence measures of data generated from Asia network. The

shorthand m_X , m_Y or m_Z depicts the median of $\frac{\hat{H}_{ZY|X}}{\hat{H}_{ZY}}$, $\frac{\hat{H}_{ZX|Y}}{\hat{H}_{ZX}}$ or $\frac{\hat{H}_{YX|Z}}{\hat{H}_{YX}}$, respectively. The shorthand r_X , r_Y or r_Z depicts the percentage of cases, where value $\frac{\hat{H}_{ZY|X}}{\hat{H}_{ZY}}$, $\frac{\hat{H}_{ZX|Y}}{\hat{H}_{ZX}}$ or $\frac{\hat{H}_{YX|Z}}{\hat{H}_{YX}}$ achieves the maximum. All entries are calculated on the basis of 1000 replications and for a sample size of 200 or 400. The last column shows the voting according to Step 1.1 of OPB (Fig. 5.4).

5. From Magnitude of Dependences to Causal Structure

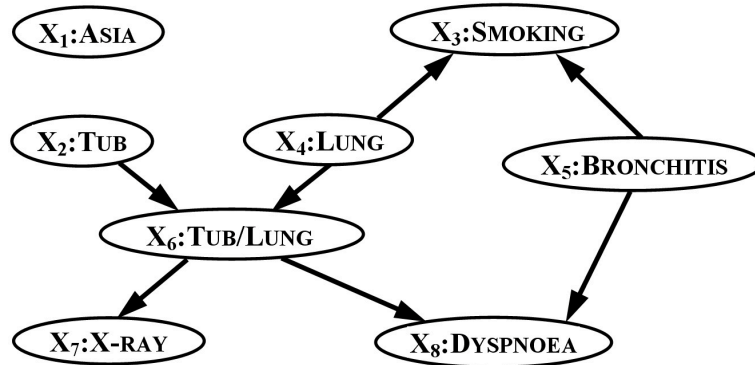


Figure 5.7.: Result of OPA+K2 (K2 with initial ancestral order induced by OPA on a complete adjacency structure) algorithm in graphical representation. The detailed statistics of structures detected by OPA+K2 is collected in Tab. C.6 in Appendix C.3.

Now, let us consider a more challenging situation, i.e., testing our orientation heuristics on a skeleton with redundant edges. As an extreme case, we took the complete adjacency structure and ran OPA (Fig. 5.4) to learn orientation without any information about the adjacency structure. As seen from Tab. C.5 in Appendix C.3, apart from unnecessary edges, three edges, namely $ASIA \rightarrow TUB$, $SMOKING \rightarrow LUNG$ and $SMOKING \rightarrow BRONCHITIS$, are often wrongly directed and the other 5 arrows can be discovered correctly.

Based on the resulting ancestral order by OPA, we could use other techniques to prune unnecessary edges, since hypothesis tests have problems due to deterministic relations between variables. More discussions about these problems of learning Asia network can be found in [180]. Setting an appropriate cut-off value for thresholding kernel dependence measure is also very difficult due to the small sample size. For purely discrete (in particular, binary) domains, K2 [40] is a well-known score-based Bayesian approach for this purpose, if an initial ordering of variables is given. The power of such score-based Bayesian approaches can efficiently take a very large number of data points into account and make pruning of edges accurate.

Since the output of OPA can contain undirected edges, the ancestral order given by the output is sometimes not unique. We start K2 with an initial ancestral order induced by OPA. If the order induced by OPA is not unique, we chose one of them randomly. The so-called OPA+K2 performs well in learning Asia network from sampled data. The output (Fig. 5.7) contains no undirected edges and the missing arc from ASIA to TUB is probably due to the weak dependency between them in datasets of such small sample sizes. Although the edges from SMOKING to LUNG and BRONCHITIS have the wrong orientation, the result contains no unnecessary edges. OPA+K2+OPB (learning adjacency structure by starting K2 with the initial ancestral order induced by OPA and then re-orienting the corresponding adjacency structure of the result of OPA+K2 by OPB) would revise both wrong directed edges into undirected.

Tab. C.6 in Appendix C.3 summarizes how often an arrow is detected by OPA+K2 after 1000 replications. An extensive comparison of well-known constraint-based and Bayesian algorithms with respect to the Asia network is provided by Leray et al. (see [101] Fig. 2 for experimental

5.4. Simulated experiments with orientation heuristics

results). We can see that OPA+K2 performs better than K2 combined with other initialization of causal orders, which indicates that OPA provides quite reliable causal ordering. Furthermore, OPA+K2 is competitive with other Bayesian methods and PC. The performance of PC is unsatisfactory in the sense that several edges are completely missing. Repeated experiments with a sample size from 500 to 5000 show that 3–5 from the total 8 edges are always missing. This result is actually traced back to the independence test of PC. Actually, with regard to the sample size, the result obtained by OPA+K2 is better than all 12 algorithms listed in Fig. 2 in [101] concerning the so-called “editing measures”,⁴ since our result has an editing measure of merely 3. Most notably, the result of such a hybrid algorithm can be reliably achieved with datasets of moderate sample sizes.

The experiments with Asia network showed that OPA is reasonably robust with respect to redundant edges. On the other hand, the given adjacency structure influences the reliability of the orientation heuristics. A fixed cut-off value for thresholding dependence measure does not work well in a complex network. The combination of learning orientation in the structure by our orientation heuristics and pruning edges by a score-based Bayesian approach alleviates the problem.

5.4.4. Simulated data from functional models

In this section, we focus on the quantitative comparison of marginal and conditional dependence measures on continuous domains. We sampled datasets from a model as shown in Fig. 5.8. X is sampled from a gamma distribution $\mathcal{G}(2, 2)$. Y is a quadratic function of X added with an independent Gaussian noise:

$$Y \propto \left(\frac{X}{10}\right)^2 + \mathcal{N}(0, \kappa).$$

Variable Z is a function of X and Y :

$$Z \propto \cos(\pi X) + \ln(|Y|) + \mathcal{N}(0, 0.01).$$

One can imagine that X influences Z like a “seasonal” cycle, whereas Y adds a logarithmic bias. The mutual dependence between X and Y decreases as $\kappa > 0$ increases. Fig. 5.9 illustrates the dependence between X and Z with the change of κ from 10^{-3} to 10^2 .

We generated 200 data points from functional models with $\kappa \in \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2\}$. We chose one random sample and computed the empirical conditional cross-covariance operators and the HS norms. We chose different regularization constants ϵ from 10^{-10} to 1 to compute the measures according to Eq. (2.11) and Eq. (2.12). Fig. 5.10 visualizes the resulting ratios of dependence measures of this example. We can observe that if the regularizer ϵ is not too large, the ratios are insensitive to the choice of ϵ in $[10^{-10}, 10^{-2}]$. Thus, we set regularizer $\epsilon = 10^{-5}$ throughout this thesis.

Tab. 5.9 summarizes the dependence measures for this example. For most values of κ , Crite-

⁴Editing measure [101] is defined as the length of the minimal sequence of operators needed to transform the original graph into the resulting one. Operators are edge-insertion, edge-deletion and edge reversal.

5. From Magnitude of Dependences to Causal Structure

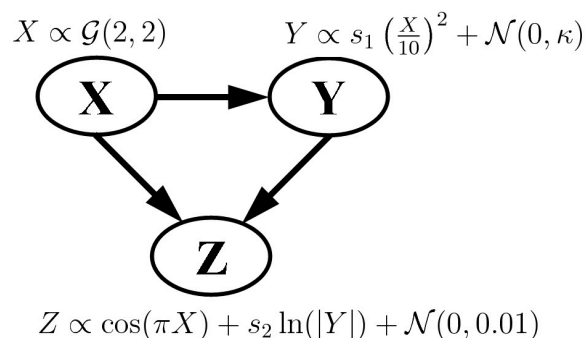


Figure 5.8.: A functional model with shielded collider structure. (s_1, s_2) takes the values from $\{(\pm 1, \pm 1)\}$ and induces both positive and negative dependence. In the first experiment, we set $(s_1, s_2) = (+1, +1)$.

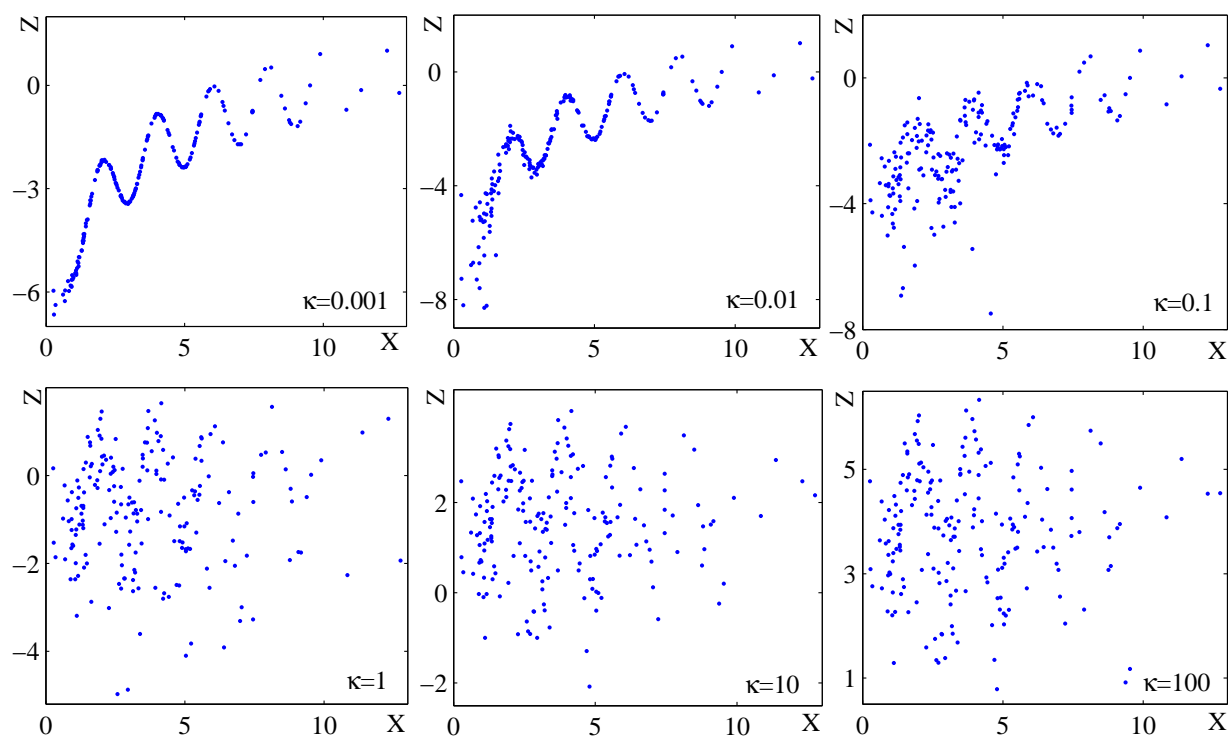


Figure 5.9.: Toy data sampled from a functional model of three continuous variables (Fig. 5.8) with 6 different parameters κ . Variable X is the “seasonal” cyclic influences of variable Z . The smaller κ is, the clearer the “seasonal” effect of X on Z can be recognized.

5.4. Simulated experiments with orientation heuristics

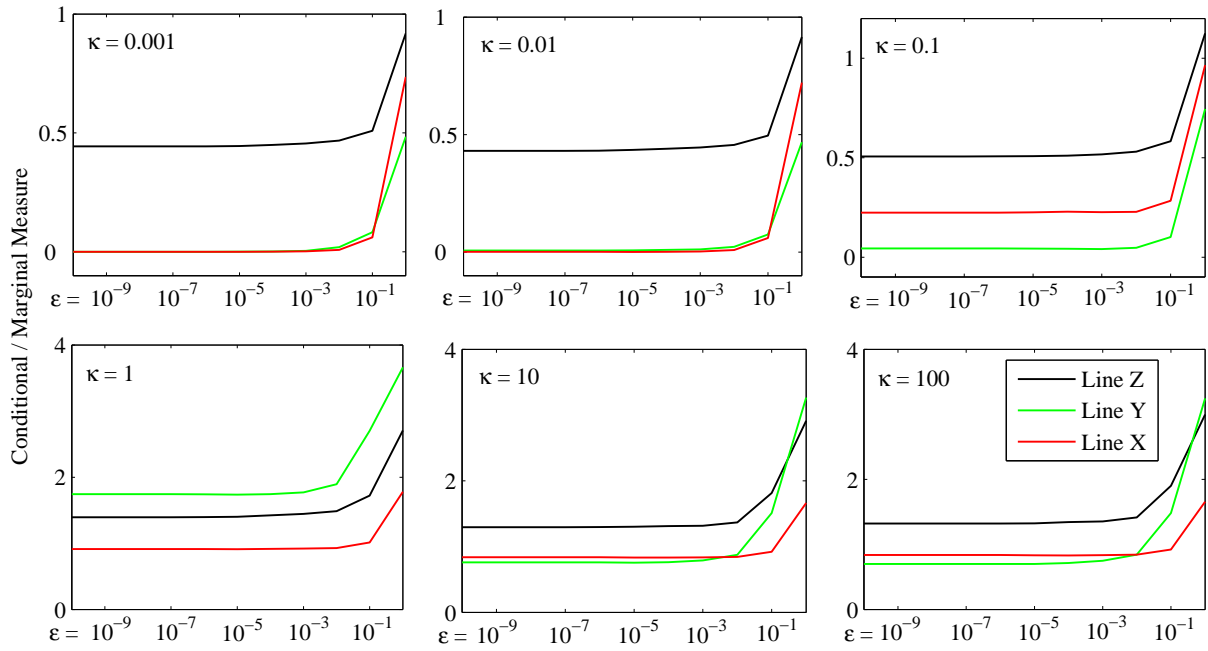


Figure 5.10.: 200 data points sampled from a functional model of three continuous variables with different κ . “Line Z” visualizes the values of ratio $\frac{\hat{\mathbb{H}}_{YX|Z}}{\hat{\mathbb{H}}_{YX}}$, computed with different regularizers ϵ from 10^{-10} to 1, “Line Y” illustrates the values of ratio $\frac{\hat{\mathbb{H}}_{ZX|Y}}{\hat{\mathbb{H}}_{ZX}}$, and “Line X” corresponds to the values of ratio $\frac{\hat{\mathbb{H}}_{ZY|X}}{\hat{\mathbb{H}}_{ZY}}$.

5. From Magnitude of Dependences to Causal Structure

κ	$\widehat{\mathbb{H}}_{ZY X} / \widehat{\mathbb{H}}_{ZY}$	$\widehat{\mathbb{H}}_{ZX Y} / \widehat{\mathbb{H}}_{ZX}$	$\widehat{\mathbb{H}}_{YX Z} / \widehat{\mathbb{H}}_{YX}$
0.001	$\frac{1.3053 \times 10^{-5}}{0.0363} = 0.0004$	$\frac{3.8863 \times 10^{-5}}{0.0572} = 0.0007$	0.0300/0.0674 = 0.4447
0.01	$\frac{3.6994 \times 10^{-5}}{0.0371} = 0.0010$	0.0004/0.0560 = 0.0077	0.0292/0.0671 = 0.4348
0.1	0.0075/0.0331 = 0.2258	0.0012/0.0276 = 0.0436	0.0238/0.0468 = 0.5077
1	0.0182/0.0200 = 0.9122	0.0020/0.0012 = 1.7367	0.0072/0.0051 = 1.4014
10	0.0186/0.0224 = 0.8326	0.0019/0.0025 = 0.7564	0.0028/0.0022 = 1.3019
100	0.0188/0.0226 = 0.8317	0.0018/0.0025 = 0.7006	0.0026/0.0020 = 1.3238

Table 5.9.: Empirical kernel dependence measures for one sample of the functional model of three continuous variables as shown in Fig. 5.8. $\frac{\widehat{\mathbb{H}}_{YX|Z}}{\widehat{\mathbb{H}}_{YX}}$ in most cases (excepting $\kappa = 1$) achieves the maximum, which indicates Z being a collider between X and Y .

tion 2 identified the correct colliders, expecting the case $\kappa = 1$, where the voting for a collider is obviously not consistent with the generating model. Note that $\widehat{\mathbb{H}}_{ZY|X}$ ($\kappa = 0.01, 0.001$) or $\widehat{\mathbb{H}}_{ZX|Y}$ ($\kappa = 0.001$) lies below the pre-specified cut-off values 10^{-4} for dependence, so that the conditional dependence would not be captured.

To demonstrate that the above conclusion is not only based on some particular sampling, we replicated our experiments 1000 times with datasets of size 200 sampled from each of the 24 functional models:

$$y = s_1 \left(\frac{x}{10}\right)^2 + \epsilon_y \quad \text{and} \quad z = \cos(\pi x) + s_2 \ln(|y|) + \epsilon_z$$

with different combinations of $(s_1, s_2) \in \{(\pm 1, \pm 1)\}$ and $\kappa \in \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2\}$. The 4 different values of (s_1, s_2) induce all combinations of negative and positive correlations. 1000 replications (see Tab. 5.10 for results) show that our method yields the same results for various combinations of (s_1, s_2) . We conjecture that the voting for a collider agrees with the generating causal structure for most values of κ and the majority of the samples. When κ is in a small interval close to 1, however, we mainly obtain wrong votes.

5.5. Kernel-based causal learning algorithm (KCL)

We have shown so far that our orientation heuristics via kernel dependence measures can provide some good initial information about the orientation in the structure. On the other side, if we can learn somewhat good adjacency structure, the performance of our orientation heuristics can be improved. For this reason, we combine the statistical test of independence with the orientation heuristics, and propose a kernel-based causal learning algorithm (KCL).

Like IC, the KCL algorithm can be broken into two phases: an adjacency phase and an orientation phase. In the adjacency phase, a complete undirected adjacency structure over all variables is initially constructed and the edges $X - Y$ are removed if some set $S_{XY} \subseteq \mathcal{V} \setminus \{X \cup Y\}$ can be found such that the constraint $X \perp\!\!\!\perp Y \mid S_{XY}$ can be verified. In search for S_{XY} , the orienta-

5.5. Kernel-based causal learning algorithm (KCL)

κ	(s_1, s_2)	$\widehat{\mathbb{H}}_{ZY X} / \widehat{\mathbb{H}}_{ZY}$	$\widehat{\mathbb{H}}_{ZX Y} / \widehat{\mathbb{H}}_{ZX}$	$\widehat{\mathbb{H}}_{YX Z} / \widehat{\mathbb{H}}_{YX}$	% Correct
0.001	(+1, +1)	0.0002 ± 0.0001	0.0009 ± 0.0006	0.3387 ± 0.0571	100%
	(-1, +1)	0.0002 ± 0.0001	0.0009 ± 0.0006	0.3387 ± 0.0572	100%
	(+1, -1)	0.0002 ± 0.0001	0.0023 ± 0.0009	0.3473 ± 0.0578	100%
	(-1, -1)	0.0002 ± 0.0001	0.0024 ± 0.0009	0.3473 ± 0.0576	100%
0.01	(+1, +1)	0.0006 ± 0.0003	0.0040 ± 0.0018	0.3482 ± 0.0582	100%
	(-1, +1)	0.0006 ± 0.0003	0.0040 ± 0.0019	0.3480 ± 0.0586	100%
	(+1, -1)	0.0009 ± 0.0005	0.0034 ± 0.0014	0.3623 ± 0.0598	100%
	(-1, -1)	0.0009 ± 0.0004	0.0034 ± 0.0015	0.3621 ± 0.0588	100%
0.1	(+1, +1)	0.1576 ± 0.0576	0.0285 ± 0.0164	0.4594 ± 0.0755	100%
	(-1, +1)	0.1559 ± 0.0588	0.0283 ± 0.0167	0.4603 ± 0.0764	100%
	(+1, -1)	0.1685 ± 0.0630	0.0334 ± 0.0219	0.4916 ± 0.0780	100%
	(-1, -1)	0.1660 ± 0.0635	0.0318 ± 0.0196	0.4889 ± 0.0787	100%
1	(+1, +1)	0.9465 ± 0.0941	1.7120 ± 0.7932	1.5100 ± 0.4535	42.0%
	(-1, +1)	0.9412 ± 0.0936	1.6951 ± 0.7907	1.4904 ± 0.4027	43.1%
	(+1, -1)	0.9546 ± 0.0936	1.7367 ± 0.7764	1.5463 ± 0.4314	44.1%
	(-1, -1)	0.9531 ± 0.0943	1.7064 ± 0.7728	1.5221 ± 0.3886	46.4%
10	(+1, +1)	1.0077 ± 0.0920	2.1533 ± 0.8812	2.4537 ± 0.9278	66.4%
	(-1, +1)	1.0095 ± 0.0930	2.1714 ± 0.8729	2.4462 ± 0.9289	66.1%
	(+1, -1)	1.0112 ± 0.0952	2.1949 ± 0.9827	2.4753 ± 0.9742	65.9%
	(-1, -1)	1.0113 ± 0.0948	2.1937 ± 1.0127	2.4610 ± 0.9544	65.7%
100	(+1, +1)	1.0100 ± 0.0924	2.1577 ± 0.8712	2.4623 ± 0.9302	66.5%
	(-1, +1)	1.0102 ± 0.0925	2.1610 ± 0.8649	2.4645 ± 0.9307	66.6%
	(+1, -1)	1.0115 ± 0.0948	2.1936 ± 0.9999	2.4841 ± 0.9864	66.1%
	(-1, -1)	1.0115 ± 0.0948	2.1929 ± 1.0046	2.4862 ± 0.9882	65.8%

Table 5.10.: Experiments with 200 data points sampled from each of 24 functional models (Fig. 5.8) with 6 different κ and 4 different (s_1, s_2) . Shorthands “ $m \pm \sigma$ ” denote the median m and standard deviation σ after 1000 replications. The last column shows how often (in percentage) $\frac{\widehat{\mathbb{H}}_{YX|Z}}{\widehat{\mathbb{H}}_{YX}}$ achieves the maximum, which indicates Z being a collider between X and Y .

5. From Magnitude of Dependences to Causal Structure

tion heuristics OPA presented in Fig. 5.3 is performed to learn an auxiliary graph without any independence decision. We reduce the search space of S_{XY} by the potential ancestor condition (Definition 21) with respect to the auxiliary graph.

The auxiliary graph is used in the following iterative scheme. We apply the orientation heuristics to an adjacency structure and check the relevant conditioning subsets according to the potential ancestor condition with respect to the auxiliary graph. If some conditional independence is verified by the kernel independence test, the corresponding edge will be removed. Hence, a new adjacency structure is obtained for the next iteration. The iterative loop with arcs progressively removed converges if no more edges can be removed. Consequently, the absence of an edge in the final output represents the presence of conditional independence, but not vice versa.

Once the adjacency structure over all observed variables has been estimated by the first phase, the orientation phase OPB (Fig. 5.4) is begun. The first step of the orientation is to examine unshielded triples, i.e., $X-Z-Y$, and consider whether to orient them as an unshielded colliders on Z via the voting procedure by a unanimous vote. Once all such unshielded triples have been checked, a series of orientation rules (see Fig. B.1 in Appendix B.3) is applied to orient any edges whose directions are implied by previous directions. If there are still remaining undirected edges, we examine all shielded triples and identify the colliders of them via the voting procedure by a unanimous vote again. The complete scheme of KCL is shown in Fig. 5.11. Fig. 5.12 illustrates stepwise results of learning the causal structure as shown in Fig. 1.1, when KCL is applied.

Due to the potential ancestor condition and the auxiliary graph \mathcal{G} learned by OPA, the number of hypothesis tests in the adjacency phase can be reduced, since only these constraints will be tested, which are consistent with the directed auxiliary graph. The orientation phase (step 2 of KCL) will terminate in the worst-case scenario (complete skeleton) after $\binom{N}{3}$ calls (evaluating all triples).

The final output of KCL is represented by three kinds of edges: “–” (undirected) meaning no evidence for both directions; “→” (directed) meaning consistent evidence for one direction; “↔” (bi-directed) meaning evidence for both directions. The bi-directed edges in the output of our KCL indicate conflicting voting results, which might be traced back to any violation of assumptions. The presence of hidden common causes in the true model is one possibility of such violation. Note that a bi-directed edge in a maximal ancestral graph [133] is explicitly used to represent the presence of a hidden common cause.

Note that, having taken the degree of dependence into account, the orientation of KCL is, on the one hand, less sensitive to the type I errors, because KCL does not use the conditional set S_{XY} that is found by hypothesis tests to infer the orientation. On the other hand, due to step 2.4 in the orientation step, KCL is able to provide some orientation even in the complete adjacency structure (maybe due to a high level of type I error of hypothesis tests).

5.5.1. Some implementation issues of KCL

As discussed in Section 4.2, learning structure from independence constraints has many problems in practice. In particular, as the number of variables increases, the number of possible non-trivial independence constraints grows exponentially. Only a small set of all possible constraints can

5.5. Kernel-based causal learning algorithm (KCL)

Input: Data of a set of variables \mathcal{V} .

Step 1: Learning adjacency structure.

1.1 Test unconditional independence: Initialize \mathcal{G} by a complete undirected graph. For all edges between variables X and Y , test the unconditional independence hypothesis $X \perp\!\!\!\perp Y$ via kernel dependence measures based on data. Remove edge $X - Y$ in \mathcal{G} , if the hypothesis is accepted. The result is a skeleton (an undirected graph) \mathcal{G} .

1.2 Construct auxiliary graph: Orient skeleton \mathcal{G} by OPA (Fig. 5.3), providing an auxiliary graph \mathcal{G} .

1.3 Test conditional independence: Choose an edge between variables X and Y . Test the conditional independence hypothesis $X \perp\!\!\!\perp Y \mid S_{XY}$ for all potential subsets S_{XY} via kernel dependence measures, subject to the potential ancestor condition and auxiliary graph \mathcal{G} . For several potential subsets S_{XY} , the constraint with S_{XY} of small cardinality should be first tested. If the independence is accepted, remove the edge between X and Y and change all directed edges into undirected edges, providing a skeleton \mathcal{G} , then goto step 1.2. Otherwise, repeat step 1.3 for another edge in \mathcal{G} . If all edges are checked, change all directed edges into undirected edges, providing a skeleton \mathcal{G} and continue.

Step 2: Learning orientation in structure by OPB (Fig. 5.4).

2.1 Unshielded colliders: Apply the collider condition of Eq. (5.1) to a unshielded triple $X - Z - Y$. If the collider condition is satisfied for Z , register one vote for $X \rightarrow Z$ and $Z \leftarrow Y$ respectively. Based on the voting results of all possible triples in \mathcal{G} , orient undirected edges into directed or bi-directed edges by a unanimous vote.

2.2 Non-colliders: Orient all substructures $X \rightarrow Z - Y$ (X and Y nonadjacent) into $X \rightarrow Z \rightarrow Y$.

2.3 Acyclicity: Orient all edges $X - Y$ into $X \rightarrow Y$, if a directed path from X to Y exists in \mathcal{G} .

2.4 Shielded colliders: Apply the collider condition to substructure $X - Z - Y$ (X and Y adjacent) or $X \rightarrow Z - Y$ (X and Y adjacent). If the collider condition considers Z as a collider, register one vote for $X \rightarrow Z$ and for $Z \leftarrow Y$. Based on the voting results of all possible triples in \mathcal{G} , orient the remaining undirected edges into directed or bi-directed edges by a unanimous vote.

Output: A mixed graph \mathcal{G} with un-, uni- and bi-directed edges.

Figure 5.11.: Kernel-based causal learning algorithm (KCL).

5. From Magnitude of Dependences to Causal Structure

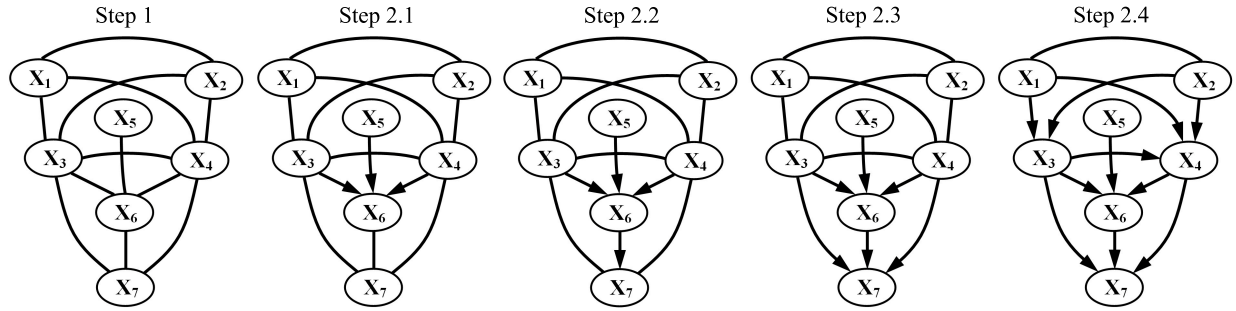


Figure 5.12.: Stepwise results of learning the structure as shown in Fig. 1.1, when KCL is applied.

actually be tested. How to control the potential errors of these tests is essential to the performance of KCL.

A constraint-based approach, in principle, removes an edge between X and Y when some set $S_{XY} \subseteq \mathcal{V} \setminus X \cup Y$ can be found which makes X and Y independent. Thus, an edge is removed when conditional independence is accepted by at least one test. An edge is present in the resulting structure when the conditional independence hypothesis between X and Y are rejected in every test. A test which wrongly yields a dependence (a type I error) between X and Y has no impact on the resulting adjacency structure as long as there is another test yielding the absence, while a type II error does. For this reason, a straightforward implementation automatically tends to remove too many edges, if type II error is not kept to an extremely low level. This phenomenon can be often observed by outputs of PC/FCI. For this reason, to keep the level of type II error made by independence tests as low as possible is essential to the performance of learning adjacency structure.

Our dependence measures benefit from the power of kernel-based approaches and can detect additional dependence in which the data are uncorrelated, yet have some more complex nonlinear dependence that simple correlation does not detect. However, unless the sample sizes are excessively large, the conditional independence tests of two variables conditional on a large set of other variables are in general not reliable. The number of errors of any statistical test increases when the sample is small or the cardinality of the conditioning set is large (see [153], p. 116). The kernel-based independence test suffers from the same problem, i.e., the dependence measure tends to be very small when the cardinality of conditioning set is large.

Instead of limiting the the cardinality of the conditioning set directly, we handle the problem in an implicit and flexible way: if the differences between the original estimator $\widehat{\mathbb{H}}_0$ and the simulated values $\widehat{\mathbb{H}}_1, \dots, \widehat{\mathbb{H}}_{n_p}$ are too small, e.g., smaller than 10^{-8} , the independence hypothesis will be rejected in favor of dependence. This way, we avoid the arbitrariness of setting an upper-bound on the cardinality of the conditioning set when testing conditional independence. When sample size is small or the conditioning set is large, our independence test will be unreliable and dependence will be assumed: lack of support for independence implies dependence. Thus, if our test rejects an independence hypothesis, it does not mean that the data are against independence,

but that there is no evidence in the data for it.

Another problem is that a naive search of every conditioning set S_{XY} that makes X and Y independent is inefficient (too many tests) and inaccurate (conditioning on too many variables). The PC algorithm provides a simple strategy for selecting S_{XY} as follows. The in-degree of the structure (maximum number of direct parents in graph) can be chosen to be bounded from above by some constant and one begins to test independence with conditioning sets of small cardinality. Thus, the PC-style selection is to first take S_{XY} with small cardinality into account and then the subsets with larger cardinality, since testing conditional independence with smaller conditioning set is more reliable. If the underlying model is indeed sparse, PC will be efficient.

In addition to the PC-style selection, using the orientation information of the auxiliary graph learned by OPA, the potential ancestor condition can reduce the search space of S_{XY} . The reduction avoids unnecessary independence tests, which could lead to a type II error.

Another critical issue of learning adjacency structure is the order of testing conditional independence, in particular if conflicting constraints can be obtained. One typical example is the non-intersection conflict (Definition 18). As an example, we consider the digoxin clearance data, which are already discussed in Section 4.2.3. The constraints verified by kernel independence tests in the digoxin clearance data are

$$(X \perp\!\!\!\perp Y \mid Z) \wedge (X \perp\!\!\!\perp Z \mid Y),$$

where X : URINE FLOW, Y : DIGOXIN CLEARANCE, Z : CREATININE CLEARANCE (see Tab. 3.8). The resulting adjacency structure depends on whether $X \perp\!\!\!\perp Y \mid Z$ or $X \perp\!\!\!\perp Z \mid Y$ is first tested. The PC algorithm does not address this problem.

Unlike RCL as shown in Section 4.5, KCL uses the kernel dependence measures to handle the conflicting constraints. In agreement with Assumption 5, we propose to first test the independence of pairs (X_i, X_j) with larger $\hat{\mathbb{H}}_{X_i X_j}$, because we are of the opinion that the screening-off effect induced by conditioning on a set of variables can be more reliably detected, if the magnitude of marginal dependence is strong. The weaker the magnitude of marginal dependences, the less reliable the conditional independence test. In the digoxin clearance data, we have $\hat{\mathbb{H}}_{YZ} > \hat{\mathbb{H}}_{YX} > \hat{\mathbb{H}}_{ZX}$ (see Tab. 3.8). The resulting structure would be

$$\text{URINE FLOW} - \text{CREATININE CLEARANCE} - \text{DIGOXIN CLEARANCE}$$

stating that digoxin clearance is independent of urine flow given creatinine clearance.

5.6. Real-world experiments with KCL

It is clear that the assumptions we made, e.g., λ -collider condition, could be violated in real-world data. Therefore, our intention was not to seek special data that would fit our algorithm, but rather to analyze how well KCL really performs on various kinds of data. In particular, KCL is able to learn structure when no independence relations are present. For the sake of evaluation, we prefer such data and variables where common sense provides some obvious prior information about the causality. Since we intend to compare KCL to the conventional constraint-based PC/FCI,

5. From Magnitude of Dependences to Causal Structure

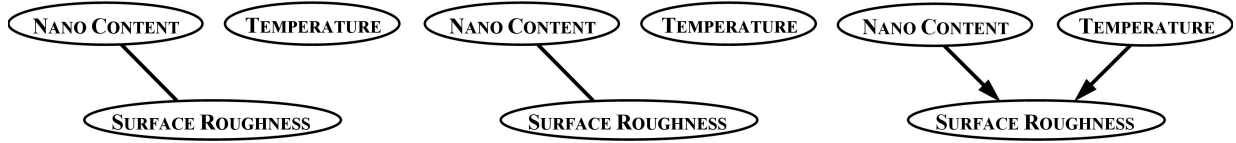


Figure 5.13.: Outputs learned by PC (left), FCI (middle), and KCL (right) from ceramic surface data.

it suggests itself to examine the datasets, which are listed as examples for its software project TETRAD on its webpage. In addition, we perform the BN-PC algorithm, WinMine toolkit for data on purely categorical domains and the LiNGAM algorithm for data on purely continuous domains.⁵

5.6.1. Ceramic surface

The influence on sintered bodies of the variation of nano powder content on sintered bodies was investigated by researchers of Research Center Karlsruhe, Germany. The nano powder was added to a powder mixture in different ratios (0%–70%), from which sintered ceramic parts were fabricated. The ceramic parts were sintered at four different temperature levels: 1300°C, 1350°C, 1400°C and 1450°C. The mixture ratios and temperatures are chosen independently. Using an optical scanning device, the surface roughness of these parts is characterized by roughness average R_a as well as roughness depth R_z^{ISO} and R_z^{DIN} , depending on ISO or DIN standards. R_a , R_z^{ISO} , R_z^{DIN} are defined in DIN EN ISO 4287, DIN 4768 (1990) and DIN 4762 (1989), respectively.

The dataset contains 80 measurements. We know that the NANO CONTENT and sintering TEMPERATURE influence the SURFACE ROUGHNESS of sintered parts and not vice versa. In our experiments, we used different vectors to characterize the SURFACE ROUGHNESS: R_a , R_z^{ISO} , R_z^{DIN} , (R_a, R_z^{ISO}) , (R_a, R_z^{DIN}) , $(R_z^{\text{ISO}}, R_z^{\text{DIN}})$, and $(R_a, R_z^{\text{ISO}}, R_z^{\text{DIN}})$.

In all 7 different vectorial definitions of SURFACE ROUGHNESS, KCL identified SURFACE ROUGHNESS as the common effect. Remark that this is an obvious advantage of KCL against PC, since the former can be extended to multidimensional domains in a straightforward way. The result of PC (Fig. 5.13, left) is less specific and plausible than KCL (Fig. 5.13, right). In the case of PC, we interpreted the SURFACE ROUGHNESS as a one-dimensional variable: R_a , R_z^{ISO} or R_z^{DIN} . All the three constructions yielded the same output (Fig. 5.13, left).

⁵See http://www.phil.cmu.edu/projects/tetrad_examples for datasets. We used default parameters of TETRAD 4.3.8 and set significance level $\alpha = 0.05$. BN-PC [28] is a constraint-based algorithm using mutual information as independence measure and implemented in BNT Structure Learning Package by Leray et al. [101], and online available at <http://banquiseasi.insa-rouen.fr/projects/bnt-slp>. WinMine toolkit [31] is a Bayesian approach using a non-informative prior on the structures and online available at <http://research.microsoft.com/~dmax/winmine/tooldoc.htm>. LiNGAM [143] is a recently developed algorithm for learning structures on continuous domains. We used the version 1.4.2, which is online available at <http://www.cs.helsinki.fi/group/neuroinf/lingam>.

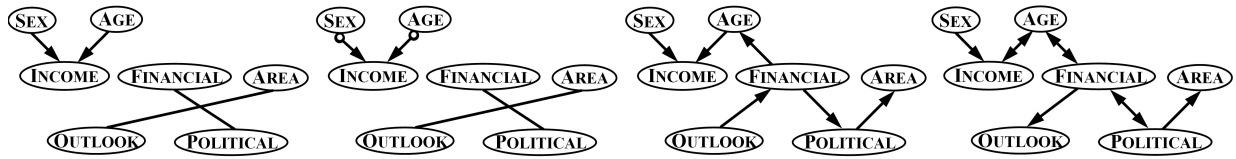


Figure 5.14.: Outputs learned by PC, FCI, BN-PC, and KCL (from left to right) from Montana outlook poll data.

5.6.2. Montana outlook poll

We tested KCL on the data of Montana outlook poll, which are already discussed in Section 4.6.3. All constraint-based algorithms detected a more or less similar adjacency structure (Fig. 5.14), since we conjecture that various independence tests achieves consistent results in purely categorical domains. Both PC/FCI (Fig. 5.14, first and second plot from left) and KCL (Fig. 5.14, rightmost) provided no obviously wrong arrow. Variables SEX, AGE, and AREA are in fact not identified as effects of any other variable, which is in agreement with our prior knowledge. The result of KCL is more plausible, in the sense that PC/FCI erroneously excluded the relation between AGE and FINANCIAL. It is noteworthy that the output of KCL some has bi-directed edges, which indicate conflicting orientation information.

Since this dataset contains only categorical variables, it is justified to run BN-PC algorithm. The result (Fig. 5.14, third plot from left) has the same corresponding skeleton as that of KCL (Fig. 5.14, rightmost). The arrow from AGE to INCOME is correctly detected by BN-PC, whereas the output of KCL is less specific in the orientation of this edge. It contains, however, obviously wrong arrows from FINANCIAL to AGE. The causal direction from OUTLOOK to FINANCIAL also seems to be less plausible. We do not speculate on the causal direction between POLITICAL and AREA. In addition, we ran score-based Bayesian algorithms with greedy search or MCMC and WinMine toolkit on this data. They returned as output, unfortunately, a trivial graph without any edges and found no structure in the data.

5.6.3. Egyptian skulls

We perform PC/FCI and KCL on the Egyptian skulls data, which are already discussed in Section 4.6.2. The output of KCL (Fig. 5.15, rightmost) is consistent with the output of RCL (Fig. 4.14). In comparison to the output of PC/FCI, the output of KCL has one edge more and is completely directed. This experiment confirms our observation that KCL/RCL tends to draw more edges than PC/FCI.

5.6.4. Cheese data

As cheese ages, various chemical processes take place that determine the taste of the final product. In a study of cheddar cheese from the LaTrobe Valley of Victoria, Australia [13] samples of cheese were analyzed for their chemical composition and were subjected to taste tests. The

5. From Magnitude of Dependences to Causal Structure

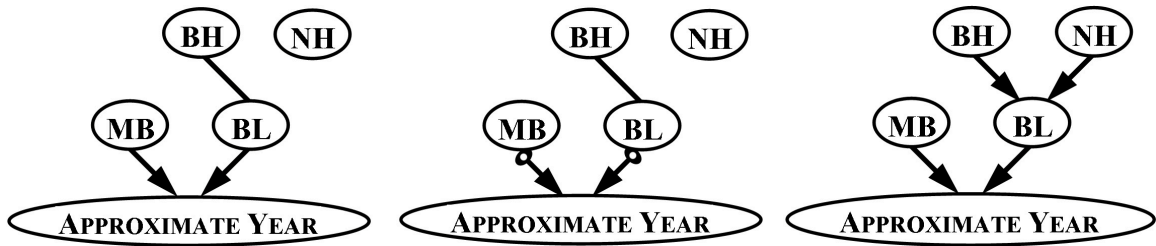


Figure 5.15.: Outputs learned by PC (left), FCI (middle), and KCL (right) from Egyptian skulls data.

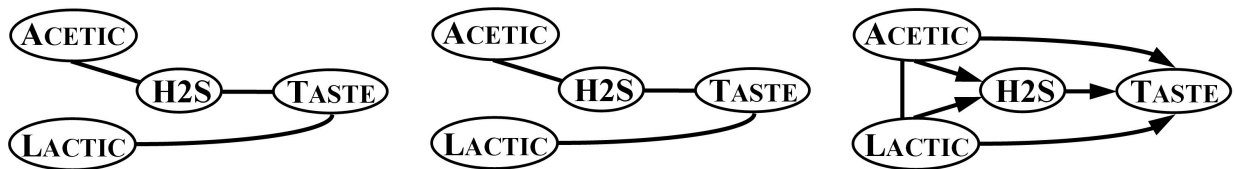


Figure 5.16.: Outputs learned by PC (left), FCI (middle), and KCL (right) from cheese data. Note that the undirected edges in output of PC and FCI have slightly different meanings, although we chose the same representation.

dataset contains concentrations of various chemicals in 30 samples of mature cheddar cheese and a subjective measure of taste for each sample. Overall TASTE scores were obtained by combining the scores from several tasters. The variables ACETIC, H2S, and LACTIC represent the concentrations of acetic acid, hydrogen sulfide, and lactic acid, respectively.

The causal graphs obtained by PC, FCI, and KCL are shown in Fig. 5.16. The output of KCL is the most specific one. The detected causal knowledge that TASTE is only an effect and not a cause of any other variable is in agreement with the ground truth. Note that, although the kernel-based approach detects no independence, KCL is able to offer some causal information. Due to our lack of chemical understanding, we do not speculate on the plausibility of the influences among the various chemicals detected by KCL, i.e., from ACETIC and LACTIC to H2S. The edge between ACETIC and LACTIC cannot be oriented. This example shows that, in spite of the fully connected skeleton, KCL can learn orientation in the structure by means of the magnitude of dependences, whereas RCL cannot learn anything from the independence relations.

Since all domains in this dataset are real-valued, LiNGAM algorithm could be applicable. LiNGAM converged with no-error-report and returned a graph without any edges as output. Hence, LiNGAM provided no information about the causal structure. We conjecture that this might be due to the fact that the statistical Wald tests [172] for pruning edges in LiNGAM may not be suitable for such a small sample size.

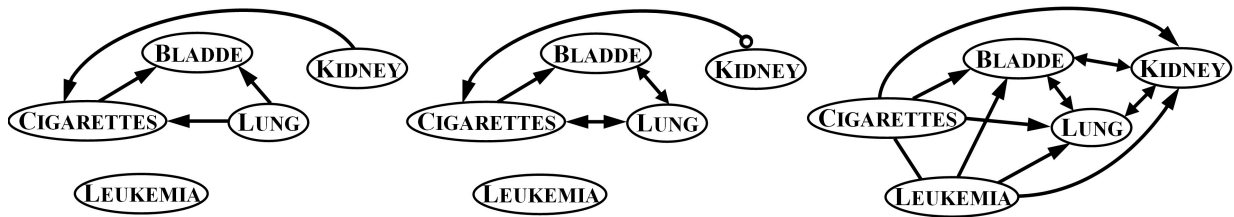


Figure 5.17.: Outputs learned by PC (left), FCI (middle), and KCL (right) from smoking and cancer data.

5.6.5. Smoking and cancer

The smoking and cancer data [58] contain the numbers of CIGARETTES (hundreds per capita) sold in 43 states in the US and the District of Columbia in 1960, together with death rates per hundred thousand population from various forms of cancer, i.e., BLADDER cancer, LUNG cancer, KIDNEY cancer, and LEUKEMIA. The fact that Nevada and the District of Columbia are outliers in the distribution of cigarette consumption contributes to the difficulty of the analysis. The ready explanation for the outliers is that cigarette sales are increased by tourism (Nevada) and commuting workers (District of Columbia).

It is generally accepted that the consumption of cigarettes is a cause of various forms of cancer, not vice versa. As seen from the right plot in Fig. 5.17, KCL discovers CIGARETTES as the common cause of BLADDER, LUNG, KIDNEY, and LEUKEMIA, which confirms the common-sense understanding of the causal influences. The causal direction between CIGARETTES and LEUKEMIA in the output of KCL remains indeterminate. Due to our lack of medical understanding, we do not speculate on the plausibility of the orientation from LEUKEMIA to other forms of cancer. Interestingly, KCL contains some bi-directed edges between BLADDER, LUNG, and KIDNEY, which might be due to some common hidden causes of BLADDER, LUNG, and KIDNEY. Obviously, KCL detects more dependences among observed variables and provides thus a considerably more complex structure than PC/FCI. The outputs of PC/FCI (Fig. 5.17, left and middle) contain significantly fewer edges and is less specific. In particular, the orientations from LUNG (by PC) and KIDNEY (by PC/FCI) to CIGARETTES are obviously wrong.

Although the output of KCL is fully connected, one independence relation is accepted by the kernel test:

$$\text{CIGARETTES} \perp\!\!\!\perp \text{LEUKEMIA} \mid \text{BLADDER, LUNG, KIDNEY},$$

if we conducted tests for all possible non-trivial independence constraints. The output of RCL will be a graph with overall undirected edges excepting the connection between CIGARETTES and LEUKEMIA. Due to the auxiliary graph learned by the orientation heuristics and the potential ancestor condition, this independence constraint was not tested and thus not considered by KCL. Consequently, the edge between CIGARETTES and LEUKEMIA will not be removed in the final output of KCL.

Since this dataset contains only purely continuous variables, we ran the LiNGAM algorithm. Like the cheese data in Section 5.6.4, LiNGAM converged with no errors. But, the output of

5. From Magnitude of Dependences to Causal Structure

LiNGAM, a completely unconnected graph, gave again no hints about the causal structure.

5.6.6. Brain size and IQ

Monozygotic (MZ) twins share numerous physical, psychological, and pathological traits. In vivo brain image acquisition and analysis make it possible to determine quantitatively whether twins share neuroanatomical traits and whether neuroanatomical measures correlate with brain size. Using magnetic resonance imaging and computer-based image analysis techniques, measurements of the BRAIN VOLUME (cm^3), CORPUS COLLASUM surface area (cm^2), CORTICAL SURFACE area (cm^2) were obtained in 10 pairs of MZ twins. HEAD CIRCUMFERENCE (cm), body WEIGHT (kg), and full-scale IQ (intelligence quotient) were also measured. Tramo et al. [165] used GENOTYPE (Pair Identifier), BIRTH ORDER, and SEX (1: Male, 2: Female) as between-subject factors to examine neuroanatomic similarities in MZ twins and their relationship to head size and IQ.

If we applied the constraint-based approaches to Brain Size and IQ data, the result of PC/FCI (Fig. 5.18, left and middle) indicates merely some correlation between BRAIN VOLUME, CORPUS COLLASUM and CORTICAL SURFACE, as well as the relation between GENOTYPE and SEX. As discussed in Section 5.1, it is difficult to detect significant dependences by statistical tests in the relatively large network from a sample with 20 data points. Therefore, it is helpful to amplify the original data and run the so-called resampling-based multiple test to balance the errors of hypothesis tests. More precisely, we resampled (with replacement) 100 subsamples of 200 data points and conducted the usual kernel test as described in Fig. 3.2 for each subsample. Then, we obtained a set of 100 p-values (one for each subsample) for each independence hypothesis. Instead of the sophisticated procedure as described in Fig. C.1 in Appendix C, we used here a simplified version of multiple testing via the median of the set of 100 p-values due to computational feasibility. If the median of the set of 100 p-values larger than $\alpha = 0.05$, the independence hypothesis will be accepted, otherwise rejected. By means of this simplified resampling-based multiple testing, we ran KCL on this dataset.

The result of KCL (Fig. 5.18, right) reveals the genetic influences on the size and shape of the human forebrain and its gross morphologic subdivisions. The fact that BIRTH ORDER and SEX is not detected as effect of these neuroanatomic measures is also consistent with our prior knowledge. Due to our lack of medical understanding, we do not speculate on the plausibility of the causal interpretation of the arrow from IQ to body WEIGHT and the three bi-directed edges, which indicate hidden common causes.

In this experiment, the relations among BRAIN VOLUME, CORPUS COLLASUM, CORTICAL SURFACE are less interesting than the effect of various factors on them. As expected, the result in [165] indicated a strong similarity of BRAIN VOLUME, CORPUS COLLASUM, and CORTICAL SURFACE in MZ twins. These brain measures are tightly correlated with one another and with HEAD CIRCUMFERENCE. In order to make the resulting structure more task-oriented, i.e., representing causal relationships between more interesting factors, we cluster the set of measured variables into groups of variables by prior knowledge, i.e., the meaning of variables, and learn the structure among these vectorial variables. Each group of variables (called latent factors, see Section 4.3 for more discussions) is represented by a single node in the final output. We performed

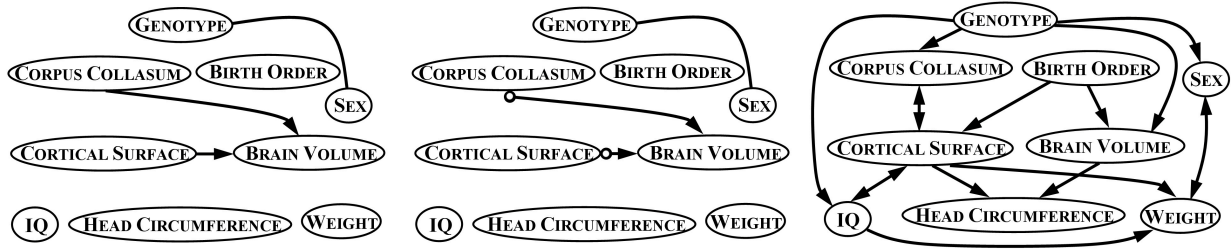


Figure 5.18.: Outputs learned by PC (left), FCI (middle), and KCL (right) from brain size and IQ data.

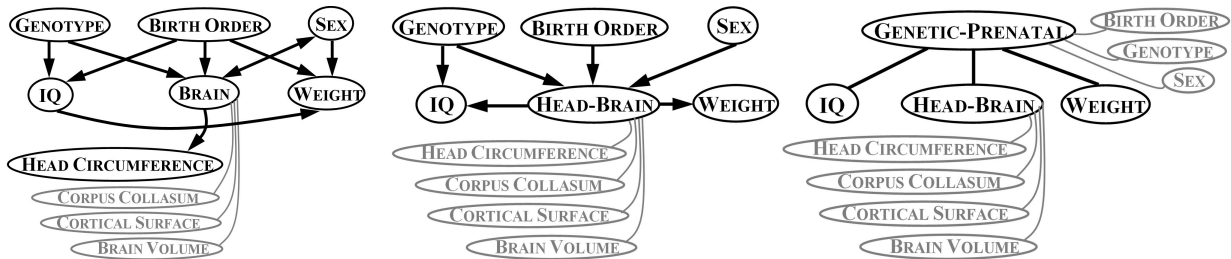


Figure 5.19.: Outputs learned by KCL from brain size and IQ data with different clusterings of variables.

KCL with three different clusterings of variables:

$$\begin{aligned}
 \text{BRAIN} & := (\text{BRAIN VOLUME}, \text{CORPUS COLLASUM}, \text{CORTICAL SURFACE}) \\
 \text{HEAD-BRAIN} & := (\text{HEAD CIRCUMFERENCE}, \text{BRAIN VOLUME}, \text{CORPUS COLLASUM}, \text{CORTICAL SURFACE}) \\
 \text{GENETIC-PRENATAL} & := (\text{GENOTYPE}, \text{BIRTH ORDER}, \text{SEX})
 \end{aligned}$$

The results based on the first two clusterings (Fig. 5.19, left and middle) are positive in the sense that GENOTYPE, BIRTH ORDER, and SEX are identified as causes. The output of the last clustering (Fig. 5.19, rightmost) is an undirected graph. The undirected structure excludes collider structures on GENETIC-PRENATAL, the output indicates the plausible fact that conditioning on GENETIC-PRENATAL makes every pair of the three measurements HEAD-BRAIN, IQ, HEAD-BRAIN independent and at least two of these three measurements are direct effects of GENETIC-PRENATAL. The obviously false hypothesis that GENETIC-PRENATAL could be a common effect of any pair of them is correctly excluded. This example shows that it seems that an appropriate clustering of variables, i.e., a meaningful construction of nodes in the output, is essential for discovering useful structures representing data. Note that PC/FCI and LiNGAM cannot treat the multi-dimensional variables at all.

5. From Magnitude of Dependences to Causal Structure

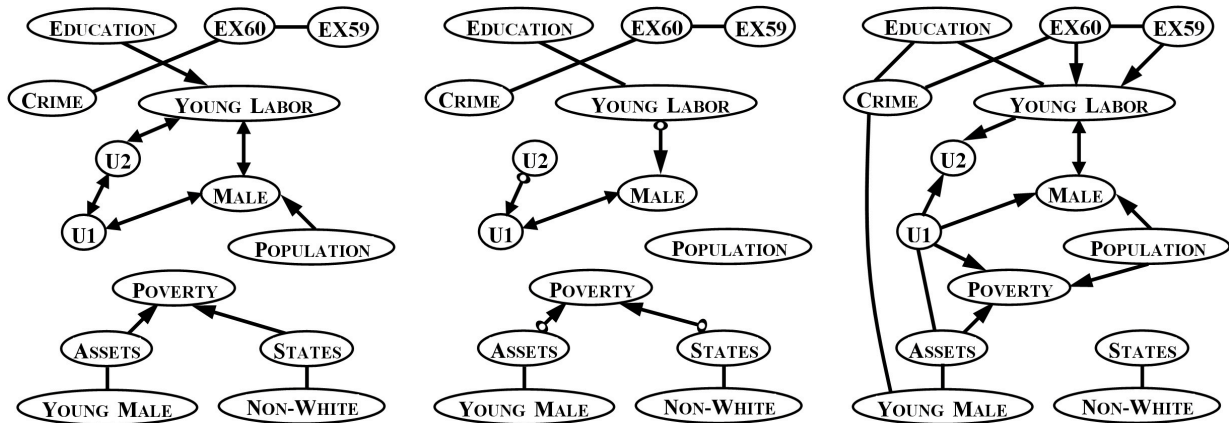


Figure 5.20.: Outputs learned by PC (left), FCI (middle), and KCL (right) from US crime data.

5.6.7. US crime data

The US crime data [167] are crime-related and demographic statistics for 47 US states in 1960. The data were collected from the FBI's Uniform Crime Report and other government agencies to determine how the variable crime rate depends on the other variables measured in the study. The dataset consists of 14 variables: CRIME rate: the number of offenses reported to police; YOUNG MALE: the number of males of age 14–24; STATES: binary indicator variable for Southern states; EDUCATION: the number of years of schooling for persons of age 25 or older; EX60: 1960 per capita expenditure on police by state and local government; EX59: 1959 per capita expenditure on police by state and local government; YOUNG LABOR: Labor force participation rate per 1000 civilian urban males age 14–24; MALE: the number of males per 1000 females; POPULATION: State population size; NON-WHITE: the number of non-whites; U1: unemployment rate of urban males of age 14–24; U2: unemployment rate of urban males of age 35–39; ASSETS: value of transferable goods and assets or family income; POVERTY: the number of families earning below 1/2 the median income.

It is remarkable that the output of PC (Fig. 5.20, left) contains 4 bi-directed edges, which are traced back to the conflicting conditional independence information. PC may fail partially due to failure of assumptions (e.g., relationships are nonlinear, the true model is cyclic, etc.) or because the sample is not large enough and some statistical decisions are inconsistent. If the resulting adjacency structure after independence tests is correct, bi-directed edges in the output of PC could also be due to latent common causes. Since PC excludes hidden common causes, it is probably better to consider the result of FCI more justified. However, if we wish to find out the causal relationships between crime rate and other factors, the results of PC/FCI (Fig. 5.20, left and middle) are both unsatisfactory, although they provide some plausible connections between the expenditure on police and crime rate, some relationships among demographic statistics. The result of KCL (Fig. 5.20, right) is more complex and does not really provide more evaluable causal information about crime rate as well, since it contains many undirected edges.

Regarding the meaning of variables, it is obvious that some variables must be strongly re-

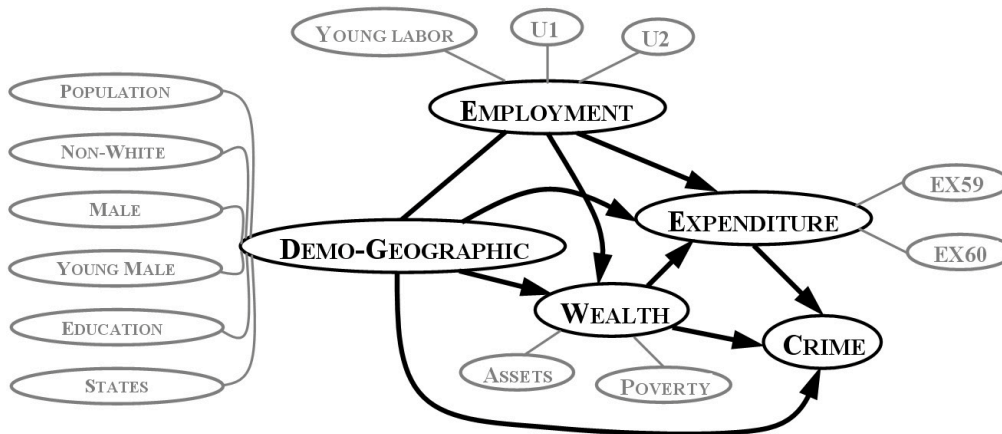


Figure 5.21.: Output learned by KCL from US crime data, given prior knowledge about the variable clustering.

lated. In order to better understand the phenomenon of crime rate, we propose to introduce an appropriate clustering of variables. We reconstruct a demographic and geographic factor, called DEMO-GEOGRAPHIC (comprising POPULATION, NON-WHITE, MALE, YOUNG MALE, and STATES), a factor called EXPENDITURE (containing EX59 and EX60), a factor called EMPLOYMENT (containing YOUNG LABOR, U1, and U2), and a factor WEALTH (containing ASSETS and POVERTY). The variable CRIME remains unchanged. The output of KCL, in which each node corresponds to a factor, is shown in Fig. 5.21. The variable CRIME is reasonably detected as the effect of distinct factors and factor DEMO-GEOGRAPHIC is not an effect of any other factors. The result suggests to consider factor EMPLOYMENT as a cause of WEALTH and EXPENDITURE and factor WEALTH as a cause of EXPENDITURE, which seems to be plausible.

5.6.8. US economy data

One of the interesting fields for our method is learning causal relationships from economic data. We conducted KCL on US economy data from January 1959 to June 2005 (Federal Reserve System <http://www.economagic.com/>). The dataset of size 559 collects money supply M1, money supply M2, REAL INCOME (disposable personal income), INDUSTRIAL PRODUCTION, UNEMPLOYMENT RATE, OIL PRICE, 90-day treasury bills (90B), 90-day commercial paper interest rates (90P), spread (difference of 90B and 90P) by the month. We group money supply M1 and money supply M2 to a 2-dimensional factor MONEY SUPPLY. The variable INTEREST RATES consists of 90B, 90P and spread. Fig. 5.22 illustrates the output of KCL based on the clustering of variables.

Due to the complexity of the US economy, we do not speculate on the correctness of the model found by KCL, since this goes beyond the scope of this thesis. Nevertheless, it seems to be plausible that the real economic activity (MONEY SUPPLY, REAL INCOME, INDUSTRIAL PRODUCTION) is considered as a cause of the indicator on the labor market (UNEMPLOYMENT

5. From Magnitude of Dependences to Causal Structure

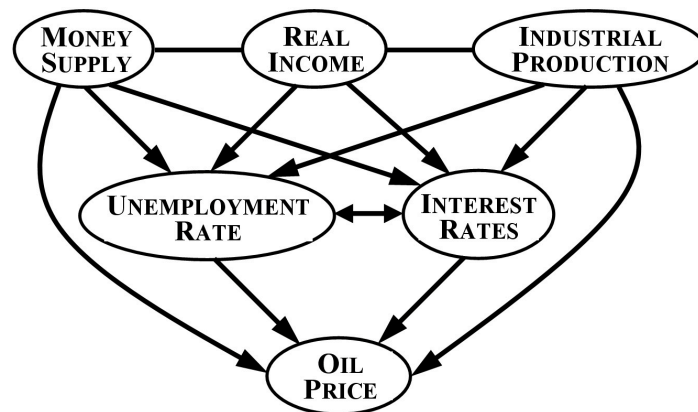


Figure 5.22.: Output learned by KCL from US economy data, given prior knowledge about the variable clustering.

RATE) and the indicator on the financial market (INTEREST RATES). The indicators of real economic activity influence the price level of commodity market (OIL PRICE). The two undirected edges exclude the unshielded collider on REAL INCOME, which represents the fact that INDUSTRIAL PRODUCTION and MONEY SUPPLY is conditionally independent, given REAL INCOME. It should be mentioned that the data are actually given in the form of time series and the result of KCL did not take into account the temporal aspect of the measurement at all.

6. Discovering Causal Order by Properties of Conditionals

We have so far showed that independence relations between variables are helpful for causal inference. If no (conditional) independence relations are present, the magnitude of dependences can be used to infer causal structure. However, both methods need at least three variables and are not capable of giving preference to either of the possible causal hypotheses, if only two dependent variables are measured. Thus, an additional inference rule, which is able to supply some evidence of the statistical asymmetry between cause and effect, would be desirable.

6.1. Motivational example

Imagine the situation that only two dependent variables X and Y . The approaches which are based on independence relations or dependence measures can prefer neither $X \rightarrow Y$ nor $X \leftarrow Y$ (assuming no confounders). Our intention is to take a close look at the dependence, which is described by the Markov kernels (introduced in Section 1.1), i.e., $\{P(X), P(Y|X)\}$ and $\{P(Y), P(X|Y)\}$, with respect to different causal directions. Having assumed some plausible properties of Markov kernels of a natural causal relationship in the real world, one can indeed find some evidence of the underlying causal direction.

To motivate our idea we consider a generating model $X \rightarrow Y$, where cause X is binary and effect Y is real-valued. A convenient assumption about the conditional distribution of the effect given the cause is that it follows a Gaussian distribution $\mathcal{N}(\mu_X, \sigma^2)$ with the same variance but different expectations for each of both values of X (Fig. 6.1, left). The marginal distribution of the effect Y could then be bimodal as shown in the right plot of Fig. 6.1.

A ready explanation of the relation between X and Y is that X shifts the expectation of Y and labels different classes of Y . Having the causal hypothesis $X \rightarrow Y$ in mind, the bimodality of $P(Y)$ has a very natural explanation. In other words, presenting the joint distribution of (X, Y) in terms of conditionals $\{P(X), P(Y|X)\}$ require less information (only the first and second moments of X and Y), whereas using $\{P(Y), P(X|Y)\}$ more than the first two moments of Y are required. Actually this idea is also a motivation for studying Gaussian mixture models: if the distribution of variable Y can be decomposed into a mixture of some Gaussian-distributed variables, it is likely that the decomposition corresponds to different ensembles that stem from different populations.

6. Discovering Causal Order by Properties of Conditionals

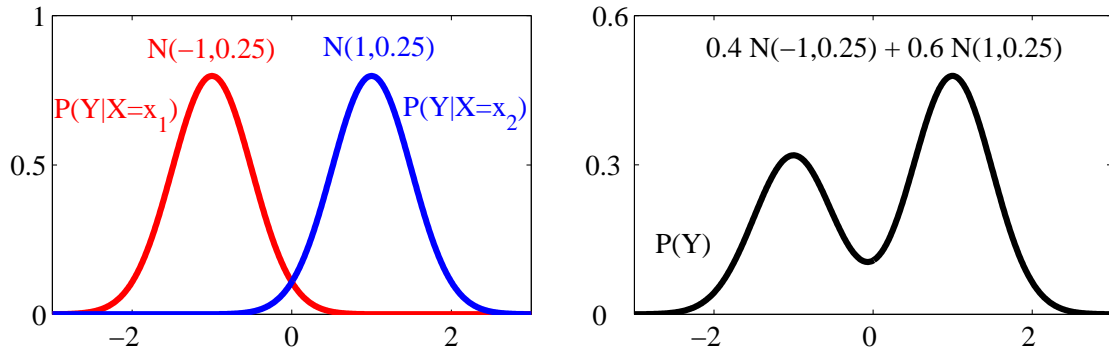


Figure 6.1.: An intuitive example for inferring causal direction via properties of conditional distributions. The generating model is that a binary X effects a real-valued Y by shifting its expectation. The conditional probabilities $P(Y|X)$ is Gaussian distributed (left plot). The corresponding marginal distribution $P(Y)$ is bi-modal (right plot).

6.2. Plausible Markov kernel assumption

The motivational example in the last section showed that a dependence between two variables could be interpreted by different models. If one of the model is able to represent the data with simpler conditionals, one tends to intuitively consider this model as the underlying causal structure. To further explain why we expect that the shape of conditionals is more likely to be simple with respect to the true causal structure, we start with a thought experiment.

Imagine a classical system whose time evolution is determined by a Markov chain (first order Markov process) in discrete time $t \in \mathbb{Z}$. Suppose X_t is the set of variables describing the system configuration at time t and we assume that the variables at time t directly influence only the variables at time $t+1$, i.e., we exclude instantaneous influence among variables within the same time step.

Now, we restrict our attention to one step in system change between two time steps t and $t+1$ and rewrite the sets of N variables X_t and X_{t+1} by $C := (C_1, \dots, C_N)$ and $E := (E_1, \dots, E_N)$, respectively. Fig. 6.2 illustrates a graphical representation evolving C and E over two time slices. The undirected edges indicates spurious dependences among C_i , which are generated by X_{t-1} . Since the time order coincides with the causal order, the asymmetry between past and future necessarily corresponds to the asymmetry between cause and effect. Hence, the arrows from C_i (“Cause”) to E_j (“Effect”), as shown in Fig. 6.2, could be interpreted causally.

Given the causal structure as shown in Fig. 6.2, the Markov condition implies that, given all direct causes $\{C_1, \dots, C_N\}$, effects $\{E_1, \dots, E_N\}$ become stochastically independent, i.e.,

$$P(E_1, \dots, E_N | C_1, \dots, C_N) = \prod_{j=1}^N P(E_j | C_1, \dots, C_N). \quad (6.1)$$

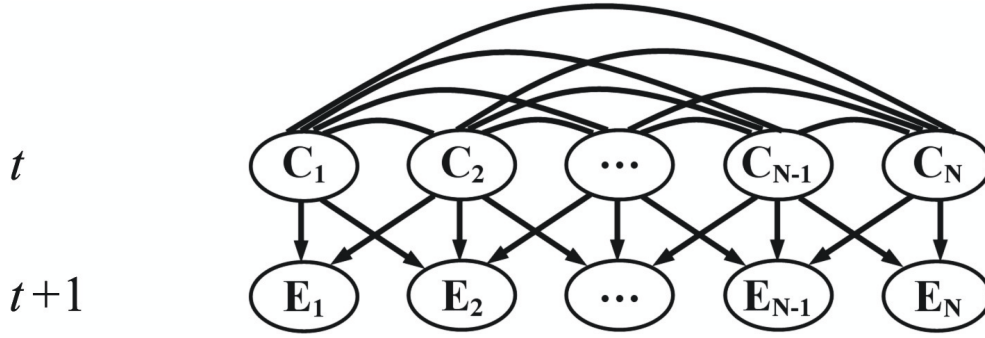


Figure 6.2.: Two time layers of a first order Markov stochastic process. The first layer represents the configuration of relevant variables at time t and the second layer the configuration of them at time $t+1$. The value of every variable at time t influences the values of itself and its neighbors at time $t+1$. The undirected edges representing spurious dependences induced by influences at time $t-1$.

It is easy to see that the conditionals of backward time do not follow the analogue statement, i.e.,

$$P(C_1, \dots, C_N | E_1, \dots, E_N) \neq \prod_{j=1}^N P(C_j | E_1, \dots, E_N), \quad (6.2)$$

otherwise the faithfulness assumption would be violated. Using the d-separation criterion, it is apparent that conditioning on any subset of $\{E_1, \dots, E_N\}$ cannot, in a faithful Bayesian network, make any subsets of $\{C_1, \dots, C_N\}$ mutually independent. The present dependences can only be canceled out by accident. In other words, if one wishes to reverse a causal link, then the total number of links can not decrease.

This difference between forward and backward time can be understood as an asymmetry of simplicity in parameters of causal models. The hypothetical causal model of forward time, is featured by a factorization into natural conditionals, which is not possible for the hypothetical causal model of backward time. The next question is how to formalize such an intuitive simplicity concept.

Suppose the number of direct parents in the underlying causal model is bounded from above by some finite number k , that means every effect E_j is influenced by at most k causes. Fig. 6.2 shows a situation of $k=3$. Then the conditional probability, which is consistent with the causal direction $C \rightarrow E$, can be written in an exponential form as

$$P(E | C) = P(E_1, \dots, E_N | C_1, \dots, C_N) = \exp \left(\sum_j^N f_j(E_j, C_{j_1}, \dots, C_{j_k}) \right),$$

where C_{j_1}, \dots, C_{j_k} are k direct parents of E_j in graph with $j_i \in \{1, \dots, N\}$. Function f_j depends on at most $k+1$ variables, i.e., E_j and $\{C_{j_1}, \dots, C_{j_k}\}$. The other conditional probability, which

6. Discovering Causal Order by Properties of Conditionals

is opposite to the causal direction, can be calculated as:

$$P(C | E) = P(C_1, \dots, C_N | E_1, \dots, E_N) = \exp \left(\sum_j^N f'_j(C_j, E_{j_1}, \dots, E_{j_{k'}}) \right),$$

where $E_{j_1}, \dots, E_{j_{k'}}$ with $j_i \in \{1, \dots, N\}$ are variables that function f'_j depends on. If the generating model is faithful, we expect $k' \geq k$. In the structure as shown in Fig. 6.2, we have even $k \ll k' = N$.

Summing up, $P(E|C)$ can be represented by functions of lower order (smaller number of input variables) than $P(C|E)$. A function with lower order is smoother. This idea will be formalized by the so-called plausible Markov kernel assumption.

Assumption 6 *Let π_1, π_2 be two distinct orders on the set of variables $\mathcal{V} := \{X_1, \dots, X_N\}$. Mk_1 and Mk_2 denote the corresponding set of Markov kernels (as introduced in Section 1.1) with respect to π_1 and π_2 . If π_1 is consistent with the ancestral ordering entailed by the underlying causal structure on \mathcal{V} (called causal order), whereas π_2 is inconsistent, then Mk_1 is more plausible than Mk_2 , in the sense that the functions in Mk_1 are smoother than those in Mk_2 .*

In other words, all Markov kernels in Mk_1 describe cause-and-effect relationship and represent the “physics” of a natural causal mechanism, whereas Markov kernels in Mk_2 are mixtures of cause-and-effect relations and prior probabilities of causes. We expect that the functions in Mk_2 is less smooth than those in Mk_1 . A first attempt is made to justify this assumption from a thermodynamic viewpoint by D. Janzing and A. Allahverdyan [93, 4]. A related framework to capture asymmetry of relationships between variables by means of Bayesian networks is presented by Comley et al. [37].

The question now is how to evaluate the plausibility of Markov kernels. In practice, there is a quite common agreement that the shape of some well-known densities, e.g., Gaussian or gamma distributions, is rather smooth. In contrast, a mixture of two Gaussians, in particular when it is obviously bimodal, is considered as less smooth. Section 6.1 showed that common sense gives us in some situations an intuitive idea about which distributions would be considered natural and which one might demand an additional explanation as being a mixture of “more natural” and “smoother” distributions. Nevertheless, quantifying and comparing smoothness of probabilities from finite data is a non-trivial problem.

6.3. Plausible Markov kernels via low-order interactions

We propose to use the constrained entropy maximization subject to statistical moments of low-order for evaluating the smoothness of conditional probabilities from finite data without estimating the density directly. More precisely, given a hypothetical causal order $\pi := (X_1, \dots, X_N)$, we define the set of smoothest Markov kernels $Mk_\pi := \{P(X_1), \dots, P(X_N | X_1, \dots, X_{N-1})\}$ via entropy maximization subject to the first and second moments.

The idea is the following. Given statistical moments, one is required to pick one distribution from the set of distributions satisfying the given moments. A natural choice is to pick the

distribution with maximum entropy, which corresponds to smooth distributions. If one restricts the constraints for entropy maximization to only very few “simple” functions, e.g., the first and second moments, the maximum entropy method aims to take the simplest or smoothest distribution, which contains no unwanted non-smooth and complex structure. We refer to Collins et al. [36] and Dowson et al. [48], who described a mathematical framework of the maximum Shannon entropy approach to assign a probability distribution on the basis of a limited number of moments.

6.3.1. Smoothest Markov kernel of cause

Suppose cause X is a vectorial variable $X := (X^{(1)}, \dots, X^{(n)})$ with possible values $x \in \mathcal{X} \subseteq \mathbb{R}^n$. The smoothest Markov kernel P_X is the following (joint) distribution that maximizes the entropy function $\mathcal{H}(X)$ of X subject to the given first and second moments.

$$\begin{aligned}
 & \text{maximize}_{P_X} & \mathcal{H}(X) & := - \sum_x P(x) \ln(P(x)) & & \text{(Entropy of } P_X) \\
 & \text{subject to} & P(x) & \geq 0 \quad \forall x \in \mathcal{X} & & \text{(Non-negativity)} \\
 & & \sum_x P(x) & = \mathbf{1} & & \text{(Normalization)} \\
 & & \sum_x x P(x) & = \mu & & \text{(1st moment)} \\
 & & \sum_x x^{(i)} x^{(j)} P(x) & = \alpha_{ij} \quad \forall i, j = 1, \dots, n & & \text{(2nd moment)}
 \end{aligned}$$

Here μ denotes the vector of empirical mean and $\alpha \equiv (\alpha_{ij})$ the empirical covariance matrix of X . Note that the entropy maximization can also be rewritten as a maximum likelihood estimation within an exponential family of distributions having polynomials of degree two in the exponent. This is basically the well-known convex duality between entropy maximization and maximum likelihood (see e.g. [6]).

6.3.2. Smoothest Markov kernel of effect given single cause

To determine the smoothest Markov kernel $P_{Y|X}$ of effect $Y := (Y^{(1)}, \dots, Y^{(m)})$ with its single cause $X := (X^{(1)}, \dots, X^{(n)})$, we maximize the entropy of the conditional distribution of Y given X constrained by the mean vector of Y , the within-block covariance of Y itself and the cross-covariance of Y with X . Hence, the smoothest Markov kernel of Y is the solution of the following optimization.

$$\begin{aligned}
 & \text{maximize}_{P_{Y|X}} & \mathcal{H}(Y|X) & := - \sum_x \sum_y P(x) P(y|x) \ln(P(y|x)) & & \text{(Entropy of } P_{Y|X}) \\
 & \text{subject to} & P(y|x) & \geq 0 \quad \forall (x, y) \in \mathcal{X} \times \mathcal{Y} & & \text{(Non-negativity)} \\
 & & \sum_y P(y|x) & = \mathbf{1} \quad \forall x \in \mathcal{X} & & \text{(Normalization)} \\
 & & \sum_x \sum_y y P(x) P(y|x) & = \mu & & \text{(1st moment)} \\
 & & \sum_x \sum_y y^{(i)} y^{(j)} P(x) P(y|x) & = \alpha_{ij} & & \text{(2nd moment)} \\
 & & \sum_x \sum_y y^{(k)} x^{(l)} P(x) P(y|x) & = \beta_{kl} & & \text{(2nd mixed moment)} \\
 & & \forall i, j, k = 1, \dots, m \quad \text{and} \quad l = 1, \dots, n
 \end{aligned}$$

In this context, the value of cause X and its distribution P_X is given. $\mu \in \mathbb{R}^m$ denotes the empirical mean of Y , $\alpha \equiv (\alpha_{ij}) \in \mathbb{R}^{m \times m}$ the empirical within-block covariance of Y and $\beta \equiv (\beta_{kl}) \in \mathbb{R}^{m \times n}$ the empirical cross-covariance of X and Y .

6. Discovering Causal Order by Properties of Conditionals

Actually, we need not take account of the non-negativity condition in the optimization explicitly, because the logarithms in the objective function imply this condition. In the continuous limit, one could take the limit of the discrete optima. It should be mentioned that, given the observed first and second moments, the optimization for $P_{Y|X}$ in general is not necessarily feasible. An example will be demonstrated in Section 6.4.2.

6.3.3. Smoothest Markov kernel of effect given multiple causes

Our definition of the smoothest Markov kernels can be straightforwardly generalized to multiple causes $\{X_1, \dots, X_{j-1}\}$ by treating them formally as one variable $AN_j = (X_1, \dots, X_{j-1})$ in the optimization, although they appear in the hypothetical causal structure as separate nodes.

Joint distribution $P(AN_j)$ of all causes X_i on set \mathcal{X}_i ($i = 1, \dots, j$) is given, e.g., it can be iteratively calculated by the optimizations described in Section 6.3.1 and Section 6.3.2. Further, μ_i (expectation of X_i) and β_{ij} (expectation of $X_i X_j$) are known. Then the smoothest Markov kernel is the conditional probability measure

$$P(X_j|X_1, \dots, X_{j-1}) = P(X_j|AN_j)$$

that maximizes the conditional entropy

$$\mathcal{H}(X_j|X_1, \dots, X_{j-1})$$

subject to the constraints

$$E[X_i] = \mu_i \quad \text{and} \quad E[X_i X_j] = \beta_{ij} \quad \forall i \leq j.$$

It can be shown that the optimization leads to a distribution of the form

$$P(x_j|an_j) = \exp\left(\gamma(an_j) + \theta_0 x_j + x_j \sum_{i=1}^j \theta_i x_i\right) \quad (6.3)$$

with appropriate Lagrange multipliers $\gamma(an_j)$ and θ_i . If $j = 1$, AN_1 is then empty (see e.g., [48] for unconditional distributions and [9] for conditional distributions). We assign $P(x_1|an_1) = P(x_1)$ and obtain

$$P(x_1) = \exp(\gamma + \theta_0 x_1 + \theta_1 x_1^2).$$

Due to the existence of awkward normalization constants $\gamma(an_j)$, namely one for each $j-1$ -tuple an_j , it is typically difficult or impossible to obtain all Lagrange multipliers analytically, if the value set of AN_j becomes very large or even infinite. Fortunately, the optimization is strictly convex [25], which ensures a unique solution (if solvable) and numerical feasibility for a finite domain and computational efficiency if the cardinality of the domain is not too large. Although variables in general might be either continuous or discrete, for the sake of simplicity, we henceforth assume that all domains are discrete and finite. For a continuous domain, the only change required under this assumption is a suitable discretization with a sufficiently small

scale. The resulting discrete values are called supports of the continuous domain. A visualization of the probability measure divided into small enough intervals gives a good intuition of the shape of density on a continuous domain. Note that nominal-categorical variables of d nominal alternatives can be treated as shown in Eq. (3.1).

In summary, given an ancestral order $\pi := (X_1, \dots, X_N)$ as well as the first and second moments of variables, the set of the smoothest Markov kernels

$$Mk_\pi := \{P(X_1), P(X_2|X_1), \dots, P(X_N|X_1, \dots, X_{N-1})\}$$

can be, in turn, calculated according to the constrained optimization problems above.

6.4. Examples of smoothest Markov kernels

To give some intuition of the smoothest Markov kernels we present some examples. In some special cases one can solve the optimizations even in a closed form.

6.4.1. Numerical solution on continuous domain

It is well known that the solution of Eq. (6.3) for continuous variables on an unbounded real-valued range leads to a Gaussian conditional. But, when there is certain restriction on the possible value of variables, it is not trivial to see which properties the smoothest Markov kernels would have. Despite of that, we numerically compute the solution on bounded continuous range to give a bit more intuition about our notion of smoothness.

We divided continuous domains into small enough intervals of equal width. Supposing that X and Y have respective supports $\{x_1, \dots, x_n\} \subset \mathcal{X}$ and $\{y_1, \dots, y_m\} \subset \mathcal{Y}$, the smoothest Markov kernel $P(X)$ in numerical implementation is represented by vectors

$$A \equiv (a_i) \in \mathbb{R}^n \quad \text{with} \quad a_i := P(X=x_i) \in [0, 1]$$

subject to $\sum_{i=1}^n a_i = 1$. And the smoothest Markov kernel $P(Y|X)$ in numerical implementation is given by matrices

$$B \equiv (b_{ij}) \in \mathbb{R}^{n \times m} \quad \text{with} \quad b_{ij} := P(Y=y_j|X=x_i) \in [0, 1]$$

subject to $\sum_{i=1}^n b_{ij} = 1$ for every j . This way, we can numerically compute the smoothest Markov kernel.

Fig. 6.3 visualizes examples of the smoothest Markov kernels $P(X)$ with the constraint parameters as listed in Tab. 6.1, one- (Fig. 6.3, plot A, B, and C) or multidimensional (Fig. 6.3, plot D, E, and F). Fig. 6.4 visualizes examples of the smoothest Markov kernels $P(Y|X)$ with the constraint parameters as shown in Tab. 6.2. We observe that having imposed the pre-specified constraints of the first two moments, the smoothest Markov kernels gain indeed a “smooth” shape. Actually, they are truncated exponential distributions of order up to 2.

6. Discovering Causal Order by Properties of Conditionals

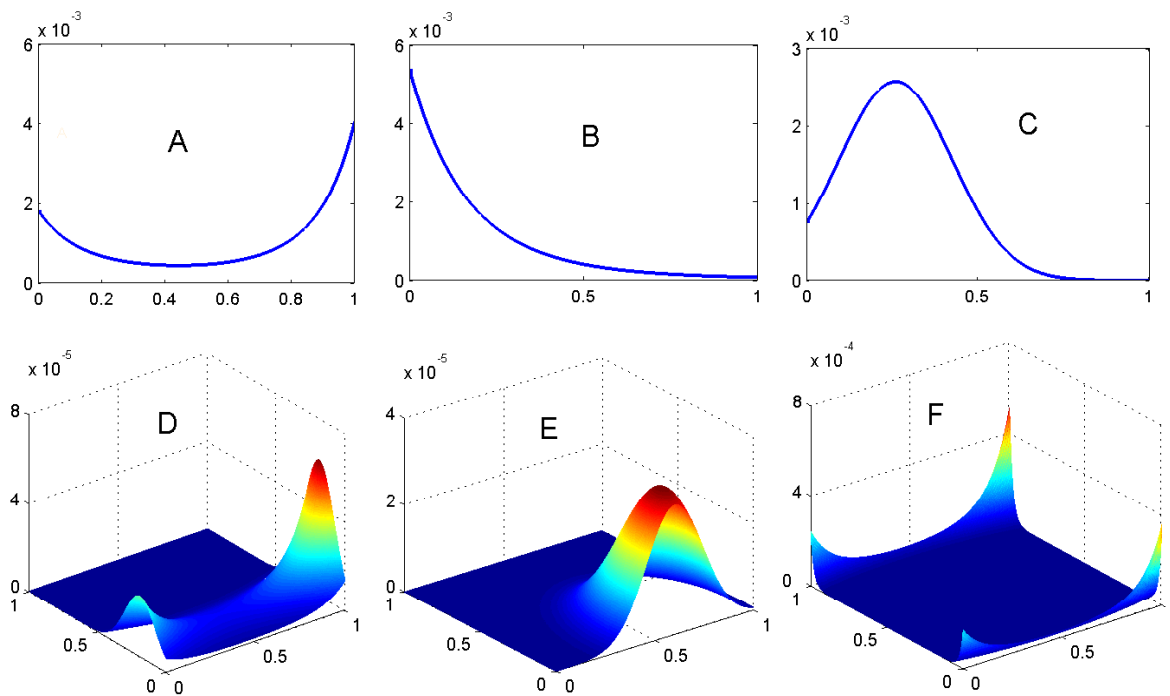


Figure 6.3.: Examples of the smoothest Markov kernel $P(X)$ for cause X with a scalar range (plot A, B, and C) and a two-dimensional range (plot D,E, and F). The constraints for first and second moments are given in Tab. 6.1.

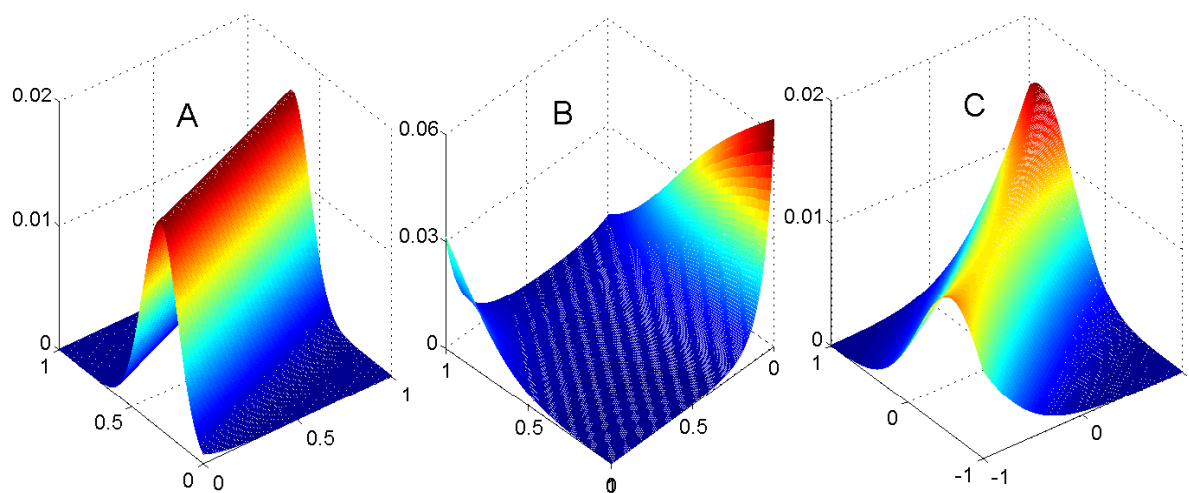


Figure 6.4.: Examples of the smoothest Markov kernel $P(Y|X)$ of effect Y given cause X . X, Y are both scalar variables. The constraints for first and second moments are given in Tab. 6.2.

6.4. Examples of smoothest Markov kernels

Plots in Fig. 6.3	A	B	C	D	E	F
\mathcal{X}	$[0, 1]$	$[0, 1]$	$[0, 1]$	$[0, 1] \times [0, 1]$	$[0, 1] \times [0, 1]$	$[0, 1] \times [0, 1]$
μ^X	0.60	0.20	0.28	$\begin{pmatrix} 0.20 \\ 0.60 \end{pmatrix}$	$\begin{pmatrix} 0.20 \\ 0.60 \end{pmatrix}$	$\begin{pmatrix} 0.58 \\ 0.60 \end{pmatrix}$
α^X	0.48	0.08	0.10	$\begin{pmatrix} 0.05 & 0.15 \\ 0.15 & 0.48 \end{pmatrix}$	$\begin{pmatrix} 0.06 & 0.13 \\ 0.13 & 0.38 \end{pmatrix}$	$\begin{pmatrix} 0.56 & 0.40 \\ 0.40 & 0.49 \end{pmatrix}$

Table 6.1.: Constraint parameters for plots in Fig. 6.3. X is assumed to be a scalar variable in columns A, B, C, and a two-dimensional vector in columns D, E, F. The first row shows the ranges, the second shows the postulated first moments, the last row the postulated second moments (from top to bottom).

Plots in Fig. 6.4	\mathcal{X}	μ^X	α^X	\mathcal{Y}	μ^Y	α^Y	β^{XY}
A	$[0, 1]$	0.64	0.43	$[0, 1]$	0.55	0.32	0.36
B	$[0, 1]$	0.65	0.55	$[0, 1]$	0.45	0.35	0.38
C	$[-1, 1]$	0.65	0.55	$[-1, 1]$	0.45	0.35	0.36

Table 6.2.: Constraint parameters for examples A, B, and C as shown in Fig. 6.4. X and Y are in all three cases scalars. The first three columns show the value range of X , its first and second moment (from left to right). The next three columns show the value range of Y , its first and second moment. The last column shows the mixed second moment of X and Y .

6.4.2. Analytical solution on hybrid (binary and real-valued) domain

An interesting example is the smoothest Markov kernel on the hybrid domain, namely binary X ($x \in \{\pm 1\}$) and real-valued Y ($y \in \mathbb{R}$). In such case, the optimization has a closed-form solution.

For hypothetical causal order “ $X \rightarrow Y$ ”, the smoothest Markov kernel of X is just the observed relative frequencies. The smoothest Markov kernels $P(Y|X)$ are Gaussian conditionals with a single variance and two different expectations. In other word, the smoothest Markov kernel presents a linear shift of the expected values by a multiple of X . We denote the kernels below with \mathcal{Q} to distinguish from those of the reversed causal order, which are denoted by \mathcal{R} .

$$\begin{aligned} \mathcal{Q}(x_{-1}) = p & \quad \text{and} \quad \mathcal{Q}(x_{+1}) = 1 - p \\ \mathcal{Q}(Y|x_{-1}) \propto \mathcal{N}(\mu_{-1}, \sigma^2) & \quad \text{and} \quad \mathcal{Q}(Y|x_{+1}) \propto \mathcal{N}(\mu_{+1}, \sigma^2) . \end{aligned}$$

For the other hypothetical causal order “ $Y \rightarrow X$ ”, the smoothest Markov kernel is a Gaussian distribution for the continuous cause Y and a hyperbolic tangent function for the binary effect X .

$$\begin{aligned} \mathcal{R}(Y) & \propto \mathcal{N}(\mu, \sigma_0^2) \\ \mathcal{R}(x_{-1}|Y) = \frac{1}{2} - \frac{1}{2} \tanh(\lambda y + \nu) & \quad \text{and} \quad \mathcal{R}(x_{+1}|Y) = \frac{1}{2} + \frac{1}{2} \tanh(\lambda y + \nu) , \end{aligned}$$

where $\lambda, \nu \in \mathbb{R}$ are chosen such that the constraints are satisfied. The derivations are provided in Appendix A.5. Note that the joint measures induced by smoothest Markov kernels subject to different hypothetical causal orders are different, since for $Y \rightarrow X$, cause Y exhibits a unimodal distribution, whereas for the reversed direction $X \rightarrow Y$ the smoothest Markov kernels can lead to a bimodal distribution for effect Y as its marginal distribution. This agrees with the motivational example in Section 6.1.

It is noteworthy that it is possible that there exist no λ, ν in the expressions for \mathcal{R} that satisfy the desired constraints of the first two moments. For example, let $X \rightarrow Y$ be the generating model given by Markov kernels $\mathcal{Q}(x_{\pm 1}) = \frac{1}{2}$ and $\mathcal{Q}(Y|x_{\pm 1}) \propto \mathcal{N}(\mu_{\pm 1}, \sigma^2)$ with very small σ^2 , i.e., the two Gaussians are highly separated. The observed data lead to the constraints $E[X] = E[Y] = 0$ and $E[X^2] = E[Y^2] = E[XY] = 1$. One can easily check that there is no kernel $P(X|Y)$ satisfying these constraints. The infeasibility here is due to the fact that the empirical distribution of Y (having almost only the values ± 1 as its value set) differs strongly from the supposed “smoothest” Gaussian distribution $\mathcal{N}(0, 1)$ for the wrong hypothetical causal model. A pragmatic way to handle infeasible constraints is therefore to consider them as hints that the true distribution differs so strongly from the supposed smoothest one that the corresponding causal hypothesis should be rejected. Note that in the case of two binary variables X, Y , i.e., $\sigma^2 = 0$, our inference rule will be indifferent for both hypothetical causal order, which we will show through a more general statement for binary domains in Section 6.4.3 later.

6.4.3. Analytical solution on binary domain

For binary domains, the meaning of the smoothest Markov kernel cannot be visualized by a “smooth” shape. It is another kind of simplicity. Without loss of generality, we henceforth assign $\{0, 1\}$ as the value set of a binary variable. We can further simplify the solution of Eq. (6.3) for a binary variable X_j and eliminate Lagrange multipliers $\gamma(an_j)$ in Eq. (6.3) to specify the smoothest Markov kernel in a convenient and elegant form, since we have

$$\frac{P(X_j=1 | an_j)}{1 - P(X_j=1 | an_j)} = \exp(\theta_0 x_j + x_j \sum_{i=1}^j \theta_i x_i) \quad \Rightarrow \quad P(X_j=1 | an_j) = \frac{1}{2} \left(1 + \tanh\left(\lambda + \sum_{i=1}^{j-1} \lambda_i x_i\right) \right)$$

with

$$\lambda = \frac{1}{2} (\theta_0 + \theta_j) \quad \text{and} \quad \lambda_i = \frac{1}{2} \theta_i \quad \text{for } i = 1, \dots, j-1.$$

The kernel can be interpreted in the following way. The influence of each ancestor X_i ($i < j$) on X_j can be characterized by the parameter λ_i . If λ_i is negative, X_i has a repressive effect on the occurrence of X_j (independent of the value assignment of the other ancestors). If λ_i is positive, X_i is conducive to X_j . Such a unique separation into repressive and conducive variables is a feature of the simplest cause-and-effect relationship. More precisely, one has to ask whether the map

$$(x_1, \dots, x_{j-1}) \mapsto P(X_j=1 | x_1, \dots, x_{j-1})$$

is simple, since smoothness of the function $x_j \mapsto P(x_j | x_1, \dots, x_{j-1})$ does not make sense for a fixed (x_1, \dots, x_{j-1}) in contrast to a real-valued variable or discrete variable on a large domain. Note that this simplicity feature of the smoothest Markov kernels makes already sense if two ancestors are present.

More generally, the simplicity feature of the smoothest Markov kernels can be naturally considered as part of a hierarchy of exponential models (see e.g., [7] for an information geometry approach for exponential hierarchies of unconditional distributions) as follows. We may represent every strictly positive Markov kernel of a binary variable X_j with ancestors AN_j by

$$P(X_j=1 | an_j) = \frac{1}{2} \left(1 + \tanh(f_j(an_j)) \right)$$

with the function

$$f_j(an_j) = \lambda + \sum_{i_1=1}^{j-1} \lambda_{i_1} x_{i_1} + \sum_{i_1, i_2=1}^{j-1} \lambda_{i_1 i_2} x_{i_1} x_{i_2} + \dots + \sum_{i_1, \dots, i_{j-1}=1}^{j-1} \lambda_{i_1 \dots i_{j-1}} x_{i_1} \dots x_{i_{j-1}},$$

since the “tanh” function is invertible in the open interval $(-1, 1)$. We define $\mathcal{K}_k^{(j)}$ as the set of conditional probability distributions $P(X_j | AN_j)$ for which all coefficients λ in f_j with more than k indices vanish and shall drop the superscript j when this will lead to no confusion. We obtain the hierarchy

$$\mathcal{K}_0 \subset \mathcal{K}_1 \subset \dots \subset \mathcal{K}_{j-1}.$$

6. Discovering Causal Order by Properties of Conditionals

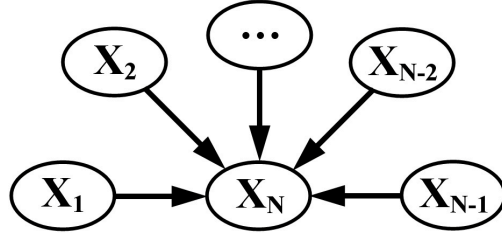


Figure 6.5.: Graphical representation of an OR gate with $n - 1$ independent input bits.

One can easily prove that the constrained entropy maximization defined above leads to terms in \mathcal{K}_k if the set of constraints is extended by terms up to moments $E[X_{i_1} X_{i_2} \dots X_{i_k}]$ of order k . We therefore consider the above hierarchy as a straightforward definition of the complexity of Markov kernels and observe that the “smoothest” kernels are in \mathcal{K}_1 which is the first non-trivial class, since for all kernels in \mathcal{K}_0 the variables AN_j do not influence X_j at all.

We define $\mathcal{M}_1^{X_1, \dots, X_n}$ as the set of joint measures on (X_1, \dots, X_n) for which all Markov kernels $P(x_j | an_j)$ are in $\mathcal{K}_1^{(j)}$. The asymmetry of the set \mathcal{M}_1 with respect to a reordering of variables is decisive for the applicability of our approach to binary domains. The next subsection elaborates on this by means of Boolean functions OR/AND as models for elementary causal mechanism.

6.4.4. Identifying causal order of OR/AND gates by Markov kernels

The Boolean functions OR/AND are ideal simplified models for many elementary causal relations in real life where an effect depends on several respective sufficient or necessary conditions. For instance, a plant grows if a sufficient amount of water, light, and fertilizer is available and it dies if at least one of these necessary conditions is not satisfied.

Remarkably, the Markov kernels describing OR/AND gates are both in the closure of class \mathcal{K}_1 . To see this, we study an OR gate in detail. Let X_1, \dots, X_{n-1} be the binary variables that correspond to the input bits of an OR gate and X_n the output (see Fig. 6.5).

The conditional probabilities of X_n can be written as

$$P(X_n=1 | an_n) := 1 - \prod_{i=1}^{n-1} (1 - x_i) .$$

Defining

$$P_k(X_n=1 | an_n) := \frac{1}{2} \left(1 + \tanh \left(-k + 2k \sum_{i=1}^{n-1} x_i \right) \right) ,$$

we have

$$\lim_{k \rightarrow \infty} P_k(x_n | an_n) = P(x_n | an_n) ,$$

i.e., $P(X_n | AN_n) \in \mathcal{K}_1^{(n)}$.

6.4. Examples of smoothest Markov kernels

Suppose the joint distribution $P(X_1, \dots, X_n)$ is generated by an OR gate when the inputs X_1, \dots, X_{n-1} are randomly chosen according to the uniform distribution, i.e., $P(x_j | an_j) = \frac{1}{2}$ for all $j < n$ and $an_j \in \{0, 1\}$. Clearly the joint measure P is in $\mathcal{M}_1^{X_1, \dots, X_n}$. Now we consider an ordering of the variables where the output is not at the end. Without loss of generality, we choose the order X_2, \dots, X_n, X_1 . We have

$$P(X_1=1 | x_2, x_3, \dots, x_{n-1}, X_n=1) = \begin{cases} 1 & \text{for } x_2 = x_3 = \dots = x_{n-1} = 0 \\ \frac{1}{2} & \text{otherwise} \end{cases} \quad (6.4)$$

and

$$P(X_1=1 | X_2 = \dots = X_n = 0) = 0. \quad (6.5)$$

Note that the event $X_n = 0$ and $X_i = 1$ for some $i \in \{2, \dots, n-1\}$ does not occur and the corresponding conditional probabilities need not to be specified. We will show that there is no Markov kernel in the closure of \mathcal{K}_{n-3} that satisfies Eq. (6.4). We are particularly interested in the Markov kernel of X_1 since it depends on $n-1$ variables and is therefore the natural candidate for being the most complex Markov kernel. We write

$$P(X_1=1 | x_2, \dots, x_n) = \frac{1}{2} \left(1 + \tanh(f(x_2, \dots, x_n)) \right),$$

where f is an appropriate function. Define a function \tilde{f} with $n-2$ arguments by

$$\tilde{f}(x_2, \dots, x_{n-1}) := f(x_2, \dots, x_{n-1}, x_n=1).$$

If the kernel of Eq. (6.4) was in the closure of \mathcal{M}_{n-3} , there existed a sequence f_k with polynomials of degree $n-3$ and a corresponding sequence \tilde{f}_k of degree $n-3$ such that $\tilde{f}_k(x_2, \dots, x_{n-1})$ tended to infinity for $x_2 = x_3 = \dots = x_{n-1} = 0$ and to zero otherwise. Elementary linear algebra arguments show that the space of polynomials of degree $n-3$ would then contain the element g with

$$g(x_2, \dots, x_{n-1}) = \begin{cases} 1 & \text{for } x_2 = x_3 = \dots = x_{n-1} = 0 \\ 0 & \text{otherwise} \end{cases}$$

This is however not true since the unique function g satisfying these constraints is given by

$$g(x_2, \dots, x_{n-1}) = \prod_{i=2}^{n-1} (1 - x_i),$$

which is a polynomial of degree $n-2$. The lower bound on the degree is tight because there is indeed a sequence of polynomials of degree $n-2$ that induce Markov kernels which satisfy the constraints of Eq. (6.4) and Eq. (6.5) in the limit. The sequence $(f_k)_{k \in \mathbb{N}}$ of functions, given by

$$f_k(x_2, \dots, x_n) := k \left(2(x_n - 1) + \prod_{i=2}^{n-1} (1 - x_i) \right),$$

6. Discovering Causal Order by Properties of Conditionals

tends to $-\infty$ for $x_n=0$ and the induced conditional probability satisfies therefore the constraint in Eq. (6.5). Moreover, the condition of Eq. (6.4) is also satisfied.

This shows that an OR gate induces a joint measure that is in \mathcal{M}_1 when considered with respect to the correct generating causal order. By inverting input and output one can instantly see that this is true also for an AND gate. The remarks show that for $n \geq 4$ the set \mathcal{M}_1 is not invariant with respect to a reordering of n variables and kernels in \mathcal{K}_1 can lead to joint distributions defining kernels which are in \mathcal{K}_{n-2} but not in \mathcal{K}_{n-3} . This implies that our inference proposal can in principle identify the output of an OR/AND gate as the effect and its random inputs as causes whenever the number of inputs is at least 3. Of course, the number of data points in the sample should be large enough to allow a reliable estimation of the joint measure.

This theoretical result is actually not very surprising. The intuition behind the result is that reversing at least one of the arrows in the causal model as shown in Fig. 6.5 generates dependence among the inputs, which can only be canceled out by accident. This dependence can only be described by sophisticated high-order terms of inputs. This is why the concatenated entropy maximization leads to conditionals that, in turn, generate different joint measures when different orders are chosen in the maximization procedure. It will be apparent later that this difference among the joint measures is essential for our inference rule.

Based on independence relations, the constraint-based approach, e.g., PC, is also capable of distinguishing output from inputs, if the inputs of OR/AND gates are indeed independent. Our hope is that the evidence of high-order terms (non-smoothness or non-simplicity) will survive when the inputs are dependent in a simple manner, because such noisy OR gates have many of the properties of linear systems (see [124, 29, 70] for more discussions). Although it is hard to generally justify Assumption 6, in the sense that the majority of natural causal relationships have such properties of simplicity, numerical experiments in Section 6.8 will show that there are some real life cases where our assumption appears to be reasonable.

6.5. pIMK causal order discovery algorithm

Having defined the smoothest Markov kernels, we move to the issue of model selection. The idea is that we evaluate the goodness of fit to finite n data points by means of the joint measures implied by the corresponding smoothest Markov kernels with respect to different hypothetical causal orders. For this purpose, we introduce maximum likelihood and minimum distance estimation to select models.

One possible approach to prefer one of the hypothetical causal orders is the maximum likelihood method. The method assigns an order π to be causal, if its derived smoothest Markov kernels lead to a joint measure that has the maximum log-likelihood score l_π given data. We briefly describe the case of estimating the causal direction between only two observed variables X, Y . The joint measures \mathcal{Q} and \mathcal{R} induced by the smoothest Markov kernels corresponding to two hypothetical causal orders are given. Based on the n observed data points (x_i, y_i) , we

calculate the respective log-likelihood scores

$$l_{X \rightarrow Y} := \sum_{i=1}^n \ln \left(\mathcal{Q}(x_i, y_i) \right) \quad \text{and} \quad l_{Y \rightarrow X} := \sum_{i=1}^n \ln \left(\mathcal{R}(x_i, y_i) \right)$$

and prefer the causal order with larger log-likelihood. The extension to more than two hypothetical orders is trivial.

We shall interpret the maximum score merely as showing that the set of Markov kernels with respect to that particular causal order seem to be the closest to the smoothest one. Note that we do not expect at all that the calculated joint measures is a good approximation for the true probabilities, it is more likely that the joint measures could all be rejected with high level of confidence. The task is nonetheless to decide which measure provides the best fit.

Another possibility to check which joint measure fits better to the observed data is the concept of minimum distance estimate as described by Devroye et al. [47]. To explain the idea, we first consider the situation with two distinct resulting joint measures \mathcal{Q} and \mathcal{R} . The set

$$\mathcal{A} = \{(x, y) \in \mathcal{X} \times \mathcal{Y} : \mathcal{Q}(x, y) > \mathcal{R}(x, y)\}$$

is defined as the Scheffé set [136] for an ordered pair of distributions $(\mathcal{Q}, \mathcal{R})$. Moreover,

$$\mu_n(\mathcal{A}) := \frac{1}{n} \sum_{i=1}^n 1_{\mathcal{A}}((x_i, y_i)),$$

where $1_{\mathcal{A}}$ denotes the characteristic function of a set \mathcal{A} . Then, $\mu_n(\mathcal{A})$ is the observed relative frequency for the set \mathcal{A} after n observations. By preferring the measure for which the probability of \mathcal{A} is closer to the observed relative frequency we have a good chance to prefer the measure that has the smaller maximum variation distance to the true distribution. In the selection problem with two candidates, we say that \mathcal{Q} wins against \mathcal{R} when

$$d_{X \rightarrow Y} := \left| \sum_{\mathcal{A}} \mathcal{Q} - \mu_n(\mathcal{A}) \right| < \left| \sum_{\mathcal{A}} \mathcal{R} - \mu_n(\mathcal{A}) \right| =: d_{Y \rightarrow X}. \quad (6.6)$$

Our inference rule assigns the causal hypothesis that corresponds to the so-called ‘‘Scheffé tournament’’ winner as ‘‘true’’.

For a selection problem within k alternatives P_π , we run a competition, the so-called ‘‘Scheffé tournament’’, with $k(k-1)/2$ matches among them, one for each ordered pair. For each P_π , we total the number of wins and declare the measure P_π with the maximum number of wins the tournament winner. If there is more than one winner, repeat the competition within the winners so long as no winners can be eliminated any more. In the end, we consider the tournament winners most plausible, supported by the given dataset \mathcal{D} and hence interpret the corresponding orders, induced by π , as causal.

In comparison to the maximum log-likelihood method, the advantage of the minimum distance method is that it is less sensitive to small deviations between true and hypothetical probabilities

6. Discovering Causal Order by Properties of Conditionals

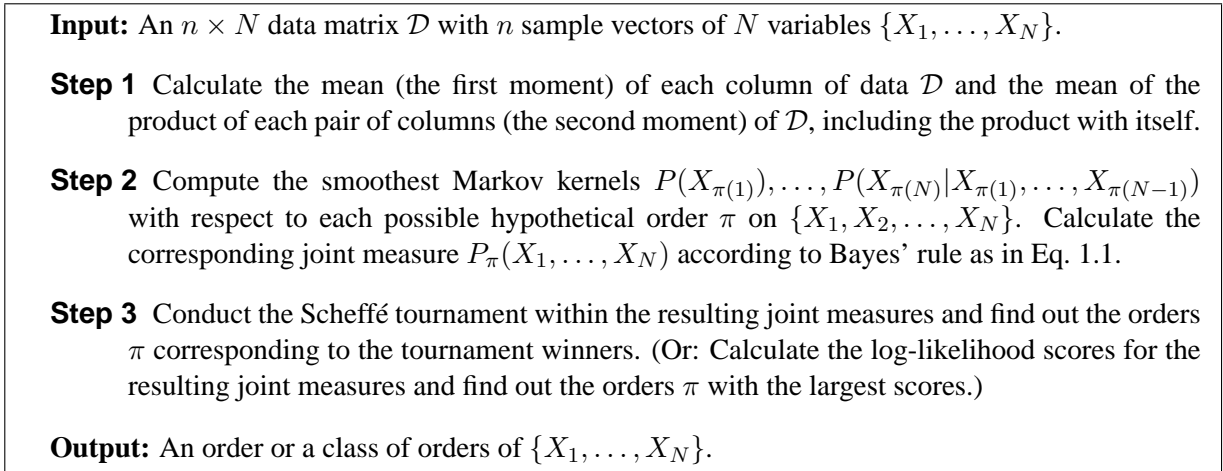


Figure 6.6.: pMK causal order discovery algorithm.

if both measures differ within a region of small probability, although in all our experiments the estimated causal directions coincide. Because the difference of the minimum distance between the true and wrong causal models are much more significant than that of the log-likelihood scores, we prefer in the following the minimum distance estimation to the maximum likelihood method. Now, we summarize the ideas and describe the pMK causal order discovery algorithm as shown in Fig. 6.6, under the plausible Markov kernel (short: pMK) assumption.

6.6. Experiments with data on binary domains

Section 6.4.4 showed some theoretical consequences of the smoothest Markov kernels on binary domains, which claimed that pMK is capable of discovering OR/AND gates with independent inputs. Simulated experiments shall explore how robust pMK behaves with regard to noises and sample sizes.

6.6.1. Simulated noisy OR data

We sampled data of sample sizes ranging from 20 to 200 from three 3-bit noisy OR gates as shown in Tab. 5.2 and ran pMK algorithm to learn the causal ordering of the four variables $\{X_1, X_2, X_3, X_4\}$. In this experiment, the output of pMK is 6 orderings of them. The last variable in the orderings is always the same. The ordering of the first three is arbitrary. We transform the resulting order of variables into a partially directed graph to make the evaluation comparable with other algorithms. The graph is fully connected, since pMK is not capable of removing unnecessary edges.

To evaluate pMK regarding different samples, we introduce the following scoring system for all directions in the structure. Suppose pMK identified X_4 as the output. We assign the probability score 100% to the direction $X_i \rightarrow X_4$ and 0% to $X_i \leftarrow X_4$ ($i \in \{1, 2, 3\}$). For for the

6.6. Experiments with data on binary domains

Model	3-Bit-IndDet					3-Bit-IndPro					3-Bit-DepPro				
	20	50	100	150	200	20	50	100	150	200	20	50	100	150	200
$X_1 \rightarrow X_2$	49.1	51.5	52.3	50.4	48.3	50.4	48.0	48.7	48.4	47.4	50.4	49.4	52.1	51.2	52.8
$X_1 \leftarrow X_2$	50.9	48.5	47.7	49.6	51.7	49.6	52.0	51.3	51.6	52.6	49.6	50.6	47.9	48.8	47.2
$X_1 \rightarrow X_3$	49.4	51.2	52.5	51.2	50.8	49.7	48.3	47.1	47.4	46.0	50.5	49.3	51.9	55.8	61.6
$X_1 \leftarrow X_3$	50.6	48.8	47.5	48.8	49.2	50.3	51.7	52.9	52.6	54.0	49.5	50.7	48.1	44.2	38.4
$X_1 \rightarrow X_4$	69.9	85.8	92.4	95.8	97.4	62.5	71.3	76.5	78.2	79.0	53.6	64.2	66.4	60.5	58.3
$X_1 \leftarrow X_4$	30.1	14.2	7.6	4.2	2.6	37.5	28.7	23.5	21.8	21.0	46.4	35.8	33.6	39.5	41.7
$X_2 \rightarrow X_3$	50.7	49.5	49.2	49.8	51.2	49.0	50.3	48.3	48.9	48.6	50.1	50.1	49.7	54.6	58.8
$X_2 \leftarrow X_3$	49.3	50.5	50.8	50.2	48.8	51.0	49.7	51.7	51.1	51.4	49.9	49.9	50.3	45.4	41.2
$X_2 \rightarrow X_4$	69.9	85.8	92.4	95.8	97.4	61.8	72.8	77.8	79.7	81.7	53.2	64.8	64.3	59.4	55.4
$X_2 \leftarrow X_4$	30.1	14.2	7.6	4.2	2.6	38.2	27.2	22.2	20.3	18.3	46.8	35.2	35.7	40.6	44.6
$X_3 \rightarrow X_4$	69.9	85.8	92.4	95.8	97.4	62.7	72.6	79.5	80.8	83.1	53.2	64.8	64.6	54.7	46.7
$X_3 \leftarrow X_4$	30.1	14.2	7.6	4.2	2.6	37.3	27.4	20.5	16.9	10.0	46.8	35.2	35.4	45.3	53.3

Table 6.3.: Statistics of outputs learned by pMK on data sampled from noisy 3-bit OR gates as shown in Tab. 5.2. The entries in the rows of $X_i \leftarrow X_4$ ($i = 1, 2, 3$) show how often (in percentage) X_4 is correctly identified as the output of OR gates. An entry of 50% indicates an indeterminate edge by pMK, while a score of 100% indicates a deterministic orientation by pMK.

other directions $X_i \rightarrow X_j$ ($i, j \in \{1, 2, 3\}$ and $i \neq j$), we assign the probability score 50%, since we do not have any information to prefer some of them. Tab. 6.3 shows the average score for all possible arrows in the fully connected structure after 1000 replications of sampling. A score of 50% indicates an indeterminate edge by pMK, while a score of 100% indicates a deterministic orientation by pMK.

The larger the sample size, the better the pMK algorithm performed. In comparison to Tab. C.4 in Appendix C.2, we can see that the performance of the pMK algorithm (in the case of sample size of 200) is competitive with the constraint-based PC algorithm. A main shortcoming of pMK is that it is only feasible for a small number of variables. Otherwise the number of hypothetical causal orders and the dimension of the joint probability vector would lead to intractable computational problems. For this reason, we propose in the following to combine the pMK with PC to get the advantages of both approaches.

6.6.2. Personal income data

We study the relationships between annual personal income and various demographic factors. The data come from the US current population survey (CPS). One dataset is transformed from CPS 1995 with altogether 149,642 records. The binary version contains entries for 112,164 persons with age at least 16. The other dataset contains data from CPS 2001 for 13,803 per-

6. Discovering Causal Order by Properties of Conditionals

Data	CPS 1995		CPS 2001	
	Distance ($\times 10^{-3}$)	Winner	Distance ($\times 10^{-3}$)	Winner
(P_4, P_1)	$d_4 = 0.2324 < 1.6186 = d_1$	P_4	$d_4 = 0.1530 < 5.1513 = d_1$	P_4
(P_4, P_2)	$d_4 = 0.2324 < 1.6183 = d_2$	P_4	$d_4 = 0.1530 < 5.1417 = d_2$	P_4
(P_4, P_3)	$d_4 = 0.2324 < 1.6189 = d_3$	P_4	$d_4 = 0.1530 < 5.1715 = d_3$	P_4
(P_3, P_1)	$d_3 = 1.4718 < 1.4729 = d_1$	P_3	$d_3 = 3.6131 < 3.6681 = d_1$	P_3
(P_3, P_2)	$d_2 = 1.2952 < 1.2963 = d_3$	P_2	$d_3 = 1.6253 < 1.6788 = d_2$	P_3
(P_2, P_1)	$d_2 = 4.1082 < 4.1092 = d_1$	P_2	$d_2 = 10.679 < 10.701 = d_1$	P_2

Table 6.4.: Results of the Scheffé tournaments on CPS data.

sons, age 16 and over, resident in the Pacific Division of United States. The “Pacific Division” comprises the states of Alaska, California, Hawaii, Oregon, and Washington. Both datasets were transcribed by D. Freedman of the Statistics Department, UC Berkeley and are available at <http://www.stat.berkeley.edu/~census>.

The variables that we consider include X_1 : SEX (gender), X_2 : I-STATUS (immigrant status), X_3 : E-LEVEL (educational level), and X_4 : INCOME (annual personal income). For our purpose, variables were transformed into binary ones, which stand for male or female, whether being native born in the US or not, whether having more than a Bachelor’s degree or less, whether having an annual income of more than \$50,000 or less.

We know that gender can only affect the other variables but it cannot be an effect of them. Furthermore, we assume that income is rather an effect of the others than a cause even though we cannot completely exclude causal arrows in the backward direction. For both datasets the causal hypotheses generated by the plMK algorithm were indeed consistent with this prior knowledge and assumptions.

The plMK algorithm has generated 4 different joint measures P_i with $i = 1, \dots, 4$ corresponding to orderings where the variable X_i is at the end. As reported in Tab. 6.4, in both datasets (CPS 1995 and 2000) P_4 is the winner of the Scheffé tournaments between pairs of measures. Although the absolute differences between log-likelihood scores are not very large, P_4 has the largest score in both datasets. Fig. 6.7 visualizes the graphical structure corresponding to P_4 . The undirected edges depict yet unspecified causal relations. The plMK algorithm identified personal INCOME as the effect in both datasets. It is remarkable that structures with the variable SEX at the end have obtained no wins at all (see Tab. 6.5).

6.7. Combining plMK with constraint-based algorithm

The intention behind the plausible Markov kernel assumption is not replacing conventional approaches that use independence relations. It rather should provide additional hints on the orientation of structure. Our inference rule can distinguish between causal structures that generate the same set of independence constraints, whereas PC is efficient if the underlying network is indeed sparse. To benefit both advantages, we suggest to combine PC and plMK. A pre-selection

6.7. Combining pIMK with constraint-based algorithm

Data	CPS 1995	CPS 2001
SEX as last variable in the ordering	0	0
I-STATUS as last variable in the ordering	6	12
E-LEVEL as last variable in the ordering	12	6
INCOME as last variable in the ordering	18	18

Table 6.5.: Total wins of the 4 distinguishable classes of causal orders on CPS data by the Scheffé tournaments.

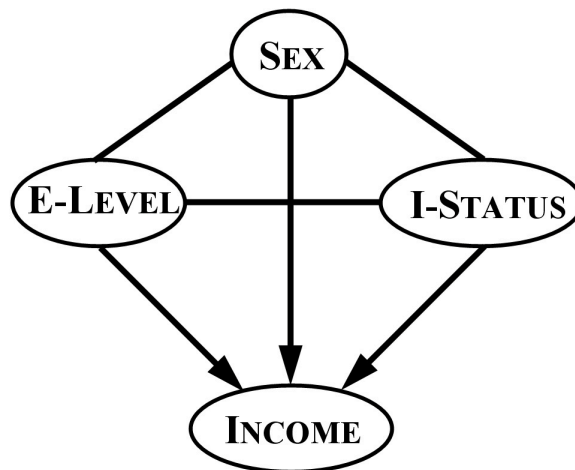


Figure 6.7.: Graphical representation of output generated by pIMK when applied to CPS data.

6. Discovering Causal Order by Properties of Conditionals

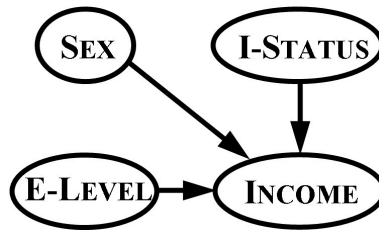


Figure 6.8.: Output generated by PC on CPS 2001.

of causal hypotheses through PC reduces the search space (possible orderings of variables) for pMK. The conventional approach tends to prefer directed graphs with small numbers of arrows. The pMK algorithm can additionally prefer those models where the corresponding Markov kernels are simple.

Now we study the demographic data in Section 6.6.2 again. We first start with the PC algorithm (with χ^2 test and $\alpha = 0.05$). Fig. 6.8 visualizes the result for CPS 2001, which contains only directed edges. The output is in agreement with the output of pMK saying that INCOME is the effect of the other variables.

The left plot of Fig. 6.9 shows the result for CPS 1995. Additional correlations between SEX and E-LEVEL and between I-STATUS and E-LEVEL are observed on CPS 1995. Due to the additional dependences the resulting graph is more complex compared to CPS 2001 and PC is incapable of making any statement about the orientation of the causal connection between E-LEVEL and INCOME. This means that a causal arrow from INCOME to E-LEVEL cannot be excluded. The pMK algorithm is here more specific since its output states that INCOME is the effect of all other variables, i.e., no arrow from INCOME to E-LEVEL is allowed.

Note, however, that the pMK is in other respects less specific than PC since it cannot distinguish (in the case of 4 binary variables) between different structures having the same variable at the end. Recall, for instance, that the results of pMK (see Fig. 6.7) did not show that SEX is not an effect of any other variables, the latter statement is only consistent with the class of preferred causal structures. Combining PC with pMK we may then orient the edge from E-LEVEL to INCOME as done in Fig. 6.9, right. This shows that a combination of PC and pMK leads to a completely determined causal structure, which states that INCOME is not a cause of any other variable and SEX/I-STATUS is not the effect of any other variable. This is consistent with our prior knowledge and assumptions.

The examples showed that the conventional PC works quite well on binary domains. The improvement of pMK in this respect is limited. However, the typical application of pMK is inference on hybrid models consisting of both continuous and discrete domains, in particular, pMK could make inference in case of only two dependent variables.

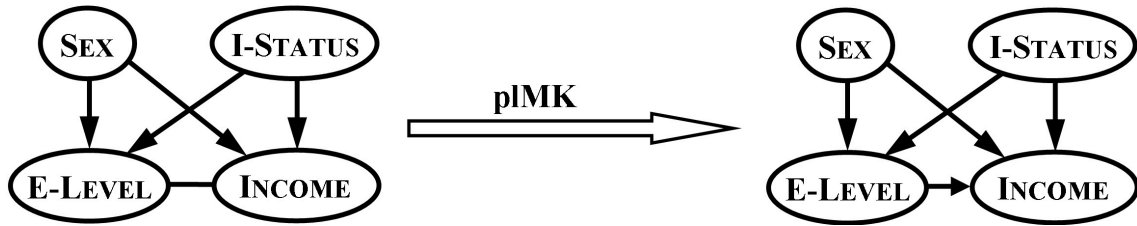


Figure 6.9.: Output generated by PC (left) and by PC+pIMK (right), when applied to CPS 1995.



Figure 6.10.: Graphical representation of relation between age and marriage status (M-Status), which is confirmed by output of pIMK.

6.8. Experiments with data on continuous domains

In this section, we demonstrate some real-world experiments to infer causal order between only two dependent variables, a case that cannot be treated by learning from independence relations or dependence measures.

6.8.1. Demographic data

We study the causal relation between the age of a person and her/his marriage status (M-Status). For this purpose, we use data from CPS 1995 and 2001 again, which are already discussed in Section 6.6.2. Only the cases with age 16 or over are considered.

The variable M-STATUS has the binary value of “never married” or else, while AGE is an integer. The observed correlations are strong, 0.4995 for CPS 1995 and 0.5238 for CPS 2001. We assume that the age of a person causally determine his/her marriage status, not vice versa (see Fig. 6.10). The outputs of pIMK learned from CPS 1995 and CPS 2001 are indeed consistent with this prior knowledge. As one can see from the Tab. 6.6, the log-likelihood scores with respect to the correct causal direction are always larger than that of the reversed one. The conducted Scheffé tournament also confirmed this result.

Another example we studied was the causal relation between SEX and annual personal INCOME, based on the CPS data. We assume that the gender of a person (binary) influences his/her personal income (real-valued), not vice versa (see Fig. 6.11). The outputs learned by pIMK from CPS 1995 and 2001 data are, however, not consistent with this prior knowledge, since the conducted Scheffé tournament did not yield the desired results. But if we took the logarithm of the continuous values of INCOME, pIMK preferred the correct causal direction. The log-likelihood scores provided qualitatively the same results. In our calculation, the untransformed as well as log-transformed continuous domains are discretized into round 5,000 intervals. The observed

6. Discovering Causal Order by Properties of Conditionals

Data	Log-likelihood scores		Distance measure based on Scheffé set	
	AGE \rightarrow M-STATUS	M-STATUS \rightarrow AGE	AGE \rightarrow M-STATUS	M-STATUS \rightarrow AGE
CPS 1995	-5.1374×10^5	-5.1440×10^5	0.0264	0.0577
CPS 2001	-6.3056×10^4	-6.3145×10^4	0.0209	0.0537

Table 6.6.: Experimental results for data from the CPS 1995 and 2001. Maximum likelihood estimation and Scheffé tournament both prefer the causal structure as shown in Fig. 6.10.

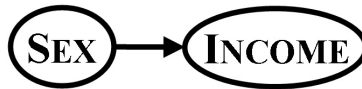


Figure 6.11.: Graphical representation of relation between gender (sex) and income, which is confirmed by output of pMK.

correlation rose from 0.2112 to 0.2502 in CPS 1995 and from 0.2828 to 0.2998 in CPS 2001 by the use of log-transformation.

Tab. 6.6 summarizes the results of the Scheffé tournament based on the original and log-transformed data. This partially negative result indicates that more flexible notions of plausible conditionals are desirable. In other words, the family of conditionals that are considered smooth should be large enough to contain, e.g., log-normal distributions but small enough to not contain mixtures of those.

It should be mentioned that the concept of plausible Markov kernels has also its limitation. Suppose an underlying model $X \rightarrow Y$ with Markov kernels $Q(x_{\pm 1}) = \frac{1}{2}$ and $Q(Y|x_{\pm 1}) \propto \mathcal{N}(\mu_{\pm 1}, \sigma^2)$ with sufficiently large σ^2 . Such situation is featured by a weak correlation between X and Y . The inference rule of pMK will in fact run into difficulties, unless there are a large number of data samples to recognize the mixture of distributions. The following example shows why this is not very surprising.

The left plot of Fig. 6.12 is the Gaussian mixture $0.5\mathcal{N}(-1, 1.5) + 0.5\mathcal{N}(1, 1.5)$, which corresponds to a correlation coefficient of about 0.5 when the Gaussian components are labeled by a

Data	with original domain		with log-transformed domain	
	SEX \rightarrow INCOME	INCOME \rightarrow SEX	SEX \rightarrow INCOME	INCOME \rightarrow SEX
CPS 1995	0.1047	0.1041	0.0523	0.0535
CPS 2001	0.1567	0.1543	0.0151	0.0156

Table 6.7.: Experimental results for data from the CPS 1995 and 2001. Scheffé tournament both prefer the causal structure as shown in Fig. 6.11. The first two columns are results for the original continuous domain of INCOME and the last two are results with the log-transformed domain.

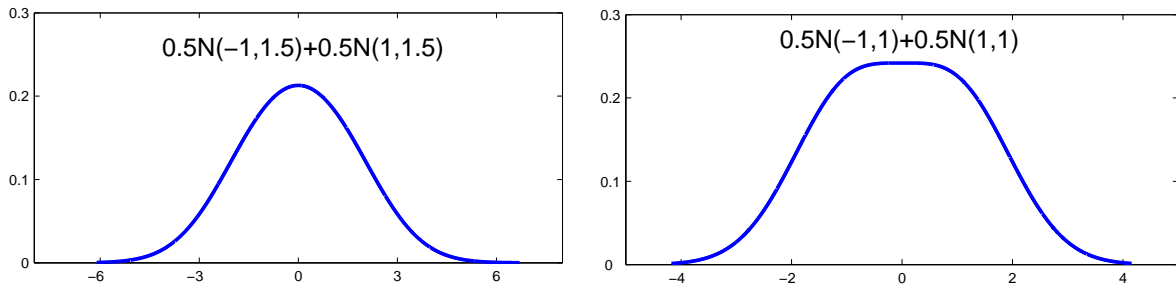


Figure 6.12.: Recognizability of mixture of two Gaussian distributions.

binary variable. The mixture is still unimodal and the shape differs only slightly from the Gaussian distribution. That is why our inference rule requires large samples for small correlations. The plateau in the right plot of Fig. 6.12 can already be taken as a hint for a mixture of ensembles, although the distribution is also unimodal. The correlation coefficient is about 0.7. This shows that the bimodal case corresponds to very large correlations.

Summing up, as long as the sample size or the underlying distribution allows us a reliable identification of mixture, the plausibility of Markov kernels in such hybrid models might help us to guess the “true” causal direction. The larger the sample size, the stronger the correlation, the better will our inference rule work.

Furthermore, we should mention the following potential objection against the inference rule of pMK. Given an effect variable Y that is in a linear way influenced by a very large number of causes X_1, \dots, X_n . Then the marginal distribution of Y is approximately Gaussian despite our claim that the distribution of the effect should typically be less smooth. However, this is a misunderstanding of our idea since the task is rather to identify dominant causes. In real life, every cause is influenced by further causes. If the latter influence the former such that the contribution of each single variable is small we would rather consider the superposition of all these small influences as background noise and set the influenced variable at the beginning of the causal order. According to such a viewpoint, we should prefer variables whose marginal distribution is stable (like, for instance, Gaussians or gamma distributions) as those that correspond to the causes, i.e., the variables at the beginning of the causal order. To develop a notion of simplicity that would also consider other stable distributions (apart from Gaussians) as “extremely smooth” with state-of-the-art machine learning methods could be interesting. The next chapter provides a kernel-based approach.

In principle, LiNGAM can also be used for causal inference between only two variables. However, the current version of LiNGAM is yet only applicable to continuous variables and cannot handle discrete, vectorial or hybrid domains. Our pMK algorithm can treat them straightforwardly. It should be mentioned that if the observed variables are in fact multivariate Gaussian-distributed, neither pMK nor LiNGAM can provide any information about causal relationships among them.

6. Discovering Causal Order by Properties of Conditionals



Figure 6.13.: Graphical representation of relation between date and temperature, which is confirmed by output of pIMK.

Variables	DATE: (X, Y)	TEMPERATURE ($^{\circ}\text{C}$)
Value set	$\{(x, y) x^2 + y^2 = 1\} \subseteq \mathbb{R}^2$	$[-23, 25] \subseteq \mathbb{R}$
1st moment	$(0.0022, -0.0009)$	5.7053
2nd moment	$\begin{pmatrix} 0.5019 & 0.0000 \\ 0.0000 & 0.4981 \end{pmatrix}$	84.6079
2nd mixed moment	$(-3.9702, -1.4548)$	

Table 6.8.: Value sets and observed statistical moments for the temperature dataset of Furtwangen.

6.8.2. Temperature data

Another example is an experiment with a meteorological dataset on continuous domains. We examined the causation between two variables, namely DATE (dates of the year) and TEMPERATURE (daily average temperatures). Common sense tells us that the seasonal cycle is a cause of temperature variation (see Fig. 6.13 for the graphical representation), not vice versa.

A dataset of daily average temperatures in Furtwangen (Black Forest, Germany) of 25 years (from Jan. 1, 1979 to Jan. 31, 2004) with 9162 entries was analyzed. The dataset contains also the temperatures at 7 am, 2 pm and 6 pm o'clock every day, as well as the daily maximum and minimum. Each day begins and ends at 6pm.

Due to the cyclic property of dates of the year, we assign the unit circle, a proper subset of \mathbb{R}^2 , to the value set of variable DATE (X, Y) with $\text{date} \in \{(x, y) | x^2 + y^2 = 1\}$. This value set can be parameterized, for example, by $x = \cos\left(\frac{2\pi}{366}k\right)$ and $y = \sin\left(\frac{2\pi}{366}k\right)$ with $k = 1, \dots, 366$ (maximum days per year). Note that we take the natural representation of data as a priori knowledge. Actually, a more natural representation of the date would be to distinguish between leap year and non-leap years and divide the angle into 366 values only for the former case. However, here we neglected leap years for reasons of convenience.

The first moment of DATE is a two dimensional vector and states the expectations in X and Y . The second mixed moment of DATE is also a two dimensional vector, which defines cross-covariance between (X, Y) and TEMPERATURE. The second moment of DATE is a symmetric matrix, which fixes the within-block covariance of (X, Y) . Tab. 6.8 summarizes all the statistical features from the data which we need for the entropy maximization described in Section 6.3.1 and Section 6.3.2.

Using these constraints we computed the plausible Markov kernels for both hypothetical causal directions. Note that in all plots the variable DATE is parameterized by the integer k . Because

6.8. Experiments with data on continuous domains

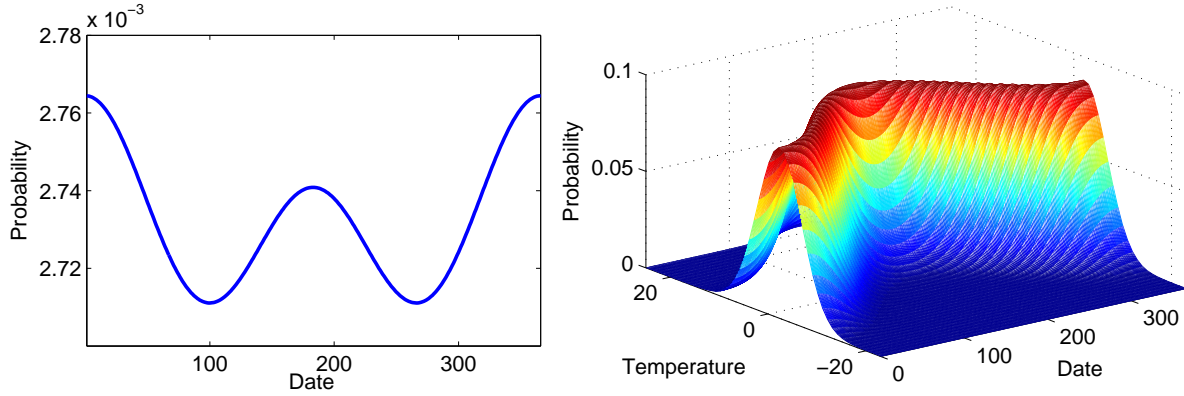


Figure 6.14.: The smoothest Markov kernels $Q(\text{DATE})$ and $Q(\text{TEMPERATURE}|\text{DATE})$ for the hypothetical causal order $\text{DATE} \rightarrow \text{TEMPERATURE}$.

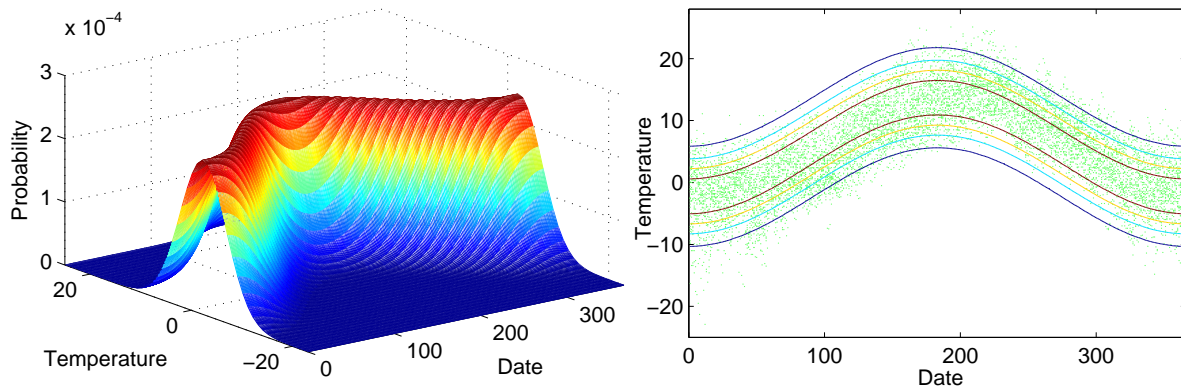


Figure 6.15.: Joint measure Q of the smoothest Markov kernels with respect to the causal order $\text{DATE} \rightarrow \text{TEMPERATURE}$. The plot on the left is a 3D illustration of the probability density; isolines of the density are drawn on the right. The green points on the right indicate the 9162 observed temperature values within a period of 25 years. The density above provides a better fit to the data than the density for the wrong causal direction in Fig. 6.17.

6. Discovering Causal Order by Properties of Conditionals

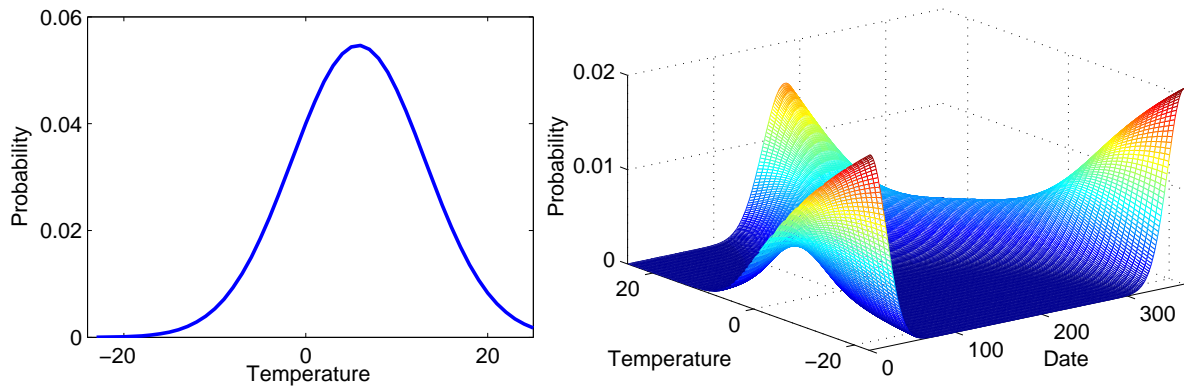


Figure 6.16.: Smoothest Markov kernels $\mathcal{R}(\text{TEMPERATURE})$ and $\mathcal{R}(\text{DATE}|\text{TEMPERATURE})$ for the hypothetical causal order $\text{TEMPERATURE} \rightarrow \text{DATE}$.

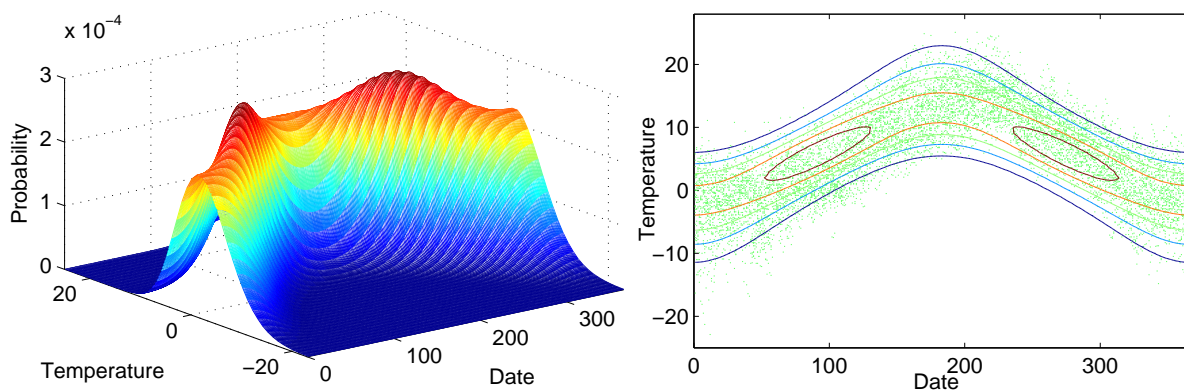


Figure 6.17.: Joint measure \mathcal{R} of the smoothest Markov kernels with respect to the hypothetical causal order $\text{TEMPERATURE} \rightarrow \text{DATE}$. As in Fig. 6.15, the left plot is a 3D plot of the computed density, the right plot shows its isolines and the green points are again the observed temperature values. The elliptic isolines indicate areas of higher probability even though the observed values do not show clustering in these areas. Accordingly, the joint density for the true causal direction in Fig. 6.15 provides a better fit.

of the non-uniform sampling of DATE (there are often only 365 days in the year and in the real dataset there is one year with more observations for the days in January), the plausible Markov kernel of the cause DATE in $\text{DATE} \rightarrow \text{TEMPERATURE}$ differs slightly from the usually expected uniform distribution (Fig. 6.14, left). For the effect variable TEMPERATURE in $\text{DATE} \rightarrow \text{TEMPERATURE}$, the plausible Markov kernel (Fig. 6.14, right) has a conditional expectation in a sinusoidal form, which traces back to the cyclic property of the cause DATE, and a Gaussian-shaped function for every given value of DATE, which is basically due to the fact that the Gaussian distribution maximizes the entropy given its variance.

In the case of the other hypothetical causal direction $\text{TEMPERATURE} \rightarrow \text{DATE}$, the cause variable TEMPERATURE has a Gaussian distribution (Fig. 6.16, left). For the effect variable DATE in $\text{TEMPERATURE} \rightarrow \text{DATE}$, we obtain a strange and non-intuitive shape for its most plausible Markov kernel (Fig. 6.16, right).

Then we calculated the joint distributions from these plausible Markov kernels based on both hypothetical causal directions.

$$\begin{aligned} \mathcal{Q}(\text{DATE}, \text{TEMPERATURE}) &= \mathcal{Q}(\text{TEMPERATURE}|\text{DATE}) \mathcal{Q}(\text{DATE}), \\ \mathcal{R}(\text{DATE}, \text{TEMPERATURE}) &= \mathcal{R}(\text{DATE}|\text{TEMPERATURE}) \mathcal{R}(\text{TEMPERATURE}). \end{aligned}$$

Fig. 6.15 (left) visualizes the resulting joint distribution \mathcal{Q} and Fig. 6.17 (left) visualizes \mathcal{R} . Our computation is based on a discretization of one day for the variable DATE and one degree for the variable TEMPERATURE. Fig. 6.15 (right) and Fig. 6.17 (right) display the isolines of both joint distributions with the observed temperature values. We note that \mathcal{Q} and \mathcal{R} have different numbers of modes and that this qualitative difference between both distributions appeared to be with respect to changes in the discretization.

Our calculation of the log-likelihood scores of the most plausible joint distribution \mathcal{Q} and \mathcal{R} shows that for given data the “true” causal direction $\text{DATE} \rightarrow \text{TEMPERATURE}$ achieves a log-likelihood score of -8.0900×10^4 , whereas the other direction gets a lower log-likelihood score of -8.1031×10^4 . If we run the Scheffé tournament, \mathcal{Q} wins clearly with

$$d_{\text{DATE} \rightarrow \text{TEMPERATURE}} = 0.0156$$

against

$$d_{\text{TEMPERATURE} \rightarrow \text{DATE}} = 0.0780.$$

The joint measure of the true causal direction provides a better fit for the data than that of the wrong causal order. It is worth to mention that we repeated our experiments also with the measured temperatures at 7 am, 2 pm and 6 pm as well as daily maximum and minimum in the place of the daily average temperature to test the causal hypothesis. Our inference rule yielded in all cases the correct causal direction, as desired.

7. Discovering Causal Direction by Complexity Measure of Distributions

Causal inference by means of plausible Markov kernels uses the properties of conditional distributions. The motivation is that statistic dependences between cause and effect which are generated by natural causal mechanism should typically lead to “simple or smooth” expressions for $P(\text{effect}|\text{cause})$ but will not necessarily generate simple expressions for $P(\text{cause}|\text{effect})$. In last chapters, we showed a first attempt for evaluating the smoothness or simplicity of the true measure. How to quantify the smoothness and simplicity of a conditional distribution in a more general framework is our main concern. In this chapter, we propose to measure the complexity of a distribution by a Hilbert space seminorm of the logarithm of the density. The function is an element of an RKHS and its seminorm can therefore be computed by usual kernel methods. In contrast to common machine learning applications, this complexity measure plays not only the role of a regularizer used to avoid overfitting of describing finite data points. It is rather considered as an interesting quantity in its own right since it should provide hints on the causal direction. For this purpose, it is essential to choose a definition of complexity measure which is well-behaved in some respects.

7.1. Defining complexity measure by Hilbert space seminorms

Before we introduce the complexity measure for conditional densities, we define it for unconditional densities. Let us ignore for the moment the sampling issue and assume that the density P_X of some random variable X (probably vectorial) is perfectly known. For the sake of convenience and in order to avoid some technical problems, we assume that the value set \mathcal{X} of X is finite. Now, we introduce a complexity measure on the space of densities on \mathcal{X} as follows.

Definition 23 (Complexity of Marginals) *Let \mathcal{X} be a probability space, X be a random variable on \mathcal{X} , and P_X a density on \mathcal{X} . Furthermore, let \mathcal{H} be a Hilbert space of real-valued functions on \mathcal{X} containing the set of constant functions. Then we define the complexity of P_X as*

$$C(P_X) := \min \left\{ \|\phi\|^2 \mid \phi \in \mathcal{H} \text{ with } P_X(x) = \exp(\phi(x) - \ln z_\phi) \right\}$$

7.1. Defining complexity measure by Hilbert space seminorms

with the partition function $z_\phi := \sum_x \exp(\phi(x))$. Here $\|\cdot\|$ denotes a seminorm on \mathcal{H} given by

$$\|\phi\| := \sqrt{B(\phi, \phi)},$$

where B denotes a positive definite (but not necessarily strictly positive) bilinear form $B : \mathcal{H} \times \mathcal{H} \rightarrow \mathbb{R}$.

In the following, we will use the following terminology: we call two vectors $v, w \in \mathcal{H}$ orthogonal if $B(v, w) = 0$. For a subspace V we define

$$V^\perp := \{w \mid B(w, v) = 0 \ \forall v \in V\}.$$

Since V and V^\perp may have non-trivial intersection, we avoid the term ‘‘orthogonal complement’’. The term ‘‘orthogonal’’ will always refer to the bilinear form B unless something else is explicitly stated. An orthogonal projection R is said to be an projection onto V^\perp if $R\mathcal{H} \subseteq V^\perp$ and $Rw = w - v$ for some $v \in V$ that minimizes $\|w - v\|$. We have

$$C(P) = \|Q(\ln P)\|^2, \tag{7.1}$$

where Q denotes the projection onto $\mathbf{1}^\perp$. This is due to $\|\phi\| = \|Q(\phi - z_\phi \mathbf{1})\| = \|Q(\ln P)\|$. We show the following lemma.

Lemma 1 (Additivity) *Let \mathcal{H}_1 and \mathcal{H}_2 be spaces of functions on \mathcal{X}_1 and \mathcal{X}_2 , respectively. Furthermore, let C_1 and C_2 be complexity measures on the densities on \mathcal{X}_1 and \mathcal{X}_2 , respectively, defined by the corresponding seminorms in \mathcal{H}_1 and \mathcal{H}_2 . Assume that a complexity measure C on the density on \mathcal{X} is based on the seminorm of $\mathcal{H} := \mathcal{H}_1 \otimes \mathcal{H}_2$ that satisfies the embedding property $\|a \otimes \mathbf{1}\| = \|a\| = \|\mathbf{1} \otimes a\|$, where $\mathbf{1}$ denotes the function taking the constant value 1. Then we have the following additivity rule: Let P be defined by a product of densities P_1 and P_2 , i.e., $P(x_1, x_2) = P_1(x_1)P_2(x_2)$ for all x_1 and x_2 . Then the complexity of the product measure satisfies $C(P) = C_1(P_1) + C_2(P_2)$.*

Proof Let Q, Q_1, Q_2 denote the projections onto the space of functions orthogonal to $\mathbf{1}$ for the spaces $\mathcal{H}, \mathcal{H}_1, \mathcal{H}_2$, respectively. Then we have

$$\begin{aligned} \|Q(\ln P_1 \otimes \mathbf{1} + \mathbf{1} \otimes \ln P_2)\|^2 &= \|Q_1(\ln P_1) \otimes \mathbf{1} + \mathbf{1} \otimes Q_2(\ln P_2)\|^2 \\ &= \|Q_1(\ln P_1)\|^2 + \|Q_2(\ln P_2)\|^2, \end{aligned}$$

where the last equality is due to Pythagoras’ theorem after taking into account that the vectors $Q_1(\ln P_1) \otimes \mathbf{1}$ and $\mathbf{1} \otimes Q_2(\ln P_2)$ are mutually orthogonal. \blacksquare

Now we move to the definition of the complexity of conditional probabilities:

Definition 24 (Complexity of Conditionals) *Let \mathcal{X} and \mathcal{Y} be the respective value sets of random variables X and Y , and $P_{X,Y}$ be a joint density on $\mathcal{X} \times \mathcal{Y}$. Let $P_{Y|X}$ be the corresponding conditional density. We define the complexity of $P_{Y|X}$ as*

$$C(P_{Y|X}) := \min \left\{ \|\phi\|^2 \mid \phi \in \mathcal{H} \text{ with } P_{Y|X}(y|x) = \exp(\phi(x, y) - \ln z_\phi(x)) \right\}$$

7. Discovering Causal Direction by Complexity Measure of Distributions

with the partition function $z_\phi(x) := \sum_y \exp(\phi(x, y))$.

Similarly to the reformulation of Definition 23 in Eq. (7.1), the definition of the complexity of a conditional density can also be given in a more explicit form:

$$C(P_{Y|X}) = \|(\mathbf{id} \otimes Q_2)(\ln P_{Y|X})\|^2, \quad (7.2)$$

where “ \mathbf{id} ” denotes the identity map and Q_2 is as in the proof of Lemma 1. Under the assumptions of Definition 24, we have:

Lemma 2 (Consistency) *Let X and Y be stochastically independent with respect to the joint density P , i.e., $P_{Y|X} = P_Y$. Let C be a complexity measure based on a seminorm in $\mathcal{H} = \mathcal{H}_X \otimes \mathcal{H}_Y$ satisfying the embedding property in Lemma 1. Then we have $C(P_{Y|X}) = C_2(P_Y)$.*

Proof Let ϕ be some function on $\mathcal{X} \times \mathcal{Y}$ such that $P_{Y|X}(y|x) = \exp(\phi(x, y) - \ln z_\phi(x)) = P_Y(y)$. We choose an arbitrary value y_0 and set $f(x) := \phi(x, y_0) - \ln P_Y(y_0)$ and $g(y) := \ln P_Y(y)$. Then we have $\phi(x, y) = f(x) + g(y)$. Thus

$$\|(\mathbf{id} \otimes Q_2)(\phi)\|^2 = \|(\mathbf{id} \otimes Q_2)(f \otimes \mathbf{1} + \mathbf{1} \otimes g)\|^2 = \|Q_2(g)\|^2.$$

Therefore, we conclude $C(P_{Y|X}) = C_2(P_Y)$. ■

Lemma 2 is essential, if one intends to compare the complexity of marginal densities to that of conditional densities. The following causal inference principle stands behind such a comparison: having factorized a joint density $P_{X,Y}$ into $P_{Y|X}P_X$ and $P_{X|Y}P_Y$ based on both possible hypothetical causal orders, one calculates the sums of the complexities $C(P_{Y|X}) + C(P_X)$ and $C(P_{X|Y}) + C(P_Y)$ with respect to the different hypotheses. The intention is to consider the sums as the “total complexity” of the causal models $X \rightarrow Y$ and $X \leftarrow Y$ respectively and to prefer the causal direction that corresponds to the smaller total complexity. For doing so, it is crucial to make $C(P_Y)$ and $C(P_{Y|X})$ comparable. An essential property of the complexity measure is that we have

$$C(P_{Y|X}) + C(P_X) \neq C(P_{X|Y}) + C(P_Y)$$

in the generic case. The following lemma provides some deeper understanding why this is the case.

Lemma 3 (Relation to Complexity of Partition Function) *Under the assumptions of Definition 24, the following inequalities hold:*

$$\begin{aligned} C(P_{X,Y}) &\geq C(P_{Y|X}) + C(P_X) + C(R) - 2\sqrt{C(P_X)C(R)}, \\ C(P_{X,Y}) &\leq C(P_{Y|X}) + C(P_X) + C(R) + 2\sqrt{C(P_X)C(R)}, \end{aligned}$$

where R is the following measure on X : Set $R(x) := c \cdot z_f(x)$ with an appropriate normalization factor c and the partition function $z_f(x) = \sum_y \exp(f(x, y))$ which is derived from $f := (\mathbf{id} \otimes Q_2)(\ln P_{Y|X})$.

Proof Write

$$P(y|x) = \exp(f(x, y) - \ln z_f(x))$$

7.1. Defining complexity measure by Hilbert space seminorms

where f satisfies by definition $(\text{id} \otimes Q_2)(f) = f$. Furthermore, we set

$$P(x) = \exp(g(x) - \ln z)$$

with $Q_1(g) = g$ and normalization constant z . We observe that f is orthogonal to all functions that depend only on x since the latter have the form $h \otimes \mathbf{1}$ (where h is an arbitrary function). We have

$$\ln P_{X,Y} = \ln P_X + \ln P_{Y|X} = (-\ln z_f + g) \otimes \mathbf{1} + f - \ln z.$$

Due to the above remarks we have $f \perp (-\ln z_f + g) \otimes \mathbf{1}$. To compute the complexity of $P_{X,Y}$, we observe

$$\begin{aligned} C(P_{X,Y}) &= \|Q(f + (-\ln z_f + g) \otimes \mathbf{1} + \ln z (\mathbf{1} \otimes \mathbf{1}))\|^2 \\ &= \|f + Q_1(-\ln z_f + g) \otimes \mathbf{1}\|^2. \end{aligned}$$

Since the projected term is still orthogonal to f (note that it is a function that depends only on x) we have

$$C(P_{X,Y}) = \|f\|^2 + \|Q_1(-\ln z_f + g)\|^2 = \|f\|^2 + \|Q_1(\ln z_f) + g\|^2. \quad (7.3)$$

By elementary geometry we obtain

$$\begin{aligned} \|Q_1(-\ln z_f) + g\|^2 &\geq \|Q_1(\ln z_f)\|^2 + \|g\|^2 - 2\|Q_1(\ln z_f)\| \|g\|, \\ \|Q_1(-\ln z_f) + g\|^2 &\leq \|Q_1(\ln z_f)\|^2 + \|g\|^2 + 2\|Q_1(\ln z_f)\| \|g\|. \end{aligned}$$

Having $C(R) = \|Q_1(\ln z_f)\|^2$, we finally conclude

$$\begin{aligned} C(P_{X,Y}) &\geq C(P_{Y|X}) + C(P_X) + C(R) - 2\sqrt{C(P_X)C(R)}, \\ C(P_{X,Y}) &\leq C(P_{Y|X}) + C(P_X) + C(R) + 2\sqrt{C(P_X)C(R)}. \end{aligned}$$

■

Note that in high dimensional spaces the angle between two vectors is typically close to 90 degree. Therefore, it is likely that the vectors $Q_1(\ln z_f)$ and g in Eq. (7.3) satisfy $B(\ln z_f, g) \approx 0$. We then have

$$C(P_{X,Y}) \approx C(P_{Y|X}) + C(P_X) + C(R).$$

In other words, the complexity of the joint density is typically the sum of the complexities of the conditional densities and the complexity of a measure defined by the partition function. The basic idea behind our inference rule is that simple causal mechanism may generate conditional densities $P_{Y|X}$ which are simple up to a rather complex X -dependent normalization constant, i.e., the partition function. Note that the joint density could be complex even when P_X is simple due to the additional complexity of the partition function.

7.2. Calculation of seminorm using kernel methods

We have shifted the problem of defining the complexity of densities into the definition of seminorms. We will rewrite our definition such that seminorms can be calculated in an implicit way. With the so-called “kernel trick” different seminorms can be chosen by simply replacing the kernel (see [137, 22]).

Let $k_1, k_2: (\mathcal{X} \times \mathcal{Y}) \times (\mathcal{X} \times \mathcal{Y}) \rightarrow \mathbb{R}$ be positive definite symmetric kernels and $\mathcal{X} \times \mathcal{Y}$ the probability space under consideration. Let \mathcal{H}_j for $j = 1, 2$ be the Hilbert spaces given by the completion of the spans of the functions $k_j((x, y), \cdot)$ with the inner product

$$\langle k_j((x, y), \cdot), k_j((x', y'), \cdot) \rangle = k_j((x, y), (x', y')). \quad (7.4)$$

Hilbert spaces defined this way are usually referred to as RKHSs. We assume that \mathcal{H}_2 is a subspace of \mathcal{H}_1 . The vector ϕ in Definition 23 and Definition 24 can be approximated by

$$\phi(x, y) := \sum_{j=1}^n c_j k((x_j, y_j), (x, y)) = \left\langle \sum_{j=1}^n c_j k((x_j, y_j), \cdot), k((x, y), \cdot) \right\rangle \quad (7.5)$$

with appropriate coefficients c_j and points (x_j, y_j) .

We define our seminorm by

$$\|\phi\| := \|R(\phi)\|_{\mathcal{H}_1},$$

where R is the projector onto the subspace orthogonal to \mathcal{H}_2 with respect to the inner product in \mathcal{H}_1 . The idea of using such a seminorm is that the space \mathcal{H}_2 contains simple functions (for instance polynomials of low degree) that should not contribute to the complexity measure at all. This corresponds to the use of conditionally positive definite kernels in semiparametric models [147, 171]. Let $P_{Y|X}$ be a conditional density, given by

$$\begin{aligned} P_{Y|X}(y|x) &= \exp \left(\sum_{j=1}^n c_j^{(1)} k_1((x_j, y_j), (x, y)) \right. \\ &\quad \left. + \sum_{j=1}^n c_j^{(2)} k_2((x_j, y_j), (x, y)) - \ln z_{\mathbf{c}}(x) \right) \end{aligned} \quad (7.6)$$

with the appropriate partition function $z_{\mathbf{c}}(x)$. The complexity $C(P_{Y|X})$ is then defined by the minimum of $\sum_{j,j'=1}^n c_j^{(1)} c_{j'}^{(1)} k_1((x_j, y_j), (x_{j'}, y_{j'}))$, i.e., the square of the norm of the shortest component in \mathcal{H}_1 , see Eq. (7.4), over all vectors $\mathbf{c} := (c_1^{(1)}, \dots, c_n^{(1)}, c_1^{(2)}, \dots, c_n^{(2)}) \in \mathbb{R}^{2n}$ for which Eq. (7.6) holds. The vector with coefficients $k_1((x_j, y_j), (x, y))$ and $k_2((x_j, y_j), (x, y))$ with $j = 1, \dots, n$ can be interpreted as the vector of sufficient statistics of an exponential model.

The framework introduced can also be considered as a method of density estimation with kernel methods. To make this method tractable in practice, there are some issues of implementation to be addressed. The choice of kernels k_1 and k_2 will be discussed in the next section. Given k_1 and k_2 described in the next section, our remarks above specified the choice of points (x_j, y_j) for

7.3. Estimating densities from finite data with kernels

$j = 1, \dots, n$ in the range. Our experiments show that the seminorm is not sensitive against the choice of n , if n is not too small and the points (x_j, y_j) are somewhat evenly distributed over the whole range. The results of all our experiments in this chapter are based on the choice of $n = 7$ for unconditional (one-dimensional) cases and $n = 49$ for conditional (two-dimensional) cases. The 7 points for each dimension are chosen equidistantly in percentile over the whole observed range. For a binary case, $n = 2$.

To ensure that the embedding property $\|a \otimes \mathbf{1}\| = \|a\| = \|\mathbf{1} \otimes a\|$ is satisfied we proceed as follows. We choose the kernel k_1 as the product

$$k_1((x_j, y_j), (x_{j'}, y_{j'})) = k_X^{(1)}(x_j, x_{j'}) k_Y^{(2)}(y_j, y_{j'}).$$

Thus, the corresponding RKHSs have the form $\mathcal{H}_2 := \mathcal{H}_2^X \otimes \mathcal{H}_2^Y$ and $\mathcal{H}_1 := \mathcal{H}_1^X \otimes \mathcal{H}_1^Y$. We choose the kernels $k_X^{(2)}$ and $k_Y^{(2)}$ and the domains \mathcal{X} and \mathcal{Y} such that \mathcal{H}_2^X and \mathcal{H}_2^Y contain the constant functions and normalize $k_X^{(1)}$ and $k_Y^{(1)}$ such that the constant functions $\mathbf{1}$ on \mathcal{X} and \mathcal{Y} satisfy $\|\mathbf{1}\|_{\mathcal{H}_1^X} = 1$ and $\|\mathbf{1}\|_{\mathcal{H}_1^Y} = 1$, respectively.

To this end, we define the matrix $K_X := k_X^{(1)}(x_j, x_{j'})$ and calculate its inverse K_X^{-1} . Let $c := (K_X^{-1})\mathbf{1}$ be the vector of coefficients of the constant function $\mathbf{1}$. This yields the normalization condition $\langle c | K_X c \rangle = 1$, i.e., the sum of all entries of K_X^{-1} are 1. The same procedure is also applied to $k_Y^{(1)}$. The seminorm of $a \otimes \mathbf{1}$ is given by the Hilbert space norm of its component in $(\mathcal{H}_2^X \otimes \mathcal{H}_2^Y)^\perp$. Let R_X and R_Y be the orthogonal projections onto $(\mathcal{H}_2^X)^\perp$ and $(\mathcal{H}_2^Y)^\perp$, respectively. Due to $R_Y(\mathbf{1}) = 0$ the relevant component of $a \otimes \mathbf{1}$ is given by $R_X(a) \otimes \mathbf{1}$. The Hilbert space norm of this function is given by $\|R_X(a)\|_{\mathcal{H}_1^X}$ which coincides with the seminorm of a . Similar arguments apply to $\mathbf{1} \otimes a$.

7.3. Estimating densities from finite data with kernels

To calculate the complexity of a density we first use regularized maximum likelihood estimation to fit the observed data points using exponential models. A general framework for applying the kernel approach to exponential families can be found in [26]. Without regularizer, the method works as follows. Introducing the map $\psi: \mathcal{X} \times \mathcal{Y} \rightarrow \mathcal{H}_1$ with

$$\psi(x, y) := k_1((\cdot, \cdot), (x, y))$$

we define the family of conditional densities $P_\phi(y|x) = \exp(\langle \phi | \psi(x, y) \rangle - \ln z_\phi(x))$. For N observed data points (x_i, y_i) , the maximum likelihood estimation selects ϕ by

$$\max_{\phi \in \mathcal{H}_1} \left\{ \frac{1}{N} \sum_{i=1}^N (\langle \phi | \psi(x_i, y_i) \rangle - \ln z_\phi(x_i)) \right\}. \quad (7.7)$$

7. Discovering Causal Direction by Complexity Measure of Distributions

In order to avoid overfitting we include a regularizer and then obtain the expression

$$\max_{\phi \in \mathcal{H}_1} \left\{ \frac{1}{N} \sum_{i=1}^N (\langle \phi | \psi(x_i, y_i) \rangle - \ln z_\phi(x_i)) - \epsilon \|\phi\| \right\}. \quad (7.8)$$

The regularizer, the norm itself and not its square (as opposed to our complexity measure), is in agreement with the choice in [6]. The authors of [6] propose to use a value of ϵ that is proportional to $1/\sqrt{N}$. In our experiments, we chose $\epsilon = 1/\sqrt{N}$. Note, as an aside, that the regularized maximum likelihood estimation for unconditional densities can also be interpreted as maximizing the entropy of the density subject to the expectations of $\psi(X, Y)$ coinciding with the observed means of $\psi(X, Y)$ up to an error of ϵ (see [6]).

For the sake of numerical stability, we normalize the observed data for X, Y respectively. The data are linearly transformed such that the points ± 1 of the normalized data have the same percentiles as ± 3 of a standard normal distribution, respectively. Thus the normalized data points with continuous range will be located mostly in the interval $[-1, 1]$. A normalized binary variable then takes values ± 1 . We choose a discretization of 0.1 to count the relative frequencies and calculate the sum in optimization. For the experiments described in the next section we use a sum of the Gaussian kernel

$$k_\sigma((x, y), (x', y')) = \exp\left(-\frac{\|(x, y) - (x', y')\|^2}{2\sigma^2}\right)$$

to define the space \mathcal{H}_1 and a polynomial kernel

$$k_{a,b,\tilde{a},\tilde{b}}((x, y), (x', y')) = \left(\frac{\langle x \cdot x' \rangle}{a} + b\right) \left(\frac{\langle y \cdot y' \rangle}{\tilde{a}} + \tilde{b}\right)^2,$$

to define \mathcal{H}_2 . The additional scaling parameters $a, b, \tilde{a}, \tilde{b}$ are used to ensure a numerically stable training. We choose $a, b, \tilde{a}, \tilde{b}$ so that the entries of $k_{a,b,\tilde{a},\tilde{b}}$ take the value between $[-1, 1]$. Since the normalized data have the value mostly between -1 and 1 , we choose $a = \tilde{a} = 2$ and $b = \tilde{b} = \frac{1}{2}$, if x, y are one-dimensional. The formulation of both kernels for the unconditional case is straightforward. Assuming that the range of random variables is compact, the space \mathcal{H}_2 (induced by a Gaussian kernel) contains the space \mathcal{H}_1 (induced by a polynomial kernel).

The idea behind the choice of kernels is the following: if x and y are one-dimensional, the second kernel induces a space of functions spanned by the monomials $1, x, xy, xy^2, y, y^2$. We consider these as sufficiently smooth such that they should not contribute to the complexity measure. In particular, we can then obtain Gaussian distributions whose expectations and variance changes linearly with the given variable X . The Gaussian kernel and the polynomial kernel induces, on the one hand, enough flexibility to fit various global and local structure of density. On the other hand, the density estimated this way is smooth. For a discussion of smoothing properties of Gaussian and polynomial kernels we refer to [115, 147].

Our experience suggest that we have to learn appropriate values σ for the Gaussian kernel by optimizing Eq. (7.8), otherwise we could not obtain reasonable fits. Clearly, we cannot directly compare the complexity values corresponding to kernels with different values for σ . However,

we may define the complexity by the minimum over all seminorms squared within some given family of RKHSs. Denoting by \mathcal{H}_i the Hilbert space given by the kernel k_i we may define $C(P)$ by $C(P) := \inf_{i \in I} \{C_i(P)\}$, where C_i refers to the complexity measure defined by the seminorm in \mathcal{H}_i . In order to ensure additivity with respect to product measures in product spaces for the redefined C we need to define a family of spaces by $\mathcal{H}_i^{(1)} \otimes \mathcal{H}_j^{(2)}$ and optimize over all pairs (i, j) . Due to a combinatorial explosion such an optimization will only be feasible for a small set I and few tensor components. In the experiments described in the next section we have therefore used the same σ for the Hilbert spaces for X and Y .

If we run the optimization procedure in Eq. (7.8) over all Hilbert spaces (i.e., all reasonable values σ) the procedure will choose the vector ϕ from the Hilbert space that leads to the smallest norm among all those that yield the same value in the non-regularized optimization given by Eq. (7.7). We shall therefore consider the optimum of Eq. (7.8) over all kernels taken from a given family as an estimation of the minimal norm of the density over all Hilbert spaces under consideration. Since the optimization problem with σ is no longer convex, one should choose the start value of σ properly. In our experiments we chose 200 equidistant starting values in the range $(0, \frac{2}{3})$. The value which leads to the maximum of Eq. (7.8) will then be taken as the start value of a subsequent optimization via gradient descent.

7.4. Experiments with simulated and real-world data

Some simulated experiments show the intuitive meaning of our complexity measure, while the real-world examples show that this complexity measure could be helpful for inferring the causal direction between two variables.

7.4.1. Unconditional densities

We first sampled 1000 data points from various unconditional distributions as shown in Fig. 7.1. The underlying density P_1 follows a standard normal distribution; P_2, P_3, P_4, P_5 are various mixtures of 2 Gaussians; P_6, P_7, P_8 are mixtures of 3, 4, 5 Gaussians respectively. P_9 is a mixture of a Gaussian and a gamma distribution. P_{10} follows a single gamma distribution and P_{11}, P_{12} are mixtures of 2, 3 gamma distributions respectively. As expected, we see that the complexity of a single Gaussian is 0. A single gamma distribution has a very small complexity value. The measure increases as the number of components increases. This holds even for the unimodal mixture P_2, P_{11}, P_{12} .

Moreover, we examined the smoothness (complexity) of a real-world temperature dataset (Daily average temperatures from 1979 through 2004, Furtwangen, Germany) with 9162 entries. The estimated density (see Fig. 7.2) has a complexity of 0.0265, which suggests that the density of temperatures is more complex than a single normal or gamma distribution. We observe slightly larger complexity values for a gamma distribution than for a Gaussian. We leave the question open whether this property is desirable.

7. Discovering Causal Direction by Complexity Measure of Distributions

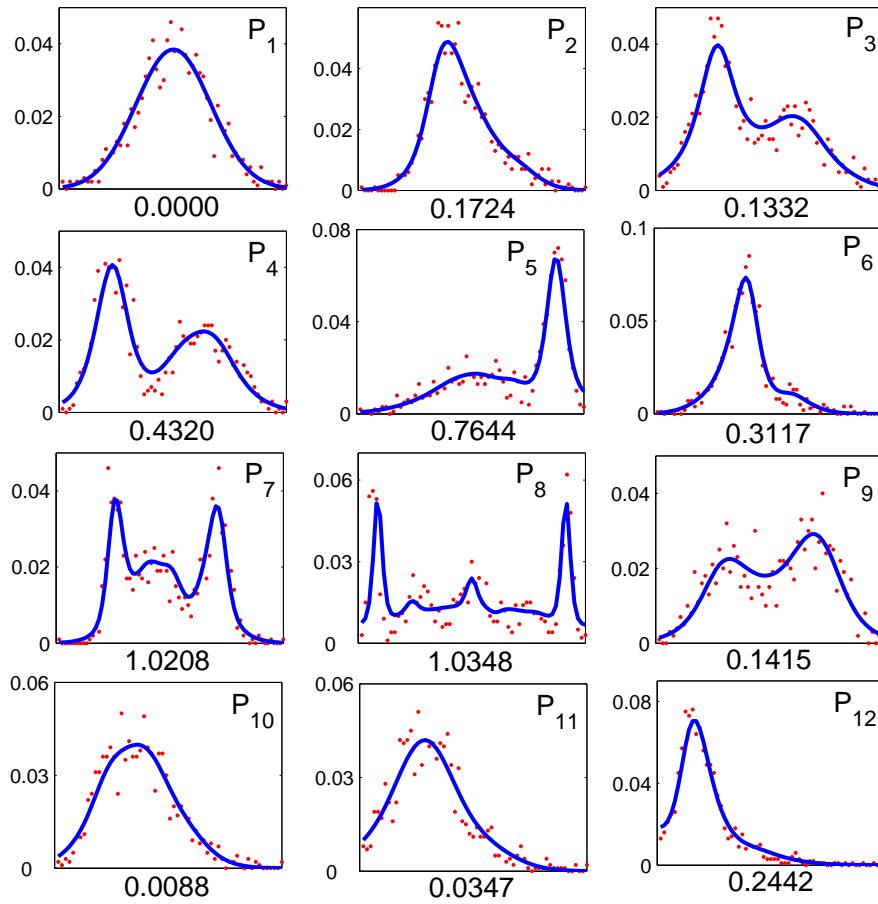


Figure 7.1.: 12 toy data sets sampled by distributions P_1, \dots, P_{12} (see text). The dots indicate the observed relative frequencies, the solid lines the estimated densities. The calculated complexity values are shown below each plot.

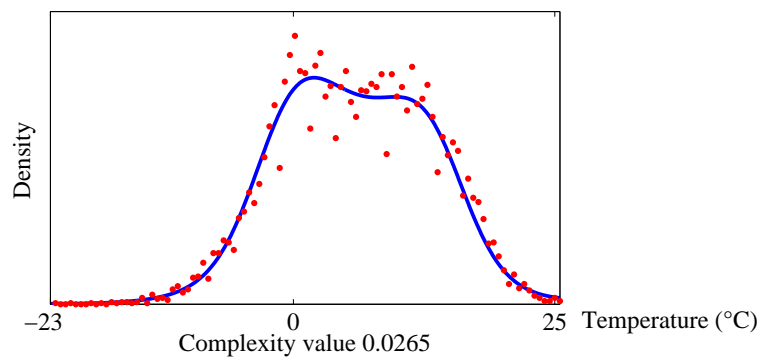


Figure 7.2.: Daily average temperatures from 1979 through 2004, Furtwangen, Germany. The dots indicate the observed relative frequencies, the solid lines the estimated density. The calculated complexity value is shown below the plot.

7.4. Experiments with simulated and real-world data

	P_2	P_3	P_4	P_9
$P_i(Y X = x_1)$	$\mathcal{N}(-1, 1)$	$\mathcal{N}(-2, 1)$	$\mathcal{N}(-3, 1)$	$\mathcal{N}(8, 1)$
$P_i(Y X = x_2)$	$\mathcal{N}(1, 4)$	$\mathcal{N}(2, 4)$	$\mathcal{N}(3, 4)$	$\mathcal{G}(9, 0.5)$
$C(P_X)$	0.0000	0.0000	0.0000	0.0000
$C(P_{Y X})$	0.0000	0.0000	0.0000	0.0004
$C(P_Y)$	0.1724	0.1332	0.4320	0.1415
$C(P_{X Y})$	0.0234	0.0000	0.0000	0.0000

Table 7.1.: Complexity of conditional densities in binary mixture models.

7.4.2. Conditional densities

The main motivation behind this complexity measure is to develop a tool for causal inference based on observed data by quantifying smoothness of conditional distributions. Intuitively, having observed a bimodal distribution after large sampling, one would prefer to interpret the observation as a mixture of two populations. It is rather implausible to assume that a probability density with such a shape should stem from a homogeneous statistical ensemble. There is a broad variety of applications where the detection of mixtures is crucial for data analysis (see e.g. [43, 54, 112]).

If we define a density on a binary variable X and a continuous variable Y by

$$P(y) = 0.5 P(y | X = x_1) + 0.5 P(y | X = x_2),$$

where both conditionals $P(y|X = x_i)$ are Gaussian, the total complexity of the model $X \rightarrow Y$ is zero since the kernel k_2 induces such a density. Note that due to our choice of kernel the complexity of the density of a binary variable is always 0. We checked on randomly generated data with 1000 points whether this result is also obtained in finite sampling. We furthermore confirmed that the model $X \rightarrow Y$ was also preferred when the conditional $P(Y|X = x_2)$ was the gamma distribution and $P(Y|X = x_1)$ was a Gaussian. In a similar way, we defined joint densities on X and Y corresponding to the mixture models P_2, P_3, P_4, P_9 in Fig. 7.1 by using a binary variable X to indicate which one of the two pure ensembles is taken. The complexity values in Tab. 7.1 show that we indeed obtained the expected results.

Since the causal inference problem was the motivation for the construction of our complexity measure, its performance with respect to some real-world data is the best criterion for judging whether it seems appropriate or not. We performed experiments with datasets from the Current Population Survey (CPS) 2001 (see Section 6.6.2 for data) on the relation between sex (binary variable) and income (continuous variable) in the US. Statistical methods show that income and gender are indeed correlated. Common sense tells us that we can exclude that the personal income influences the gender, whereas the reverse causal direction makes sense. We found that the density of the income marginalized over both genders is more complex than the density for both genders separately.

First we intended to check to what extent the complexity measure recognizes mixtures as more

7. Discovering Causal Direction by Complexity Measure of Distributions

complex. We found

$$C(P_{\text{Income}|\text{Sex}=\text{"male"}}) < C(P_{\text{Income}}),$$

and the same for $P_{\text{Income}|\text{Sex}=\text{"female"}}$. Note that left side of the inequality can also be considered as the complexity of an unconditional density since we assigned a specific value to the conditioning variable.

However, to check the performance of our causal inference principle we have to compute the total complexity of both hypothetical causal directions. Using one subsample of 10% of the data points from 13,803 entries, we found the following complexity values:

$$C(P_{\text{Sex}}) = 0.0000, C(P_{\text{Income}|\text{Sex}}) = 0.4632,$$

and

$$C(P_{\text{Income}}) = 0.6725, C(P_{\text{Sex}|\text{Income}}) = 0.0000,$$

i.e., the sum of the first two values (corresponding to the true causal direction) is indeed smaller than the sum of the last two.

Using the same dataset, we consider another example where a continuous variable causally influences a binary variable. We examine the continuous variable “Age” and the binary variable marriage status (short “M-Status”, it takes the two values: “never married” or “married, widowed, divorced or separated”). A 10% subsample leads to the following results:

$$C(P_{\text{Age}}) = 0.0023, C(P_{\text{M-Status}|\text{Age}}) = 0.0012,$$

and

$$C(P_{\text{M-Status}}) = 0.0000, \quad \text{and} \quad C(P_{\text{Age}|\text{M-Status}}) = 0.0164.$$

The sum of the first two values (corresponding to the true causal direction) is smaller than the sum of the last two. Our causal inference rule would then favor the causal hypothesis that the age should be a cause of marriage status of a person, not vice versa.

We repeated these experiments using different subsamples of 10% of the whole dataset. All subsamples yielded the same result with regarding to both causal hypotheses. However, the complexity values were slightly different for different samples. Therefore, we should not overrate the meaning of the absolute value of the complexity measure. Its relevance consists rather in allowing us to compare complexity values for different causal directions.

The third example that we tested is a data set of handwritten numerals [119] containing PCA components of the pixel vectors for the symbols “0”-“9”. We considered the symbols “0” and “1” and interpreted them as the values of a binary random variable X . For each symbol there are 200 instances. We chose a PCA coefficient as a continuous random variable Y . We assume that X is the cause of Y because the person first had the intention to write the digit “1” or “0” and wrote it afterward. Hence the PCA coefficient Y is the effect.

We applied our inference rule to several coefficients. Their correlation with X attained, among others, the values $\rho = 0.8661, -0.8079, 0.3233, 0.5674, 0.1086, -0.0601, -0.2547$. For the cases with strong correlations we obtained results that were consistent with the ground truth, i.e., $C(P_X) + C(P_{Y|X}) < C(P_Y) + C(P_{X|Y})$. When the correlation coefficient was 0.3 or smaller, we

7.4. Experiments with simulated and real-world data

have also observed several failures of the causal inference rule since $C(P_Y)$ and $C(P_{Y|X})$ are extremely small for these cases. This is because a density is hard to recognize as a mixture of two distributions if they are not sufficiently different.

In summary, experiments with real-world and simulated data show that mixtures of two simple distributions like Gaussians and gamma distributions are recognized as more complex than the corresponding conditional probability given the binary variable that labels the mixed components. Moreover, the complexities of conditionals that correspond to the true causal direction were in major cases of our limited examples smaller than the complexity of the wrong causal direction. Note that the information of causal directions can be helpful for e.g., feature selection [81]. It should be stressed that the intuitive relevance of the absolute value of complexity should not be overrated.

8. Summary and Outlook

This thesis coped with the problem of learning causality and proposed two different approaches that can be served as a basis (independent or combined together) for automatically building intelligent systems for reasoning under uncertainty. The potential application of such influence diagrams (causal structures) is that they can be used to plan or design a strategy of future interventions or manipulations.

In the spirit of the conventional constraint-based IC algorithm, we first proposed two kernel-based versions of collider identification to learn causal structure. The so-called RCL algorithm is based on kernel hypothesis tests of independence, while the so-called KCL algorithm additionally takes the magnitude of dependences into account. In RCL, we prefer constraints with small conditioning sets and explicitly treat the violations of transitivity and intersection properties of a faithful Bayesian network. RCL is particularly suitable for learning sparse networks. In KCL, we restrict the number of potential conditioning sets for the independence test by learning an auxiliary graph via kernel dependence measures. In our approach, Type II error of hypothesis tests and its impact on learning the adjacency structure can be kept at a low level. The impact of the potential type I error on learning causal directions is alleviated by using the magnitude of dependences measured by kernels. RCL and KCL takes nonlinear relationships into account and refines the IC algorithm in a computationally tractable way and provides unifying methods for learning causal structure over different kinds of (even hybrid) domains. Various experiments showed that our methods are reliable in case of small sample sizes.

In association with this work, several open problems have been suggested for further research. First, regarding the issue of measuring dependence, Fukumizu et al. [63] recently defined the kernel dependence measure with other normalization which makes the measure asymptotically independent of the choice of kernels. It is an intriguing direction to explore the possibility of improving the performance of structural learning via this measure, although there are still some numerical problems in the implementation. Moreover, mutual information is the most popular dependence measure that is able to capture nonlinear relationships. Thus, it is natural to ask the question how it relates to the kernel measure. A first result proved by Gretton et al. [75] showed that the HS-norm \mathbb{H}_{YX} approximates the mutual information $\mathbb{I}(X, Y)$ to first order near independence. A recent paper of Fukumizu et al. [63] gave some insights into the connection between mutual information and his normalized kernel measure. In spite of these works, a general relation between kernel measure and mutual information is not established yet.

In respect of statistical tests, a more efficient test statistics, rather than generating null distribution by random permutations, could be useful. In this direction, Gretton et al. [74] recently made a first attempt for the tests in unconditional cases. An efficient and reliable test statistics for conditional cases would be desirable.

One of the main challenges of constraint-based approaches is to represent independence rela-

tions verified by some statistical test in a simple and robust way, e.g., a faithful Bayesian network. In order to make such faithful representation possible, we proposed in RCL a strategy of variable clustering to handle conflicting information among these relations. Various real-world experiments showed that an appropriate clustering of variables is helpful for constructing a causally meaningful structure. It raises an interesting question how to evaluate the clustering from the point of view of causal relationships. Silva et al. [145] did some work in a linear framework in this direction. Our constraint-based clustering algorithm, which is causal-structure-oriented, provides an entirely different approach to deal with the problem of causal clustering. Nonetheless, the causal clustering is still a not well-studied problem, although clustering itself is very active research field in machine learning.

As seen from various experiments with simulated data, constraint-based approaches outperformed currently popular score-based Bayesian approaches in many of our examples. Nevertheless, a Bayesian approach in principle has some advantages that our RCL or KCL does not have. For instance, a Bayesian approach can straightforwardly incorporate prior knowledge. A Bayesian approach can be efficiently implemented and is well scalable with respect to the sample size. A related work of detecting collider candidates via a Bayesian scoring function is done by Steck [157]. Bayesian scoring function can also be used to detect independence [107, 109]. In addition, a Bayesian approach can in principle be applied to searching over a number of different latent variable models within Markov equivalent classes [86]. Therefore, modifying RCL/KCL to incorporate prior knowledge (combining with experimental data or using temporal information of time series) is a useful direction of further research. More reliable identification of colliders via a well-justified scoring function based on kernel dependence measure would be interesting. A more ambitious goal is to explore the theoretical possibility for discovering latent variables or even learning ancestral graphs [133] in a kernel-based way.

Another practical issue of further work is to make RCL/KCL efficient for learning on a huge network or from a huge dataset. Some techniques of estimating HS norm by randomly selecting or sampling a subset of Gram matrix entries [1, 50] to measure independence in a huge dataset are discussed by Jugelka et al. [94]. Unfortunately, it sometimes provides unsatisfactory results, particularly in the case of close-to-independence (see [94] for experiments).

Beyond independence, the last two chapters of this thesis dealt with the problem of causal inference when there are no independence relations are detected, i.e., a fully connected adjacency structure. In particular, if only two dependent variables are measured, approaches based on independence relations or dependence measures will fail. Our model-based approach assumes that the conditionals that are consistent with the correct causal order should be of a smooth shape or simple, since such conditionals describe indeed the present natural causal mechanism. The so-called plausible Markov kernel assumption.

Our first attempt to capture the plausibility of conditional distributions is to introduce the smoothest Markov kernel by maximizing the conditional entropy subject to the observed first and second moments. The intuition behind this attempt is to define a simple cause-effect interaction that is “as linear as possible”. The “most linear” effect on the binary variable is therefore to generate the desired correlation such that the conditional distribution has maximal uncertainty. This way, we captured the potential “simplest” influence among the variables considered by the smoothest Markov kernels.

8. Summary and Outlook

Experiments with simulated and real-world data indicated that the so-called pMK algorithm can provide useful hints on the causal direction without using independence relations and dependence measures. In other words, pMK provided a tool to select causal hypotheses which are Markov equivalent and thus indistinguishable by a constraint-based approach. Unfortunately, pMK is only computationally efficient for domains of small cardinality, because the method took the domain information directly into account. For these reasons, we proposed to use a constraint-based approach, e.g., PC, for preselecting hypothetical causal graphs and then apply pMK to small subsets of variables for directing remaining undirected edges.

The essential shortcoming of the concept of the smoothest Markov kernel is the fact that such this simplicity criterion for conditional distributions, in general, depends on the representation of data (expecting for binary domains). To represent, for example, dates of the year on a unit circle requires an intuitive understanding about what representation of the data is natural. Likewise, the shape of the distribution of a real-valued variable could be changed by using logarithmic scaling of the data. An important pre-decision is the choice of the most “natural” scaling. Our hope is that for a large causal network criteria on the simplicity of Markov kernels could be developed that are not too sensitive under such rescaling operations as long as they are in some sense not too unreasonable.

Nevertheless, this notion of simplicity that uses entropy maximization subject to the two moments should rather be considered as a first attempt instead of the right one. In order to establish a more general framework to construct complexity measures for conditional probabilities, we proposed a kernel method to estimate the complexity of distributions from finite sampling. The complexity measure is based on an RKHS seminorm of logarithm of the distribution. Since the optimization of Eq. (7.8) requires calculating the partition function, the method presented is computationally rather expensive. Evaluating conditionals with general continuous domains or with more than two random variables seems (from the current perspective) to be feasible only after a coarse discretization. In spite of this shortcoming, experiments showed this complexity measure could provide hints on the causal direction between only two variables where a constraint-based approach fails. Moreover, kernel methods seem to be quite flexible for designing better complexity measures for further research.

In summary, this thesis focused on two aspects of learning causality from statistical data: learning by independence constraints and learning by properties of conditional distributions. The respective assumptions took in these two approaches are faithfulness assumption and plausibility of Markov kernels. Actually, both assumptions can be related via some kind of simplicity principle on Markov kernels of the desirable structure. The faithfulness requires simplicity (roughly speaking, minimum links) in the structure, which means that each Markov kernels depends on minimum number of variables, while the plausibility requires that each link represents a simple Markov kernel.

Although the experimental results obtained so far seem quite promising, we do not intent to claim that these principles (in particular the specific definition of plausible Markov kernels which allow a space of functions spanned by certain simple monomials) is universally valid, since we do not expect that all real-life causal relationships always exhibit such property. Different applications may require different dependence and complexity measure. A final judgment on the performance of these inference rules actually requires a large number of real-world examples.

Nevertheless, we are of the opinion that kernel methods provide a promising tool for designing appropriate dependences or complexity measures.

Note that, in most experiments in this thesis, we used the prior knowledge to judge the quality of structures. Actually, evaluating the output of a structural learning algorithm in respect of the causal interpretation still remains an open problem. In real-world applications, the final judgment lies in the usability of the causal structure for designing interventions or manipulations.

A. Appendix

A.1. Denseness of RKHS given by Gaussian RBF kernels

Lemma 4 *The RKHS \mathcal{H}_σ given by the Gaussian RBF kernel k_σ defined in Eq. (2.2) is dense in $L^2(P)$ for any probability measure P on \mathbb{R}^m .*

Proof For notational simplicity, the proof is given only for $m = 1$. The extension to the general case is trivial (see [63], Theorem 2). First we show that the function $x \mapsto e^{\sqrt{-1}\omega x}$ ($\omega \in \mathbb{R}$) is approximated by a function in \mathcal{H}_σ with respect to the $L_2(P)$ -norm in an arbitrary accuracy.

Let f be a function in $L^2(\mathbb{R})$ and its Fourier transform be $\tilde{f}(u)$. Because it is known [69] that the condition

$$\int |\tilde{f}(u)|^2 e^{\frac{\sigma^2}{2}u^2} du < \infty$$

implies $f \in \mathcal{H}_\sigma$, we see that the function $x \mapsto e^{-\frac{1}{2\tau^2}x^2} e^{-\sqrt{-1}\omega x} \in \mathcal{H}_\sigma$ for $\tau > \sigma/\sqrt{2}$ and any $\omega \in \mathbb{R}$. From the bounded convergence theorem, we have

$$\mathbb{E}_{P_X} \left[\left| e^{\sqrt{-1}\omega x} - e^{\sqrt{-1}\omega x} e^{-\frac{1}{2\tau^2}x^2} \right|^2 \right] \rightarrow 0 \quad (\tau \rightarrow \infty).$$

Thus, it suffices to show that any function $f \in L^2(P)$ can be arbitrarily approximated in $L^2(P)$ by a function in the linear hull of the class $\{e^{-\sqrt{-1}\omega x} \mid \omega \in \mathbb{R}\}$.

Let f be an arbitrary function in $L^2(P)$. We can assume f is continuously differentiable with a compact support, because those functions are dense in $L^2(P)$. Let $\epsilon > 0$ be an arbitrary positive constant and $M = \sup_{x \in \mathbb{R}} |f(x)|$. Take an interval $[-A, A]$ with a positive number A so that it contains the support of f and $P([-A, A]) > 1 - \epsilon/4M^2$. By the standard theory of Fourier inversion (see [130], Theorem II.8), we know that the series of periodic functions

$$f_N(x) = \sum_{n=-N}^N c_n e^{\frac{\pi\sqrt{-1}}{A} nx}$$

converges uniformly to $f(x)$ on $[-A, A]$ as N goes to infinity, where c_n is given by the Fourier coefficient

$$c_n = \frac{1}{2A} \int_{-A}^A f(x) e^{-\frac{\pi\sqrt{-1}}{A} nx} dx.$$

It follows that $|f(x) - f_N(x)|^2 < \epsilon/2$ on $[-A, A]$ for sufficiently large N , and the periodicity of

A. Appendix

$f_N(x)$ ensures

$$\sup_{x \in \mathbb{R}} |f_N(x)|^2 < (M + \sqrt{\epsilon/2})^2 < 2M^2.$$

We obtain $\mathbb{E}_P [|f - f_N|^2] < \epsilon$, which completes the proof. \blacksquare

A.2. Proof of Theorem 2

The conditional covariance operator can be considered as a special case of a conditional cross-covariance operator (see Definition 11), when X equals Y . Furthermore, the operator $\Sigma_{YY|Z}$ captures the expectation of the conditional variance of a random variable. This is shown in the following theorem, proved by [62].

Theorem 7 *Under Assumption 2 we have*

$$\langle g, \Sigma_{YY|Z} g \rangle_{\mathcal{H}_Y} = \mathbb{E}_Z [\text{Var}_{Y|Z} [g(Y)|Z]]$$

for all $g \in \mathcal{H}_Y$.

Based on this property of the conditional covariance operator, we attempt to prove Theorem 2, an analogous expression for conditional cross-covariance operator $\Sigma_{YX|Z}$ (X does not equal Y).

First, we use the polar identity to compute the conditional cross-covariance operator in terms of the conditional covariance operator. Let (\mathcal{H}_U, k_U) and (\mathcal{H}_Z, k_Z) be respective RKHSs on measurable spaces \mathcal{U} and \mathcal{Z} , (U, Z) a random vector on $\mathcal{U} \times \mathcal{Z}$, and $\Sigma_{UU|Z}$ a conditional covariance operator. Thus, due to the polarization identity and Theorem 7, we have

$$\begin{aligned} \langle \tilde{g}, \Sigma_{UU|Z} \tilde{f} \rangle_{\mathcal{H}_U} &= \frac{1}{4} \left(\langle (\tilde{g} + \tilde{f}), \Sigma_{UU|Z} (\tilde{g} + \tilde{f}) \rangle - \langle (\tilde{g} - \tilde{f}), \Sigma_{UU|Z} (\tilde{g} - \tilde{f}) \rangle \right) \\ &= \frac{1}{4} \mathbb{E}_Z \left[\text{Var}_{U|Z} [(\tilde{g} + \tilde{f})(U)|Z] - \text{Var}_{U|Z} [(\tilde{g} - \tilde{f})(U)|Z] \right] \\ &= \frac{1}{4} \mathbb{E}_Z \left[\left(\text{Var}_{U|Z} [(\tilde{g}(U)|Z)] + 2\text{Cov}[\tilde{f}(U), \tilde{g}(U)|Z] + \text{Var}_{U|Z} [(\tilde{f}(U)|Z)] \right) \right. \\ &\quad \left. - \left(\text{Var}_{U|Z} [(\tilde{g}(U)|Z)] - 2\text{Cov}[\tilde{f}(U), \tilde{g}(U)|Z] + \text{Var}_{U|Z} [(\tilde{f}(U)|Z)] \right) \right] \\ &= \mathbb{E}_Z \left[\text{Cov} [\tilde{f}(U), \tilde{g}(U)|Z] \right] \end{aligned}$$

for arbitrary functions $\tilde{f}, \tilde{g} \in \mathcal{H}_U$.

Now we set $U := (X, Y)$ and define the kernel by a direct sum of reproducing kernels [10, 137]:

$$k_U(u, u') = k_U((x, y), (x', y')) := k_X(x, x') + k_Y(y, y')$$

The RKHS corresponding to this kernel is spanned by functions $k_X(x, \cdot)$ depending only on x and functions $k_Y(y, \cdot)$ that only depend on y . In a straightforward way, we may consider this space as a space of functions with the domain $\mathcal{X} \times \mathcal{Y}$, i.e., the domain of U . For $f \in \mathcal{H}_X$ and

$g \in \mathcal{H}_y$, we observe that $\tilde{f} := f \oplus \mathbf{0}$ and $\tilde{g} := \mathbf{0} \oplus g$ are elements in \mathcal{H}_u , where $\mathbf{0}$ denotes the zero function. We may write

$$\mathbb{E}_Z [\text{Cov}[f(X), g(Y)|Z]] = \mathbb{E}_Z [\text{Cov}[\tilde{f}(U), \tilde{g}(U)|Z]] = \langle \tilde{g}, \Sigma_{UU|Z} \tilde{f} \rangle_{\mathcal{H}_u}$$

Using Eq. (2.4), it is easy to check that

$$\begin{aligned} \langle \tilde{g}, \Sigma_{UU} \tilde{f} \rangle_{\mathcal{H}_u} &= \langle g, \Sigma_{YX} f \rangle_{\mathcal{H}_y}, \\ \langle \tilde{g}, \Sigma_{UZ} h \rangle_{\mathcal{H}_u} &= \langle g, \Sigma_{YZ} h \rangle_{\mathcal{H}_y}, \\ \langle h, \Sigma_{ZU} \tilde{f} \rangle_{\mathcal{H}_z} &= \langle h, \Sigma_{ZX} f \rangle_{\mathcal{H}_z}, \end{aligned}$$

for any $\tilde{f}, \tilde{g} \in \mathcal{H}_u$ and $h \in \mathcal{H}_z$. Due to the representation of conditional cross-covariance operators in Eq. (2.6) and Eq. (2.7), we have therefore

$$\langle \tilde{g}, \Sigma_{UU|Z} \tilde{f} \rangle_{\mathcal{H}_u} = \langle g, \Sigma_{YX|Z} f \rangle_{\mathcal{H}_y}.$$

In summary, we conclude

$$\langle g, \Sigma_{YX|Z} f \rangle_{\mathcal{H}_y} = \mathbb{E}_Z [\text{Cov}[f(X), g(Y)|Z]],$$

which completes the proof of Theorem 2. ■

A.3. Proof of Theorem 4

It is sufficient to prove that we have

$$\langle \ddot{g}, \Sigma_{\tilde{Y}\tilde{X}|Z} \ddot{f} \rangle_{\mathcal{H}_y \otimes \mathcal{H}_z} = \langle \ddot{g}, (\Sigma_{YX} \otimes T_Z) \ddot{f} \rangle_{\mathcal{H}_y \otimes \mathcal{H}_z},$$

for all $\ddot{f} \in \mathcal{H}_x \otimes \mathcal{H}_z$ and $\ddot{g} \in \mathcal{H}_y \otimes \mathcal{H}_z$ of the form $\ddot{f} := f \otimes h_1$ and $\ddot{g} := g \otimes h_2$. T_Z is defined by Eq. (2.8). Recall that for two RKHSs \mathcal{H}_x and \mathcal{H}_z on \mathcal{X} and \mathcal{Z} , respectively, the tensor product $\mathcal{H}_x \otimes \mathcal{H}_z$ is the RKHS on $\mathcal{X} \times \mathcal{Z}$ with the positive definite kernel $k_x \otimes k_z$ [10]. The same applies to $\mathcal{H}_y \otimes \mathcal{H}_z$. We find

$$\begin{aligned} \langle (g \otimes h_2), \Sigma_{\tilde{Y}\tilde{X}|Z} (f \otimes h_1) \rangle_{\mathcal{H}_y \otimes \mathcal{H}_z} &= \mathbb{E}_Z [\text{Cov}[f(X)h_1(Z), g(Y)h_2(Z) | Z]] \\ &= \mathbb{E}_Z [\text{Cov}[f(X), g(Y)] h_1(Z)h_2(Z)] \\ &= \text{Cov}[f(X), g(Y)] \mathbb{E}_Z [h_1(Z)h_2(Z)] \\ &= \langle g, \Sigma_{YX} f \rangle_{\mathcal{H}_y} \langle h_2, T_Z h_1 \rangle_{\mathcal{H}_z}. \end{aligned}$$

The first equality uses the definition of \tilde{X} and \tilde{Y} as well as the facts that $f \in \mathcal{H}_x$, $g \in \mathcal{H}_y$, $h_1, h_2 \in \mathcal{H}_z$. The second equality uses $Z \perp (X, Y)$ and that for every specific given value of Z , the variables $h_1(Z)$ and $h_2(Z)$ reduce to constants. The second statement of this theorem follows

A. Appendix

then directly from the definitions of $\|\Sigma_{YX}\|_{\text{HS}}^2$ and $\|\Sigma_{\check{Y}\check{X}|Z}\|_{\text{HS}}^2$. ■

A.4. Proof of Theorem 6

The following proof is based on the similar idea to the proof of Lemma 7 and Lemma 10 in [62]. However, we cannot directly use their convergence proofs since the latter refer to the convergence of traces of conditional covariance operators and we have to show convergence of traces of the squares of the conditional cross-covariance operators. Moreover, our dependence measure uses an appropriate renormalization.

First of all, according to the definition of the renormalization factor β_Z in Eq. (2.8) and its estimator $\hat{\beta}_Z^{(n)}$ in Eq. (2.10), Hoeffding's inequality [91] implies that

$$|\hat{\beta}_Z^{(n)} - \beta_Z| = O_p(n^{-1/2}). \quad (\text{A.1})$$

Furthermore, we have

$$\begin{aligned} & \|\Sigma_{YX|Z}\|_{\text{HS}}^2 - \|\widehat{\Sigma}_{YX|Z}^{(n,\epsilon)}\|_{\text{HS}}^2 \\ &= \left(\|\Sigma_{YX|Z}\|_{\text{HS}} - \|\widehat{\Sigma}_{YX|Z}^{(n,\epsilon)}\|_{\text{HS}} \right) \left(\|\Sigma_{YX|Z}\|_{\text{HS}} + \|\widehat{\Sigma}_{YX|Z}^{(n,\epsilon)}\|_{\text{HS}} \right). \end{aligned} \quad (\text{A.2})$$

If we could show that the first term converges to zero as in order $O_p(\epsilon^{-1}n^{-1/2})$, then the second term is consequently bounded in n . Thus, it remains merely to proof the convergence of the first term. Due to the triangle inequality, it is clear that

$$\|\Sigma_{YX|Z}\|_{\text{HS}} - \|\widehat{\Sigma}_{YX|Z}^{(n,\epsilon)}\|_{\text{HS}} \leq \|\Sigma_{YX|Z} - \widehat{\Sigma}_{YX|Z}^{(n,\epsilon)}\|_{\text{HS}}.$$

Using the definitions in Eq. (2.6) and Eq. (2.11) of $\Sigma_{YX|Z}$ and $\widehat{\Sigma}_{YX|Z}^{(n,\epsilon)}$, respectively, the right-hand side is bounded from above by

$$\|\Sigma_{YX} - \widehat{\Sigma}_{YX}^{(n)}\|_{\text{HS}} + \|\Sigma_{YZ}(\Sigma_{ZZ} + \epsilon I)^{-1}\Sigma_{ZX} - \widehat{\Sigma}_{YZ}^{(n)}(\widehat{\Sigma}_{ZZ}^{(n)} + \epsilon I)^{-1}\widehat{\Sigma}_{ZX}^{(n)}\|_{\text{HS}}.$$

The first summand converges to zero as in order $O_p(n^{-1/2})$ (see [60], Lemma 5). The second term has the form

$$\|AB^{-1}C - \widehat{A}_{(n)}\widehat{B}_{(n)}^{-1}\widehat{C}_{(n)}\|_{\text{HS}} \quad (\text{A.3})$$

if we use the shorthands $A := \Sigma_{YZ}$, $B := \Sigma_{ZZ} + \epsilon I$, $C := \Sigma_{ZX}$, and $\widehat{A}_{(n)}$, $\widehat{B}_{(n)}$, $\widehat{C}_{(n)}$ for the respective estimators from n data points.

Due to the triangle inequality, the term in Eq. (A.3) is then bounded from above by

$$\|(A - \widehat{A}_{(n)})\widehat{B}_{(n)}^{-1}\widehat{C}_{(n)}\|_{\text{HS}} + \|A(\widehat{B}_{(n)}^{-1} - B^{-1})C\|_{\text{HS}} + \|A\widehat{B}_{(n)}^{-1}(C - \widehat{C}_{(n)})\|_{\text{HS}}. \quad (\text{A.4})$$

The first and the third term converge to zero at speed $\epsilon^{-1}n^{-1/2}$, because of the fact that $\|\widehat{A}_{(n)} - A\|_{\text{HS}} = O_p(n^{-1/2})$, $\|\widehat{C}_{(n)} - C\|_{\text{HS}} = O_p(n^{-1/2})$ and the spectra of $\widehat{B}_{(n)}$ and B are both bounded

from below by ϵ . Note that $\widehat{C}_{(n)}$ is uniformly bounded in n since the operators themselves converge even in HS-norm. It remains therefore to analyze the convergence of the second term in Eq. (A.4). We have

$$\begin{aligned} \|A(B^{-1} - \widehat{B}_{(n)}^{-1})C\|_{\text{HS}} &= \|AB^{-1/2}(B^{1/2}\widehat{B}_{(n)}^{-1}B^{1/2} - I)B^{-1/2}C\|_{\text{HS}} \\ &\leq \|AB^{-1/2}\| \|B^{-1/2}C\| \|B^{1/2}\widehat{B}_{(n)}^{-1}B^{1/2} - I\|_{\text{HS}} \\ &= \|AB^{-1/2}\| \|B^{-1/2}C\| \|\widehat{B}_{(n)}^{-1/2}B\widehat{B}_{(n)}^{-1/2} - I\|_{\text{HS}}, \end{aligned} \quad (\text{A.5})$$

where the last equality follows from the fact that the spectrum of $\widehat{B}_{(n)}^{-1/2}B\widehat{B}_{(n)}^{-1/2}$ coincides with that of $B^{1/2}\widehat{B}_{(n)}^{-1}B^{1/2}$. Since we have the bounds

$$\|B^{-1/2}C\| = \|(\Sigma_{ZZ} + \epsilon I)^{-1/2}\Sigma_{ZZ}^{1/2}V_{ZX}\| \leq 1$$

and

$$\|AB^{-1/2}\| = \|V_{YZ}\Sigma_{ZZ}^{1/2}(\Sigma_{ZZ} + \epsilon I)^{-1/2}\| \leq 1,$$

the second term in (A.5) is then bounded from above by

$$\begin{aligned} \|\widehat{B}_{(n)}^{-1/2}B\widehat{B}_{(n)}^{-1/2} - I\|_{\text{HS}} &= \|\widehat{B}_{(n)}^{-1/2}(B - \widehat{B}_{(n)})\widehat{B}_{(n)}^{-1/2}\|_{\text{HS}} \\ &\leq \|\widehat{B}_{(n)}^{-1}\| \|B - \widehat{B}_{(n)}\|_{\text{HS}}. \end{aligned} \quad (\text{A.6})$$

Using the upper bound ϵ^{-1} for the spectrum of $\widehat{B}_{(n)}^{-1}$, the last term of Eq. (A.6) is bounded from above by

$$\epsilon^{-1}\|B - \widehat{B}_{(n)}\|_{\text{HS}}$$

Due to $\|B - \widehat{B}_{(n)}\|_{\text{HS}} = O_p(n^{-1/2})$ we have shown that the left-hand side of Eq. (A.2) converges to zero as in order $O_p(\epsilon^{-1}n^{-1/2})$.

Let us summarize the results above to study the convergence of $\mathbb{H}_{YX|Z}^{(n,\epsilon)}$.

$$\begin{aligned} &\left| \widehat{\mathbb{H}}_{YX|Z}^{(n,\epsilon)} - \mathbb{H}_{YX|Z} \right| \\ &= \left| \widehat{\beta}_Z^{(n)} \|\widehat{\Sigma}_{YX|Z}^{(n,\epsilon)}\|_{\text{HS}}^2 - \beta_Z \|\Sigma_{YX|Z}\|_{\text{HS}}^2 \right| \\ &\leq \|\widehat{\Sigma}_{YX|Z}^{(n,\epsilon)}\|_{\text{HS}}^2 |\widehat{\beta}_Z^{(n)} - \beta_Z| + \beta_Z \left| \|\widehat{\Sigma}_{YX|Z}^{(n,\epsilon)}\|_{\text{HS}}^2 - \|\Sigma_{YX|Z}\|_{\text{HS}}^2 \right|. \end{aligned} \quad (\text{A.7})$$

Due to Eq. (A.1), the first term of Eq. (A.7) is of order $O_p(n^{-1/2})$ and the second term converges in probability at the rate of $O_p(\epsilon^{-1}n^{-1/2})$. In summary, we have a convergence speed of $O_p(\epsilon^{-1}n^{-1/2})$ in probability, which completes the first part of our proof.

For the other part of Theorem 6, we show that $\Sigma_{YX|Z}^{(\epsilon)}$ converges to $\Sigma_{YX|Z}$ in HS-norm for

A. Appendix

$\epsilon \rightarrow 0$. We have

$$\begin{aligned}
& \left\| \Sigma_{YX|Z} - \Sigma_{YX|Z}^{(\epsilon)} \right\|_{\text{HS}}^2 \\
&= \left\| \Sigma_{YY}^{1/2} V_{YZ} (I - \Sigma_{ZZ}^{1/2} (\Sigma_{ZZ} + \epsilon I)^{-1} \Sigma_{ZZ}^{1/2}) V_{ZX} \Sigma_{XX}^{1/2} \right\|_{\text{HS}}^2 \\
&= \left\| \Sigma_{YY}^{1/2} V_{YZ} (\epsilon (\Sigma_{ZZ} + \epsilon I)^{-1}) V_{ZX} \Sigma_{XX}^{1/2} \right\|_{\text{HS}}^2 \\
&= \text{Tr} \left(\Sigma_{XX}^{1/2} V_{XZ} (\epsilon (\Sigma_{ZZ} + \epsilon I)^{-1}) V_{ZY} \Sigma_{YY} V_{YZ} (\epsilon (\Sigma_{ZZ} + \epsilon I)^{-1}) V_{ZX} \Sigma_{XX}^{1/2} \right) \\
&= \text{Tr} \left(\sqrt{\epsilon} (\Sigma_{ZZ} + \epsilon I)^{-1/2} V_{ZY} \Sigma_{YY} V_{YZ} \epsilon (\Sigma_{ZZ} + \epsilon I)^{-1} V_{ZX} \Sigma_{XX} V_{XZ} \sqrt{\epsilon} (\Sigma_{ZZ} + \epsilon I)^{-1/2} \right)
\end{aligned}$$

The last equality follows from the fact that $\text{Tr}(B^*TB) = \text{Tr}(TBB^*)$ for any positive trace class operator T and bounded operator B .

With a complete orthogonal system $\{\phi_i\}_{i=1}^{\infty}$ for \mathcal{H}_Z subject to $\Sigma_{ZZ}\phi_i = \lambda_i\phi_i$ with eigenvalues $\lambda_i \geq 0$, the equation above can be rephrased as follows:

$$\begin{aligned}
& \sum_{i,j=1}^{\infty} \langle \phi_i, \sqrt{\epsilon} (\Sigma_{ZZ} + \epsilon I)^{-1/2} V_{ZY} \Sigma_{YY} V_{YZ} \sqrt{\epsilon} (\Sigma_{ZZ} + \epsilon I)^{-1/2} \phi_j \rangle_{\mathcal{H}_Z} \\
& \quad \langle \phi_j, \sqrt{\epsilon} (\Sigma_{ZZ} + \epsilon I)^{-1/2} V_{ZX} \Sigma_{XX} V_{XZ} \sqrt{\epsilon} (\Sigma_{ZZ} + \epsilon I)^{-1/2} \phi_i \rangle_{\mathcal{H}_Z} \\
&= \sum_{i,j=1}^{\infty} \frac{\epsilon}{\lambda_i + \epsilon} \frac{\epsilon}{\lambda_j + \epsilon} \langle \phi_i, V_{ZY} \Sigma_{YY} V_{YZ} \phi_j \rangle_{\mathcal{H}_Z} \langle \phi_j, V_{ZX} \Sigma_{XX} V_{XZ} \phi_i \rangle_{\mathcal{H}_Z} \quad (\text{A.8})
\end{aligned}$$

The absolute value of each summand in Eq. (A.8) is bounded from above by

$$\left| \langle \phi_i, V_{ZY} \Sigma_{YY} V_{YZ} \phi_j \rangle_{\mathcal{H}_Z} \langle \phi_j, V_{ZX} \Sigma_{XX} V_{XZ} \phi_i \rangle_{\mathcal{H}_Z} \right|,$$

which does not depend on ϵ , and due to the Cauchy-Schwartz inequality the infinite sum of these terms

$$\sum_{i,j=1}^{\infty} \left| \langle \phi_i, V_{ZY} \Sigma_{YY} V_{YZ} \phi_j \rangle_{\mathcal{H}_Z} \langle \phi_j, V_{ZX} \Sigma_{XX} V_{XZ} \phi_i \rangle_{\mathcal{H}_Z} \right|,$$

is bounded from above by

$$\left(\sum_{i,j=1}^{\infty} \langle \phi_i, V_{ZY} \Sigma_{YY} V_{YZ} \phi_j \rangle_{\mathcal{H}_Z}^2 \right)^{1/2} \left(\sum_{i,j=1}^{\infty} \langle \phi_j, V_{ZX} \Sigma_{XX} V_{XZ} \phi_i \rangle_{\mathcal{H}_Z}^2 \right)^{1/2},$$

which is finite because $V_{ZY} \Sigma_{YY} V_{YZ}$ and $V_{ZX} \Sigma_{XX} V_{XZ}$ are Hilbert-Schmidt. Thus, from the dominated convergence theorem, the limit $\epsilon \rightarrow 0$ commutes with the infinite sum in Eq. (A.8). Since each summand of pair (i, j) in Eq. (A.8) converges to zero for $\epsilon \rightarrow 0$, the HS-norm between

A.5. Plausible Markov kernels between binary and real-valued variable

$\Sigma_{YX|Z}^{(\epsilon)}$ and $\Sigma_{YX|Z}$ converges to zero for $\epsilon \rightarrow 0$. This results in

$$\left| \mathbb{H}_{YX|Z}^{(\epsilon)} - \mathbb{H}_{YX|Z} \right| = |\beta_Z| \left| \left\| \Sigma_{YX|Z}^{(\epsilon)} \right\|_{\text{HS}}^2 - \left\| \Sigma_{YX|Z} \right\|_{\text{HS}}^2 \right| \rightarrow 0 \quad (\epsilon \rightarrow 0),$$

which completes the proof of Theorem 6. ■

A.5. Plausible Markov kernels between binary and real-valued variable

Here we derive the plausible Markov kernels of the causation between a binary variable X with $x \in \{-1, +1\}$ and a real-valued variable Y with $y \in \mathbb{R}$. For the sake of simplicity, we denote $x_{\pm 1}$ for the cases $X = \pm 1$. Assuming a hypothetical causal direction $X \rightarrow Y$, the plausible Markov kernel $\mathcal{Q}(X)$ is determined only through the constraint of its first moment μ^X . Note that the second moment of X is the constant 1. Defining

$$\mathcal{Q}(x_{+1}) = \frac{1}{2}(1 + \mu^X) =: q,$$

we have

$$\mathcal{Q}(x_{-1}) = \frac{1}{2}(1 - \mu^X) = 1 - q.$$

To determine the plausible Markov kernel $\mathcal{Q}(Y|X)$ we maximize the entropy function \mathcal{H}

$$\mathcal{H}(Y|X) = q \mathcal{H}(Y|x_{+1}) + (1 - q) \mathcal{H}(Y|x_{-1}) \quad (\text{A.9})$$

subject to the constraints

$$q E_{+1} + (1 - q) E_{-1} = \mu^Y \quad (\text{A.10})$$

$$q E_{+1} - (1 - q) E_{-1} = \beta^{XY} \quad (\text{A.11})$$

$$q (E_{+1})^2 + (1 - q) (E_{-1})^2 + q \text{Var}_{+1} + (1 - q) \text{Var}_{-1} = \alpha^Y \quad (\text{A.12})$$

Here μ^Y is the first moment of Y , β^{XY} the second mixed moment of X and Y , α^Y the second moment of Y . These values are known. $E_{\pm 1}$ denote the expectations of the conditional variable $(Y|x_{\pm 1})$ and $\text{Var}_{\pm 1}$ the variances of $(Y|x_{\pm 1})$, respectively. These values are yet to be determined. However, $E_{\pm 1}$ can be uniquely determined from Eq. (A.10) and Eq. (A.11):

$$\begin{aligned} E_{+1} &= \frac{\mu^Y + \beta^{XY}}{1 + \mu^X}, \\ E_{-1} &= \frac{\mu^Y - \beta^{XY}}{1 - \mu^X}. \end{aligned}$$

A. Appendix

Therefore only one more constraint remains to be satisfied:

$$q \text{Var}_{+1} + (1 - q) \text{Var}_{-1} =: \sigma^2 \quad (\text{A.13})$$

where

$$\begin{aligned} \sigma^2 &= \beta^Y - (q (E_{+1})^2 + (1 - q) (E_{-1})^2) \\ &= \beta^Y - \frac{(\mu^Y + \beta^{XY})^2}{2(1 + \mu^X)} - \frac{(\mu^Y - \beta^{XY})^2}{2(1 - \mu^X)}. \end{aligned}$$

Here σ^2 can be calculated directly from all known values. The maximization of the function in Eq. (A.9) with satisfying the constraint in Eq. (A.13) obviously has the unique solution that $\mathcal{Q}(Y|x_{+1})$ and $\mathcal{Q}(Y|x_{-1})$ are both Gaussian:

$$\mathcal{Q}(Y|x_{+1}) \propto \mathcal{N}(E_{+1}, \text{Var}_{+1}) \quad \text{and} \quad \mathcal{Q}(Y|x_{-1}) \propto \mathcal{N}(E_{-1}, \text{Var}_{-1}).$$

Otherwise it would be inconsistent with the well known fact that a normal distribution maximizes the entropy for given expectation and variance. The maximal entropy of $\mathcal{Q}(Y|X)$ of Eq. (A.9) in such case can be formulated as follows:

$$\mathcal{H}(Y|X) = \frac{1}{2} \ln(2\pi e) + \frac{q}{2} \ln(\text{Var}_{+1}) + \frac{1-q}{2} \ln(\text{Var}_{-1}) \quad (\text{A.14})$$

since the entropies of both Gaussian distributions are $\frac{1}{2} \ln(2\pi e \text{Var}_{+1})$ and $\frac{1}{2} \ln(2\pi e \text{Var}_{-1})$ respectively. Substituting Eq. (A.13) into Eq. (A.14), to achieve the maximum the first-order derivative must vanish and the second-order derivative should be negative, so that we obtain

$$\text{Var}_{+1} = \text{Var}_{-1} = \sigma^2$$

which means $\mathcal{H}(Y|X)$ achieves its maximum if and only if

$$\mathcal{Q}(Y|x_{-1}) \propto \mathcal{N}(\mu_{-1}, \sigma^2) \quad \text{and} \quad \mathcal{Q}(Y|x_{+1}) \propto \mathcal{N}(\mu_{+1}, \sigma^2).$$

The Markov kernels $\mathcal{R}(Y)$ and $\mathcal{R}(X|Y)$ for the other causal direction $X \rightarrow Y$ can also be determined analytically. Firstly, it is known that for fixed first (μ^Y) and second moment (α^Y), the Gaussian distribution $\mathcal{N}(\mu^Y, \alpha^Y - (\mu^Y)^2)$ maximizes the differential entropy of the real-valued variable Y . To determine $\mathcal{R}(X|Y)$ we maximize the entropy function

$$\mathcal{H}(X|Y) = - \int (\mathcal{R}(x_{+1}|y) \ln(\mathcal{R}(x_{+1}|y)) + \mathcal{R}(x_{-1}|y) \ln(\mathcal{R}(x_{-1}|y))) \mathcal{R}(y) dy$$

A.5. Plausible Markov kernels between binary and real-valued variable

subject to the constraints

$$\mathcal{R}(x_{+1}|y) + \mathcal{R}(x_{-1}|y) = 1 \quad \forall y \in \mathbb{R} \quad (\text{A.15})$$

$$\int (\mathcal{R}(x_{+1}|y) - \mathcal{R}(x_{-1}|y)) \mathcal{R}(y) dy = \mu^X \quad (\text{A.16})$$

$$\int y (\mathcal{R}(x_{+1}|y) - \mathcal{R}(x_{-1}|y)) \mathcal{R}(y) dy = \beta^{XY} \quad (\text{A.17})$$

$$\int (\mathcal{R}(x_{+1}|y) + \mathcal{R}(x_{-1}|y)) \mathcal{R}(y) dy = \alpha^X \equiv 1 \quad (\text{A.18})$$

Here μ^X and α^X are the known first and second moments of X . Eq. (A.18) holds trivially. Through the substitution of Eq. (A.15) in Eq. (A.16) and Eq. (A.17) only the following two constraints remain:

$$\int (2\mathcal{R}(x_{+1}|y) - 1) \mathcal{R}(y) dy = \mu^X \quad (\text{A.19})$$

$$\int y (2\mathcal{R}(x_{+1}|y) - 1) \mathcal{R}(y) dy = \beta^{XY} \quad (\text{A.20})$$

By introducing two positive Lagrange multipliers λ and ν the solution of $\mathcal{R}(X|Y)$ must be of the form

$$\begin{aligned} \mathcal{R}(x_{-1}|y) &= \frac{e^{-(\lambda y + \nu)}}{e^{\lambda y + \nu} + e^{-(\lambda y + \nu)}} = \frac{1}{2} - \frac{1}{2} \tanh(\lambda y + \nu), \\ \mathcal{R}(x_{+1}|y) &= \frac{e^{\lambda y + \nu}}{e^{\lambda y + \nu} + e^{-(\lambda y + \nu)}} = \frac{1}{2} + \frac{1}{2} \tanh(\lambda y + \nu). \end{aligned}$$

Together with Eq. (A.19) and Eq. (A.20) the unknowns λ and μ should satisfy the following equations system

$$\begin{aligned} \int \tanh(\lambda y + \nu) \mathcal{R}(y) dy &= \mu^X \\ \int y \tanh(\lambda y + \nu) \mathcal{R}(y) dy &= \beta^{XY} \end{aligned}$$

where $\mathcal{R}(y) \propto \mathcal{N}(\mu^Y, \alpha^Y - (\mu^Y)^2)$. Solving this nonlinear system, we will be able to determine λ and μ , and therefore $\mathcal{R}(X|Y)$ for every given μ^X and β^{XY} .

In summary, we obtain a closed-form solution for the causation between a binary and a real-valued variable. For one causal direction $X \rightarrow Y$, we have plausible Markov kernels in the form of

$$\begin{aligned} \mathcal{Q}(x_{-1}) &= \frac{1}{2} (1 - \mu^X) & \text{and} & & \mathcal{Q}(x_{+1}) &= \frac{1}{2} (1 + \mu^X) \\ \mathcal{Q}(Y|x_{-1}) &\propto \mathcal{N}(\mu_{-1}, \sigma^2) & \text{and} & & \mathcal{Q}(Y|x_{+1}) &\propto \mathcal{N}(\mu_{+1}, \sigma^2) \end{aligned}$$

A. Appendix

where

$$\mu_{-1} = \frac{\mu^Y - \beta^{XY}}{1 - \mu^X}, \quad \mu_{+1} = \frac{\mu^Y + \beta^{XY}}{1 + \mu^X} \quad \text{and} \quad \sigma^2 = \alpha^Y - \frac{(\mu^Y + \beta^{XY})^2}{2(1 + \mu^X)} - \frac{(\mu^Y - \beta^{XY})^2}{2(1 - \mu^X)}.$$

For the other causal direction $Y \rightarrow X$, the plausible Markov kernels have the form

$$\begin{aligned} \mathcal{R}(Y) &\propto \mathcal{N}\left(\mu^Y, \alpha^Y - (\mu^Y)^2\right) \\ \mathcal{R}(x_{-1}|y) &= \frac{1}{2} - \frac{1}{2} \tanh(\lambda y + \nu) \quad \text{and} \quad \mathcal{R}(x_{+1}|y) = \frac{1}{2} + \frac{1}{2} \tanh(\lambda y + \nu). \end{aligned}$$

Having computed these plausible Markov kernels, the corresponding joint distributions

$$\begin{aligned} \mathcal{Q}(X, Y) &= \mathcal{Q}(Y|X) \mathcal{Q}(X) && \text{(with respect to causal direction } X \rightarrow Y) \\ \mathcal{R}(X, Y) &= \mathcal{R}(X|Y) \mathcal{R}(Y) && \text{(with respect to causal direction } Y \rightarrow X) \end{aligned}$$

can be calculated. The question is whether \mathcal{Q} could equal \mathcal{R} under certain conditions, because if the equation $\mathcal{Q} = \mathcal{R}$ holds, the causal directions ($X \rightarrow Y$ and $Y \rightarrow X$) can no longer be distinguished from one another, based on our ‘‘principle of plausible Markov kernels’’. However, one may verify that whenever there exists correlation between X and Y , our method with most plausible Markov kernels leads always to different joint distributions. This is because the marginal distribution of Y based on the causal direction $X \rightarrow Y$ is a convex sum of two Gaussian distributions which have different expected values for non-vanishing correlation between X and Y . This distribution cannot coincide with the marginal distribution of Y based the causal direction $Y \rightarrow X$ since the latter is Gauss distributed.

B. Appendix

B.1. Pseudocode of Orientation Procedure A

Given a adjacency structure \mathcal{G}_{input} , the main procedure **FixOrientation** is applied to all possible distinct triples (X_a, X_b, X_c) from variables X_1, \dots, X_N . If edges $X_a - X_c$ and $X_c - X_b$ are present in \mathcal{G}_{input} , the procedure **FixOrientation** focuses on the undirected subgraph $X_a - X_c - X_b$ (X_a and X_b possibly adjacent) and calls the procedure **ProposeCollider** to test whether X_c can be accepted as a candidate for being a common effect of X_a and X_b . This is done on the basis of Criteria 1 and 2 in turn.

The essential data structure during the subroutines are partially directed graphs on N nodes, stored as $N \times N$ -matrices. An undirected edge $X_a - X_b$ is represented by an entry “1” at the positions (a, b) and (b, a) of the representing matrix. A directed edge $X_a \rightarrow X_b$ corresponds to a negative entry at position (a, b) and a “0” at (b, a) . “0” at both positions (a, b) and (b, a) indicates the absence of the edge between X_a and X_b . During the voting procedure it is decreased or increased by 1, depending on whether the current vote agrees or disagrees with the current value at the corresponding position of the matrix. The negative value quantifies the current evidence for one direction. If “0” is reached during counting, the entries at the positions (a, b) and (b, a) are both reset to “1”, i.e., there is again no evidence for either of both directions. The given adjacency structure \mathcal{G}_{input} is stored in the matrix M_0 . If one intends to incorporate prior causal knowledge such as temporal ordering, M_0 can be used to store it. However, the following procedures need to be slightly modified. The matrix M_i represents voting according to assumption $i = 1, 2$. After all these votes are counted, the main procedure performs the orientation. First, the arrows are directed using the votes in M_1 , then the votes in M_2 are used to direct the remaining edges if the majority principle leads to a definite direction. This results in a (partially) directed graph \mathcal{G}_{output} .

The procedure **Graph2Matrix** encodes the graph into a matrix, where the entry of a negative integer indicates a directed edge, the entry “1” indicates an undirected edge and the entry “0” indicates an absence of an edge. The inverse procedure **Matrix2Graph** is straightforward and thus omitted.

B.1.1. Procedure FixOrientation

Input: An undirected graph \mathcal{G}_{input} with edges $-$;

Output: A (partially) directed graph \mathcal{G}_{output} with edges $\rightarrow, -$.

(1) Initialize $M_0, M_1, M_2 := \text{Graph2Matrix}(\mathcal{G}_{input})$.

// Initialize $M_{0,1,2}$ with the input skeleton. M_0 will remain unaffected. $M_{1,2}$ will count the votes.

(2) **for** $a, b, c = 1$ to N

B. Appendix

```

// Check for all possible distinct ordered triples  $(X_a, X_b, X_c)$  taken from  $X_1, \dots, X_N$  whether they represent colliders.
if  $M_0(X_a, X_c) = 1$  and  $M_0(X_b, X_c) = 1$  // For all substructures  $X_a - X_c - X_b$  in  $M_0$ .
  then for  $i = 1, 2$ 
    Criterion :=  $i$ ;
     $M_i := \text{ProposeCollider}(\text{Criterion}, M_i, \text{Candidate } X_a - X_c - X_b)$ ;
    // Update  $M_i$  by adding a vote for  $X_a \rightarrow X_c$  and  $X_c \leftarrow X_b$ , respectively,
    // if Criterion  $i$  considers  $X_c$  as a collider candidate
  end for;
end if;
end for.
(3) for  $i, j = 1$  to  $N$  // Given the skeleton in  $M_0$ , use the voting results in  $M_{1,2}$  in turn to direct all edges with
  // unbalanced results.  $M_1$  has priority over  $M_2$ . Store these orientations in  $M_0$ .
  Case:  $M_0(i, j) = 1$  and  $M_1(i, j) \leq -1$  // If  $M_0$  contains  $X_i - X_j$  and  $M_1$  contains  $X_i \rightarrow X_j$ .
    set  $M_0(i, j) := -1$  and  $M_0(j, i) := 0$ ; // Direct  $X_i \rightarrow X_j$  in  $M_0$ .
  Case:  $M_0(i, j) = 1$  and  $M_2(i, j) \leq -1$  // If  $M_0$  contains  $X_i - X_j$  and  $M_2$  contains  $X_i \rightarrow X_j$ .
    set  $M_0(i, j) := -1$  and  $M_0(j, i) := 0$ , // Direct  $X_i \rightarrow X_j$  in  $M_0$ .
end for.
(4)  $\mathcal{G}_{output} = \text{Matrix2Graph}(M_0)$ . // Decode the resulting matrix  $M_0$  into graph  $\mathcal{G}_{output}$  as output.

```

B.1.2. Procedure ProposeCollider

Input: (i) A matrix M_{in} with integer entries ≤ 1 representing a partially directed graph;
(ii) A substructure $X_a - X_c - X_b$;
(iii) An integer $i = 1, 2$ determining which of the criteria is taken.

Output: matrix M_{out} with integer entries ≤ 1 representing a (partially) directed graph .

```

(1) Initialize  $M_{out} := M_{in}$  and ProposeAccepted := false .
(2) Case: Criterion = 1 // Collider test by the  $\lambda$ -collider condition with a very large  $\lambda = 100$  .
  compute  $h_c := \widehat{\mathbb{H}}_{X_a X_b | X_c} / \widehat{\mathbb{H}}_{X_a X_b}$  ;
  if  $h_c \geq 100$  then ProposeAccepted := true , end if ;
  Case: Criterion = 2 // Collider test by the  $\lambda$ -collider condition with a smaller  $\lambda$  .
  compute  $h_c := \widehat{\mathbb{H}}_{X_a X_b | X_c} / \widehat{\mathbb{H}}_{X_a X_b}$  ,  $h_b := \widehat{\mathbb{H}}_{X_a X_c | X_b} / \widehat{\mathbb{H}}_{X_a X_c}$  ,  $h_a := \widehat{\mathbb{H}}_{X_b X_c | X_a} / \widehat{\mathbb{H}}_{X_b X_c}$  ;
  if  $h_c \geq \max\{h_b, h_a\}$  then ProposeAccepted := true , end if ;
(3) if ProposeAccepted // A vote for  $X_a \rightarrow X_c$  and  $X_b \rightarrow X_c$ , respectively.
  for  $i = a, b$ 
    Case:  $M_{out}(i, c) = 1$  and  $M_{out}(c, i) = 1$  // If  $M_{out}$  contains the undirected edges  $X_{a,b} - X_c$ .
      set  $M_{out}(i, c) := -1$  and  $M_{out}(c, i) := 0$ ; // Orient  $X_{a,b} \rightarrow X_c$ .
    Case:  $M_{out}(i, c) \leq -1$  and  $M_{out}(c, i) = 0$  // New vote coincides with orientation stored in  $M_{out}$ .
      set  $M_{out}(i, c) := M_{out}(i, c) - 1$ ;
      // Leave the orientation untouched and increase counter for  $X_i \rightarrow X_c$  by  $-1$ .
    Case:  $M_{out}(i, c) = 0$  and  $M_{out}(c, i) < -1$ 
      // New vote is opposite to the current orientation stored in  $M_{out}$ .
      set  $M_{out}(c, i) := M_{out}(c, i) + 1$ ;
  end for ;

```

B.2. Pseudocode of Orientation Procedure B

```

// Leave the orientation untouched and decrease counter for  $X_c \rightarrow X_{a,b}$  by  $-1$ .
Case:  $M_{out}(i, c) = 0$  and  $M_{out}(c, i) = -1$ 
// If  $M_{out}$  contains the opposite orientation  $X_c \rightarrow X_{a,b}$  with a counter of  $-1$ .
set  $M_{out}(i, c) := 1$  and  $M_{out}(c, i) := 1$ ; // Reset to an undirected edge  $X_{a,b} - X_c$ .
end for;
end if.

```

B.1.3. Procedure Graph2Matrix

Input: An undirected or (partially) directed graph \mathcal{G} of X_1, \dots, X_N with edges $\rightarrow, -$;

Output: An $N \times N$ matrix M with entries “ -1 ”, “ 0 ”, and “ 1 ”;

(1) Initialize $M(i, j) := 0, i, j = 1, \dots, N$. // Start with a matrix of all-over zero.

(2) for $i, j = 1$ to N

Case: $X_i - X_j$, set $M(i, j) := 1$; // Undirected edge.

Case: $X_i \rightarrow X_j$, set $M(i, j) := -1$; // Directed edge.

end for.

B.2. Pseudocode of Orientation Procedure B

In analogy to orientation procedure A in Appendix B.1, the procedures **FixOrientation***, **ProposeCollider*** and **Graph2Matrix*** of orientation procedure B are designed to infer the orientation for the final output by a unanimous vote. The resulting graph contains \rightarrow meaning the direction is supported by a unanimous vote, $-$ meaning no votes are obtained for both directions, and \leftrightarrow meaning at least one vote is obtained for both directions. We conduct the voting procedure **ProposeCollider*** for all substructures $X_a - X_c - X_b$, where X_a and X_b are nonadjacent.¹ The final voting results are stored in M and the binary matrix L memorizes whether an orientation $X_i \rightarrow X_j$ ever obtained a vote or not. Based on the information from M and L , the main procedure **FixOrientation*** performs the orientation. The procedure **Matrix2Graph*** decodes orientation information from the matrix M and L into a mixed graph with un-, uni- and bi-directed edges.

B.2.1. Procedure FixOrientation*

Input: An undirected graph \mathcal{G}_{input} with edges $-$;

Output: A (partially) directed graph \mathcal{G}_{output} with edges $\rightarrow, -, \leftrightarrow$.

(1) Initialize $M_0, M := \text{Graph2Matrix}(\mathcal{G}_{input})$. Initialize L as zero matrix.

// Initialize M_0, M with the input skeleton. M_0 will remain unaffected. M will count the votes.

(2) for $a, b, c = 1$ to N

// Check for all possible distinct triples (X_a, X_b, X_c) taken from X_1, \dots, X_N whether they represent colliders.

¹The extension of the procedures for identifying unshielded colliders to shielded colliders is straightforward and thus omitted. We just need to apply procedures to shielded triples $X_a - X_c - X_b$, where X_a and X_b are adjacent, instead of unshielded triples.

B. Appendix

```

if  $M_0(X_a, X_c) = 1$ ,  $M_0(X_b, X_c) = 1$  and  $M_0(X_a, X_b) = 0$ 
    //  $M_0$  contains the subgraph  $X_a - X_c - X_b$ , where  $X_a$  and  $X_b$  are nonadjacent.
    // If the procedure is applied to fully connected triples, the condition  $M_0(X_a, X_b) = 1$  should be used.
     $(M, L) := \text{ProposeCollider}^*(M, \text{Candidate } X_a - X_c - X_b, L)$ ;
    // Update  $M$  by adding the votes for the arrows  $X_a \rightarrow X_c$  and  $X_c \leftarrow X_b$ , identified by  $\text{ProposeCollider}^*$ .
    //  $L$  memorizes whether a vote was ever given to a direction or not.
end if;
end for.
(3) for  $i, j = 1$  to  $N$ 
    // Given the skeleton  $M_0$ , use the voting information from  $M$  and  $L$  to direct edges in  $M_0$ .
    Case:  $M_0(i, j) = 1$ ,  $M_0(j, i) = 1$ ,  $M(i, j) \leq -1$ ,  $L(i, j) = 1$  and  $L(j, i) = 0$ 
        //  $M_0$  contains  $X_i - X_j$ ,  $M$  contains  $X_i \rightarrow X_j$ , The opposite orientation  $X_i \leftarrow X_j$  obtained no votes.
        set  $M_0(i, j) := -1$ ,  $M_0(j, i) := 0$ ; // Direct  $X_i \rightarrow X_j$  in  $M_0$ .
    Case:  $M_0(i, j) = 1$ ,  $M_0(j, i) = 1$ ,  $L(i, j) = 1$  and  $L(j, i) = 1$ 
        //  $M_0$  contains  $X_i - X_j$ , both directions  $X_i \rightarrow X_j$  and  $X_i \leftarrow X_j$  obtained at least one vote.
        set  $M_0(i, j) := -1$ ,  $M_0(j, i) := -1$ ; // Direct  $X_i \leftrightarrow X_j$  in  $M_0$ .
    end for.
(4)  $\mathcal{G}_{output} = \text{Matrix2Graph}^*(M_0)$ . // Decode the resulting matrix  $M_0$  into graph  $\mathcal{G}_{output}$  as output.

```

B.2.2. Procedure ProposeCollider*

Input: (i) A matrix M_{in} with integer entries ≤ 1 representing a partially directed graph;
(ii) A triple (X_a, X_c, X_b) ;
(iii) A matrix L with entries “0” and “1” indicating whether an orientation obtained at least one vote.

Output: matrix M_{out} with integer entries ≤ 1 representing a (partially) directed graph .

```

(1) Initialize  $M_{out} := M_{in}$  and  $\text{ProposeAccepted} := false$ .
(2) Compute  $h_c := \widehat{\mathbb{H}}_{X_a X_b | X_c} / \widehat{\mathbb{H}}_{X_a X_b}$ ,  $h_b := \widehat{\mathbb{H}}_{X_a X_c | X_b} / \widehat{\mathbb{H}}_{X_a X_c}$ ,  $h_a := \widehat{\mathbb{H}}_{X_b X_c | X_a} / \widehat{\mathbb{H}}_{X_b X_c}$ ;
    if  $h_c \geq \max\{h_b, h_a\}$  then  $\text{ProposeAccepted} := true$ , end if; // Collider test.
(3) if  $\text{ProposeAccepted}$  // A vote for  $X_a \rightarrow X_c$  and  $X_b \rightarrow X_c$ , respectively.
    for  $i = a, b$ 
        Case:  $M_{out}(i, c) = 1$  and  $M_{out}(c, i) = 1$  // If  $M_{out}$  contains the undirected edges  $X_{a,b} - X_c$ .
            set  $M_{out}(i, c) := -1$  and  $M_{out}(c, i) := 0$ ; // Orient  $X_{a,b} \rightarrow X_c$ .
        Case:  $M_{out}(i, c) \leq -1$  and  $M_{out}(c, i) = 0$ 
            // New vote coincides with orientation stored in  $M_{out}$ .
            set  $M_{out}(i, c) := M_{out}(i, c) - 1$ ;
            // Leave the orientation untouched and increase counter for  $X_i \rightarrow X_c$  by  $-1$ .
        Case:  $M_{out}(i, c) = 0$  and  $M_{out}(c, i) < -1$ 
            // New vote is opposite to the current orientation stored in  $M_{out}$ .
            set  $M_{out}(c, i) := M_{out}(c, i) + 1$ ;
            // Leave the orientation untouched and decrease counter for  $X_c \rightarrow X_{a,b}$  by  $-1$ .
        Case:  $M_{out}(i, c) = 0$  and  $M_{out}(c, i) = -1$ 

```

B.3. Orientation Rules to Make Graphs Maximally Oriented

```

// If  $M_{out}$  contains the opposite orientation  $X_c \rightarrow X_{a,b}$  with a counter of  $-1$ .
set  $M_{out}(i, c) := 1$  and  $M_{out}(c, i) := 1$ ; // Reset to an undirected edge  $X_{a,b} - X_c$ .
end for;
end if.

```

B.2.3. Procedure Matrix2Graph*

Input: An $N \times N$ matrix M with entries “ -1 ”, “ 0 ”, and “ 1 ”;

Output: A (partially) directed graph \mathcal{G} of X_1, \dots, X_N with edges $-$, \rightarrow , \leftrightarrow ;

(1) Initialize a graph \mathcal{G} with no edges.

(2) **for** $i, j = 1$ to N

Case: $M(i, j) = 1$, and $M(j, i) = 1$, set $X_i - X_j$ in \mathcal{G} ; // Undirected edge.

Case: $M(i, j) = 1$, and $M(j, i) = 0$, set $X_i \rightarrow X_j$ in \mathcal{G} ; // Directed edge.

Case: $M(i, j) = -1$, and $M(j, i) = -1$, set $X_i \leftrightarrow X_j$ in \mathcal{G} ; // Bi-directed edge.

end for.

B.3. Orientation Rules to Make Graphs Maximally Oriented

A partially directed graph is given. The orientation of the given graph is limited to v -structures. Under the assumption that there are no additional v -structures and directed cycles in the structure, rules as shown in Fig. B.1 (see [125], p. 51) are sufficient to make the given partially directed graph maximally oriented, in the sense that all edges that are common to the Markov equivalence class are oriented. Fig. B.2 visualizes these three rules.

Input: A graph G with directed (limited to v -structures) or undirected edges.

While no more edges can be oriented:

Rule 1: For each uncoupled meeting $X \rightarrow Z - Y$ (X and Y nonadjacent), orient $Z - Y$ into $Z \rightarrow Y$.

Rule 2: For each $X - Y$ such that $X \rightarrow Z \rightarrow Y$, orient $X - Y$ into $X \rightarrow Y$.

Rule 3: For each uncoupled meeting $Z_1 - X - Z_2$ (Z_1 and Z_2 nonadjacent) such that $Z_1 \rightarrow Y$, $Z_2 \rightarrow Y$, $X - Y$, orient $X - Y$ into $X \rightarrow Y$.

Output: A graph G with directed or undirected edges.

Figure B.1.: Orientation rules to make a given partially directed (limited to v -structures) graph maximally oriented, under the assumption that there are no additional v -structures and directed cycles in the structure.

B. Appendix



Figure B.2.: A partially directed (limited to v -structures) graph is given. We assume that there are no additional v -structures and directed cycles in the structure. The plots describe three substructures, which can be further oriented by orientation rules 1, 2, and 3 as shown in Fig. B.1).

C. Appendix

C.1. Numerical evidence of power increase of multiple testing

A procedure of multiple testing on a set of associated hypotheses is proposed by Benjamini et al. [19]: the so-called adaptive procedure of FDR control based on independent test statistics. In our setting, we have a set of m subsamples resampled from the original sample. An independence hypothesis is tested on each subsample and provides a set of m p-values. It is clear that the m p-values are highly dependent. To relax the precondition of independent test statistics, Benjamini et al. proposed in [21] a general correction factor for dependent test statistics. However, our experiments showed that this modification is still too conservative for our purpose of structural learning. For this reason, we employ permutation techniques to conduct the independence test. The whole multiple independence testing procedure is summarized in Fig. C.1 with a pre-specified parameter m . In our experiments, we chose $m = 100$. Step 1 runs a multiple test on the original sample. Step 2 runs multiple tests on shuffled samples by k random permutations. If some FDR $q < 0.5$ can be found, Step 3 rejects the independence hypothesis. This procedure is extremely time-consuming, but it increases the power of a statistical test.

To give some intuition how the resampling-based multiple testing procedure works, we consider an example. A dataset of 100,000 data points is sampled from an OR gate with noise $r = 0.3$ (see Fig. 3.16 and Eq. (3.6) for the definition of OR gates) and resample 100 subsamples of size 100 from the original 100,000 data points. Note that the subsamples are nearly independent of each other because of $100 \ll 100,000$. Fig. C.2 illustrates the Q-Q plots of p-values given by testing $X \perp\!\!\!\perp Y$ (left plot) and $X \not\perp\!\!\!\perp Y \mid Z$ (right plot). Here we used the likelihood ratio χ^2 test. The red dots in plots are the p-values $p_1^{(0)} \leq \dots \leq p_{100}^{(0)}$ for 100 subsamples. The lines of various colors present the p-values $p_1^{(i)} \leq \dots \leq p_{100}^{(i)}$ of 10 shuffled data samples. Our test states that, having accepted an FDR of up to 0.5, if one can always reject more hypotheses in the original sample than any of the shuffled samples, the independence hypothesis should be rejected. Graphically, if, in the subfield, the p-value line of original data (red dots in the plots) is more strongly right-skewed than those of shuffled data (lines of various colors in the plots), the independence hypothesis should be rejected, otherwise accepted. In Fig. C.2, the left plot suggests accepting independence, while the right plot suggests rejecting independence.

To give some numerical evidence of power increase of multiple testing, we show experiments with 1000 artificial datasets sampled by noisy OR gates. The same datasets are used in Section 3.3.4 for experiments with single testing. Here, we replace a single χ^2 test by the resampling-based multiple χ^2 tests. The subsamples are obtained by resampling with replacement (5-fold

C. Appendix

Input: An independence hypothesis $X \perp\!\!\!\perp Y \mid Z$ and a data sample (X, Y, Z) .

Step 1 Resample m subsamples from the original sample. For each subsample, conduct the independence hypothesis test and record the p-value. Sort the resulting set of p-values in an increasing order, i.e., $p_1^{(0)} \leq p_2^{(0)} \leq \dots \leq p_m^{(0)}$.

Step 2 Permute the original data randomly to simulate data under independency. Conduct Step 1 for simulated datasets and obtain a set of m p-values. Repeat step 2 for k times (we chose $k=10$) and obtain k sets of p-values $p_1^{(1)} \leq \dots \leq p_m^{(1)}, \dots, p_1^{(k)} \leq \dots \leq p_m^{(k)}$.

Step 3 For a given FDR q , conduct the adaptive procedure as described in [19] p. 71, and calculate the number of rejections $r_q^{(0)}, r_q^{(1)}, \dots, r_q^{(k)}$ for the sets of p-values $p_1^{(0)}, \dots, p_m^{(0)}, p_1^{(1)}, \dots, p_m^{(1)}, \dots, p_1^{(k)}, \dots, p_m^{(k)}$, respectively. If there is some $q \in (0, 0.5)$ that $r_q^{(0)} > \max\{r_q^{(1)}, \dots, r_q^{(k)}\}$, reject the independence hypothesis, otherwise accept the independence hypothesis.

Output: Accepting or rejecting $X \perp\!\!\!\perp Y \mid Z$.

Figure C.1.: Resampling-based multiple independence hypothesis test with random permutations.

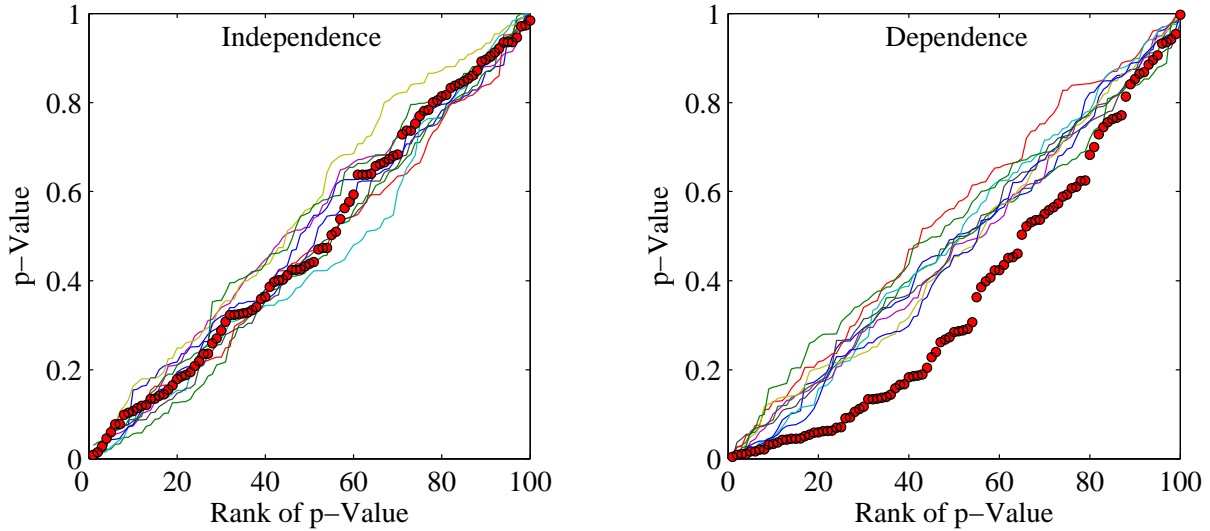


Figure C.2.: Multiple statistical independence hypothesis tests are conducted on noisy OR data. The plots are Q-Q plots of the set of p-values obtained by multiple testing. Red dots visualizes reordered p-values of original data. Lines of various colors represents reordered p-values of simulated data under independency. The left plot indicates $X \perp\!\!\!\perp Y$ and the right plot indicates $X \not\perp\!\!\!\perp Y \mid Z$.

C.1. Numerical evidence of power increase of multiple testing

	Accepting $X \perp Y$									
Original Sample Size	20		50		100		150		200	
χ^2 Test	Single	Multiple	Single	Multiple	Single	Multiple	Single	Multiple	Single	Multiple
$r = 0$	94.5	90.8	94.9	92.1	93.9	92.0	94.7	92.8	95.7	93.3
$r = 0.1$	92.7	90.0	93.8	92.3	94.2	92.4	93.7	91.0	95.3	92.9
$r = 0.2$	93.7	89.4	95.2	92.4	94.8	92.4	94.8	91.5	95.8	94.5
$r = 0.3$	93.4	89.5	94.2	91.7	93.8	90.6	96.0	93.3	94.2	91.7
Noisy OR	Rejecting $X \perp Y Z$									
$r = 0$	25.2	71.2	94.4	100	100	100	100	100	100	100
$r = 0.1$	23.6	54.1	57.0	72.6	87.5	92.8	96.6	97.9	99.2	99.7
$r = 0.2$	15.4	40.2	23.6	38.3	42.2	58.3	61.6	61.5	70.9	81.0
$r = 0.3$	11.5	29.8	11.4	20.8	13.8	23.7	16.9	27.7	22.1	35.7

Table C.1.: Numerical comparison of a single χ^2 test is replaced by resampling-based multiple χ^2 tests on discrete domains. The multiple tests use the subsample size of 5-fold of the original sample size. The generating models are noisy OR gates with 4 different noise levels $r = 0, 0.1, 0.2, 0.3$ as shown in Fig. 3.16 and Eq. (3.6). The entries show how often (in percentage) the constraint $X \perp Y$ or $X \not\perp Y | Z$ is verified after 1000 replications.

the original sample size) from the original sample.

The results of experiments with the χ^2 test on discrete domains show that a multiple test makes slightly more type I errors (upper half of Tab. C.1), but significantly less type II errors than a single test (lower half of Tab. C.1). The improvement is most impressive at small samples with less noise. When a single test is conducted, a strict control of type I error is achieved at the cost of an increase of type II error. In contrast, a multiple test benefits from the readiness to assume a bit more risk of making I errors and can keep the type II error to a lower level.

To show the power of a multiple test on continuous domains, we apply the single, multiple kernel test of independence to the Meander data (see Fig. 3.7 for the generating model). A multiple kernel test can further significantly improve the performance of a single kernel test, in the sense that type I error is strongly reduced without an increase of type II error (Tab. C.2). It is noteworthy that the resampling process on continuous domains might be more natural, if some noises on the original data points were incorporated.

Based on results of the simulation, we expect that the multiple testing procedure offers greater power than a single test and can keep both type I and II error to a relatively low and well-balanced level. Unfortunately, this procedure is extremely time-consuming in practice.

C. Appendix

Kernel Test	Single	Multiple
Rejecting $X \not\perp Y$	99.9	100
Accepting $X \perp Y Z$	34.8	99.2

Table C.2.: Numerical comparison of single and multiple kernel independence tests on continuous domains (see Fig. 3.7 for the generating model). The entries show how often (in percentage) the constraint $X \not\perp Y$ or $X \perp Y | Z$ is verified after 1000 replications.

C.2. Comparison of learning algorithms on categorical domains

Apart from the constraint-based PC [153] and BN-PC [28] algorithm there are other algorithms that could be used for causal learning, particularly on purely discrete domains. One large class is score-based Bayesian approaches proposed and described by [84, 86, 39, 118, 176]. It should be emphasized that algorithms for finding Bayesian networks are not necessarily developed for the purpose of modeling causal relationships. The goal is often merely to represent the dependence by simple structures.

The methods that we have tested are: conventional PC, information-theory-based BN-PC, Bayesian approaches using BDe (Bayesian Dirichlet equivalent) metric via exhaustive search, Greedy Search/Hill-climbing [32], and MCMC (Markov Chain Monte Carlo) [89]. We study here networks containing only 3–4 variables, which cause the search space of DAGs to be reasonably small. Exhaustive search is then tractable, allowing the computation of the posterior probability for all the DAGs. Consequently the global optimum can be determined. Other search methods do not guarantee to find the global optimum but are much more efficient. The well-known K2 [40] can actually not be used to find the causal structure, since an initial causal ordering of variables must already be given. K2 is then only able to decide which arrows can be dropped without violating the Markov condition. Heckerman et al. [85] proposed to apply the maximum weight spanning tree algorithm (MWST) [35] to initialize K2. We call it “MWST+K2”. Note that an initial order can also optionally be specified for greedy search. We call this combination “MWST+Greedy Search”. All these methods are summarized and implemented by Murphy, Leray and Francois.¹, respectively.

Since most of these algorithms are limited to discrete domains, we restrict our comparison to datasets (sample size 200) generated by 2/3-bits deterministic/noisy OR gates (see Tab. 5.2 for parameters). Tab. C.3 and Tab. C.4 collect the statistics of structures detected by all aforementioned algorithms after 1000 replications. The entries are percentages of detected arcs between two variables (X_i, X_j) within rows. For (X_i, X_j) , “• •” depicts the absence of an edge between X_i and X_j ; “• – •” depicts a present but undirected edge between them; “• → •” and “• ← •” denote “ $X_i \rightarrow X_j$ ” and “ $X_i \leftarrow X_j$ ”, respectively.

¹The BayesNet Toolbox and the BNT Structure Learning Package are online available at <http://bnt.sourceforge.net> and <http://banquiseasi.insa-rouen.fr/projects/bnt-slp>

C.3. Statistics of experiments with Asia Network

	2-Bit-IndDet				2-Bit-IndPro				2-Bit-DepPro			
	• •	•→•	•←•	•—•	• •	•→•	•←•	•—•	• •	•→•	•←•	•—•
True Model	100	0	0	0	100	0	0	0	0	100	0	0
	0	100	0	0	0	100	0	0	0	100	0	0
	0	100	0	0	0	100	0	0	0	100	0	0
OPB	96.7	0.2	0	3.1	96.7	0	0.4	2.9	12.6	0.1	0	87.3
	0	99.7	0	0.3	0	95.2	0.4	4.4	0	98.1	0	1.9
	0	99.7	0.2	0.1	0	95.2	0.2	4.8	0	98.1	0.1	1.8
PC	93.9	0	0	6.1	96.5	0	0	3.5	96.8	0	0	3.2
	0	93.9	0	6.1	0	94.1	0	5.9	0	94.2	0	5.8
	0	93.9	0	6.1	0	94.1	0	5.9	0	94.2	0	5.8
BN-PC	93.7	0	6.3	0	96.3	0	3.7	0	72.0	0.1	27.9	0
	0	93.7	6.3	0	0	82.2	17.8	0	0.1	71.4	28.5	0
	0	93.7	6.3	0	0	82.2	17.8	0	0	71.4	28.6	0
Exhaustive Search	98.5	0.4	0.5	0.6	99.3	0.1	0.2	0.4	99.4	0.1	0.2	0.3
	0	99.2	0.2	0.6	0	89.0	6.8	4.2	0	88.9	6.5	4.6
	0	99.1	0.2	0.7	0	91.7	3.5	4.8	0	91.0	4.3	4.7
Greedy Search	69.1	16.2	14.7	0	81.5	10.4	8.1	0	80.6	10.6	8.8	0
	0	83.1	16.9	0	0	68.4	31.6	0	0	66.5	31.5	0
	0	82.7	17.3	0	0	70.0	30.0	0	0	68.1	31.9	0
MWST+Greedy Search	97.2	2.5	0.3	0	98.7	1.3	0	0	98.9	1.1	0	0
	0	99.0	1.0	0	0	94.6	5.4	0	0	94.1	5.9	0
	0	97.9	2.1	0	0	89.4	10.6	0	0	88.2	11.8	0
MWST+K2	0	100	0	0	39.7	60.3	0	0	40.6	59.4	0	0
	0	100	0	0	0	100	0	0	0	100	0	0
	0	0	100	0	0	0	100	0	0	0	100	0
MCMC	69.1	16.2	14.7	0	77.9	11.3	10.8	0	77.4	11.1	10.5	0
	0	83.1	16.9	0	0	75.9	24.1	0	0	75.0	25.0	0
	0	82.7	17.3	0	0	74.1	25.9	0	0	74.4	25.6	0

Table C.3.: The underlying true model: 2-bit OR gates (first row, see Tab. 5.2 for parameters) and the structures generated by different algorithms (rows 2 to 9). 200 data points are sampled from each model. The entries are percentages of 1000 replications having the considered patterns (“• •”: no edge; “•—•”: an undirected edge; “•→•” or “•←•”: a directed edge) as output.

C.3. Statistics of experiments with Asia Network

For the sake of notational convenience, we denote X_1 : ASIA, X_2 : TUB, X_3 : SMOKING, X_4 : LUNG, X_5 : BRONCHITIS, X_6 : TUB/LUNG, X_7 : X-RAY, X_8 : DYSPNOEA in the following two tables. Tab. C.5 summarizes how often an arrow is detected by OPB after 1000 replications, given the true skeleton (Fig. 5.6, leftmost). Tab. C.6 summarizes how often an arrow is detected by OPA+K2 (K2 with a initial causal order detected by OPA) after 1000 replications.

C. Appendix

	3-Bit-IndDet				3-Bit-IndPro				3-Bit-DepPro			
	• •	•→•	•←•	•—•	• •	•→•	•←•	•—•	• •	•→•	•←•	•—•
True Model	100	0	0	0	100	0	0	0	0	100	0	0
	100	0	0	0	100	0	0	0	0	100	0	0
	0	100	0	0	0	100	0	0	0	100	0	0
	100	0	0	0	100	0	0	0	0	100	0	0
	0	100	0	0	0	100	0	0	0	100	0	0
0	100	0	0	0	100	0	0	0	100	0	0	
OPB	74.0	8.0	2.8	15.2	71.8	5.0	9.7	13.5	11.1	2.2	2.7	84.0
	76.8	8.2	2.7	12.3	74.3	4.9	12.4	8.4	1.5	51.3	19.4	27.8
	0	98.6	0.5	0.9	0.1	91.0	7.1	1.8	1.1	47.2	25.4	26.3
	76.8	4.7	3.9	14.6	71.9	4.8	11.5	11.8	1.5	67.0	10.1	21.4
	0	96.0	3.0	1.0	0	93.8	4.4	1.8	0.2	62.5	12.6	24.7
0	94.8	3.9	1.3	0	96.4	1.2	2.4	0.2	1.8	2.4	95.6	
PC	98.5	0	0	1.5	96.8	0.5	0	2.7	69.8	12.2	5.9	12.1
	98.1	0	0	1.9	98.5	0.2	0	1.3	29.3	47.2	6.9	16.6
	0	100	0	0	3.7	96.2	0	0.1	18.7	55.8	6.0	19.5
	97.4	0	0	2.6	97.1	0.1	0	2.8	20.4	54.9	10.5	14.2
	0	100	0	0	0.9	99.0	0.1	0	7.9	61.7	10.5	19.9
0	99.8	0.2	0	0.3	99.5	0.2	0	5.4	11.5	23.3	59.8	
BN-PC	97.5	0.4	2.1	0	96.9	0.7	2.4	0	71.6	9.9	18.5	0
	97.8	0.6	1.6	0	97.0	0.3	2.7	0	28.5	23.8	47.7	0
	0	74.2	25.8	0	0.6	31.3	68.1	0	20.2	27.0	52.8	0
	96.8	0.7	2.5	0	97.6	0.2	2.2	0	19.3	29.4	51.3	0
	0	65.9	34.1	0	0.8	39.8	59.4	0	8.1	32.4	59.5	0
0	48.7	51.3	0	0	47.2	52.8	0	4.9	12.6	82.5	0	
Exhaustive Search	99.3	0.1	0.2	0.4	98.8	0.2	0.5	0.5	74.6	10.7	11.2	3.5
	99.2	0	0.2	0.6	99.4	0.5	0.1	0	35.6	49.8	10.3	4.3
	0	100	0	0	2.8	49.5	40.0	7.7	24.3	62.2	8.6	4.9
	98.8	0.5	0.4	0.3	98.7	0.5	0.7	0.1	34.3	54.0	8.5	3.2
	0	100	0	0	0.9	60.5	30.3	8.3	14.3	69.3	10.7	5.7
0	100	0	0	0.4	61.1	30.6	7.9	11.8	30.2	46.5	11.5	
Greedy Search	90.1	4.4	5.5	0	97.5	1.3	1.2	0	25.3	36.0	38.7	0
	93.1	2.7	4.2	0	97.4	1.8	0.8	0	50.9	17.2	31.9	0
	0	75.2	24.8	0	2.5	32.4	65.1	0	33.7	26.2	40.1	0
	93.1	3.1	3.8	0	95.6	2.0	2.4	0	44.9	19.5	35.6	0
	0	69.0	31.0	0	1.0	42.4	56.6	0	25.8	25.8	48.4	0
0	64.5	35.5	0	0.5	47.2	52.3	0	1.3	41.2	57.5	0	
MWST+Greedy Search	97.4	2.5	0.1	0	98.7	0.9	0.4	0	43.8	46.4	9.8	0
	97.7	2.0	0.3	0	99.3	0.5	0.2	0	50.5	37.5	12.0	0
	0	94.4	5.6	0	2.5	68.5	29.0	0	33.9	50.1	16.0	0
	94.8	2.5	2.7	0	91.7	4.0	4.3	0	44.8	24.2	31.0	0
	0	82.9	17.1	0	0.9	24.5	74.6	0	25.6	39.8	34.6	0
0	78.1	21.9	0	0.5	32.2	67.3	0	1.1	41.8	57.1	0	
MWST+K2	34.1	65.9	0	0	96.4	3.6	0	0	9.8	90.2	0	0
	66.3	33.7	0	0	95.9	4.1	0	0	44.6	55.4	0	0
	0	100	0	0	3.0	97.0	16.0	0	26.1	73.9	16.0	0
	34.9	0.1	65.0	0	92.4	0.1	7.5	0	38.0	0.1	61.9	0
	0	0	100	0	0.9	0.7	98.4	0	17.4	0.2	82.4	0
0	0	100	0	0.5	0.2	99.3	0	0	41.6	58.4	0	
MCMC	86.8	6.2	7.0	0	91.7	3.4	4.9	0	37.4	29.8	32.8	0
	86.0	6.7	7.3	0	90.8	3.5	5.7	0	31.5	42.7	25.8	0
	0	99.6	0.4	0	5.8	43.8	50.4	0	22.9	46.6	30.5	0
	86.7	6.0	7.3	0	89.5	5.4	5.1	0	31.8	42.6	25.6	0
	0	99.3	0.7	0	1.9	48.8	49.3	0	13.7	52.1	34.2	0
0	99.4	0.6	0	0.7	51.3	48.0	0	9.8	38.6	51.6	0	

Table C.4.: The underlying true model: 3-bit OR gates (first row, see Tab. 5.2 for parameters) and the structures generated by different algorithms (rows 2 to 9). 200 data points are sampled from each model. The entries are percentages of 1000 replications having the considered patterns (“• •”: no edge; “•—•”: an undirected edge; “•→•” or “•←•”: a directed edge) as output.

C.3. Statistics of experiments with Asia Network

The orientation procedure A (OPA as in Fig. 5.3) is applied to the fully connected skeleton.								
Variable Pair	(X_1, X_2)	(X_2, X_6)	(X_3, X_4)	(X_3, X_5)	(X_4, X_6)	(X_5, X_8)	(X_6, X_7)	(X_6, X_8)
Correct Orientation	●→●	●→●	●→●	●→●	●→●	●→●	●→●	●→●
●→●	16.4	95.8	6.6	20.2	94.3	88.7	75.5	95.5
●←●	65.0	2.0	72.2	56.4	0.7	5.7	13.9	3.7
●—●	18.6	2.2	21.2	23.4	5.0	5.6	10.6	0.8
The orientation procedure B (OPB as in Fig. 5.4) is applied to the true skeleton.								
●→●	0.2	97.7	15.7	0.4	97.2	77.7	93.6	92.6
●←●	0.1	0.1	29.3	44.6	0.6	16.6	4.2	6.5
●—●	99.7	2.2	55.0	55.0	2.2	5.7	2.2	0.9

Table C.5.: Statistics of arcs detected by OPA and OPB. 400 data points are sampled from the Asia network. The entries are percentages of 1000 replications having the considered patterns (“● ●”: no edge; “●—●”: an undirected edge; “●→●” or “●←●”: a directed edge) as output.

C. Appendix

Variable Pair	(X_1, X_2)	(X_1, X_3)	(X_1, X_4)	(X_1, X_5)	(X_1, X_6)	(X_1, X_7)	(X_1, X_8)
Correct Pattern	$\bullet \rightarrow \bullet$	$\bullet \bullet$	$\bullet \bullet$	$\bullet \bullet$	$\bullet \bullet$	$\bullet \bullet$	$\bullet \bullet$
$\bullet \bullet$	81.0	94.9	97.1	89.2	96.9	95.4	97.6
$\bullet \rightarrow \bullet$	12.4	3.8	1.9	6.9	2.8	4.3	2.3
$\bullet \leftarrow \bullet$	6.6	1.3	1.0	3.9	0.3	0.3	0.1
$\bullet - \bullet$	0	0	0	0	0	0	0
Variable Pair	(X_2, X_3)	(X_2, X_4)	(X_2, X_5)	(X_2, X_6)	(X_2, X_7)	(X_2, X_8)	(X_3, X_4)
Correct Pattern	$\bullet \bullet$	$\bullet \bullet$	$\bullet \bullet$	$\bullet \rightarrow \bullet$	$\bullet \bullet$	$\bullet \bullet$	$\bullet \rightarrow \bullet$
$\bullet \bullet$	96.7	92.8	90.1	0	78.4	94.1	23.3
$\bullet \rightarrow \bullet$	2.7	4.8	9.1	99.1	19.2	5.7	4.7
$\bullet \leftarrow \bullet$	0.6	2.4	0.8	0.9	2.4	0.2	72.0
$\bullet - \bullet$	0	0	0	0	0	0	0
Variable Pair	(X_3, X_5)	(X_3, X_6)	(X_3, X_7)	(X_3, X_8)	(X_4, X_5)	(X_4, X_6)	(X_4, X_7)
Correct Pattern	$\bullet \rightarrow \bullet$	$\bullet \bullet$	$\bullet \bullet$	$\bullet \bullet$	$\bullet \bullet$	$\bullet \rightarrow \bullet$	$\bullet \bullet$
$\bullet \bullet$	3.5	85.4	99.1	94.4	93.2	0	76.1
$\bullet \rightarrow \bullet$	26.2	0.1	0	2.5	6.0	98.7	19.5
$\bullet \leftarrow \bullet$	70.3	14.5	0.9	3.1	0.8	1.3	4.4
$\bullet - \bullet$	0	0	0	0	0	0	0
Variable Pair	(X_4, X_8)	(X_5, X_6)	(X_5, X_7)	(X_5, X_8)	(X_6, X_7)	(X_6, X_8)	(X_7, X_8)
Correct Pattern	$\bullet \bullet$	$\bullet \bullet$	$\bullet \bullet$	$\bullet \rightarrow \bullet$	$\bullet \rightarrow \bullet$	$\bullet \rightarrow \bullet$	$\bullet \bullet$
$\bullet \bullet$	80.6	98.0	99.0	0	20.3	30.2	98.0
$\bullet \rightarrow \bullet$	18.5	0	0.4	92.0	77.5	69.8	2.0
$\bullet \leftarrow \bullet$	0.9	2.0	0.6	8.0	2.2	0	0
$\bullet - \bullet$	0	0	0	0	0	0	0

Table C.6.: Statistics of detected arrows by OPA+K2. 400 data points are sampled from the Asia network. The entries are percentages of 1000 replications having the considered patterns (“ $\bullet \bullet$ ”: no edge; “ $\bullet - \bullet$ ”: an undirected edge; “ $\bullet \rightarrow \bullet$ ” or “ $\bullet \leftarrow \bullet$ ”: a directed edge) as output.

Bibliography

- [1] D. Achlioptas, F. McSherry, and B. Schölkopf. Sampling techniques for kernel methods. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Proceedings of the 14th Neural Information Processing Systems Conference*, pages 335–342, Cambridge, MA, 2001. MIT Press.
- [2] M. Adams and Z. Jia. Structural and biochemical analysis reveal pirins to possess quercetinase activity. *Journal of Biological Chemistry*, 280(31):28675–28682, 2005.
- [3] A. Aizerman, E. Braverman, and L. Rozoner. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25:821–837, 1964.
- [4] A. Allahverdyan and D. Janzing. Relating the thermodynamic arrow of time to the causal arrow. <http://www.citebase.org/abstract?id=oai:arXiv.org:0708.1175>, 8 2007.
- [5] D. G. Altman. *Practical statistics for medical research*. Chapman and Hall, London, 1991.
- [6] Y. Altun and A. Smola. Unifying divergence minimization and statistical inference via convex duality. In G. Lugosi and H. Simon, editors, *Proceedings of the 19th Annual Conference on Learning Theory*, pages 139–153, Pittsburgh, PA, 2006.
- [7] S. Amari. Information geometry on hierarchy of probability distributions. *IEEE Transactions on Information Theory*, 47(5):1701–1711, 2001.
- [8] S. Andersson, D. Madigan, and M. Perlman. On the Markov equivalence classes for chain graphs, undirected graphs and acyclic digraphs. *Scandinavian Journal of Statistics*, 24(1):81–102, 1997.
- [9] B. Arnold, E. Castillo, and J. Sarabia. Conditionally specified distributions: An introduction. *Statistical Science*, 16(3):249–274, 2001.
- [10] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.
- [11] F. Bach and M. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2002.
- [12] C. Baker. Joint measures and cross-covariance operators. *Transactions of the American Mathematical Society*, 186:273–289, 1973.

Bibliography

- [13] I. Barlow, G. Lloyd, E. Ramshaw, A. Miller, et al. Correlations and changes in flavour and chemical parameters of Cheddar cheeses during maturation. *The Australian Journal of Dairy Technology*, 44:7–18, 1989.
- [14] D. Basu. On statistics independent of a complete sufficient statistics. *The Indian Journal of Statistics (Sankhyá)*, 15(4):377–380, 1955.
- [15] D. Basu. On statistics independent of sufficient statistics. *The Indian Journal of Statistics (Sankhyá)*, 20(3-4):223–226, 1958.
- [16] L. Baugh, A. Hill, D. Slonim, E. Brown, et al. Composition and dynamics of the *Caenorhabditis elegans* early embryonic transcriptome. *Development*, 130:889–900, 2003.
- [17] S. Bay, L. Chrisman, A. Pohorille, and J. Shrager. Temporal aggregation bias and inference of causal regulatory networks. *Journal of Computational Biology*, 11(5):971–985, 2004.
- [18] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 57(1):289–300, 1995.
- [19] Y. Benjamini and Y. Hochberg. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics*, 25(1):60–83, 2000.
- [20] Y. Benjamini and Abba M. Krieger D. Yekutieli. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, 93(3):491–507, 2006.
- [21] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001.
- [22] A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Kluwer Academic Publishers, Norwell, MA, 2004.
- [23] M. Bittner, P. Meltzer, Y. Chen, Y. Jiang, et al. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, 406:536–540, 2000.
- [24] K. Borgwardt, A. Gretton, M. Rasch, H. Kriegel, et al. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006.
- [25] S. Boyd. *Convex optimization*. Cambridge University Press, Cambridge, UK, 2004.
- [26] S. Canu and A. Smola. Kernel methods and the exponential family. *Neurocomputing*, 69(7-9):714–720, 2006.
- [27] N. Cartwright. *Nature's capacities and their measurement*. Clarendon Press, Oxford, 1989.

- [28] J. Cheng, R. Greiner, J. Kelly, D. Bell, et al. Learning Bayesian networks from data: An information-theory based approach. *Artificial Intelligence Journal*, 137:43–90, 2002.
- [29] P. Cheng. From covariation to causation: A causal power theory. *Psychological Review*, 104:367–405, 1997.
- [30] D. Chickering. Learning Bayesian networks is NP-complete. In D. Fisher and H. Lenz, editors, *Learning from data: Artificial intelligence and statistics V*, pages 121–130. Springer Verlag, New York, NY, 1996.
- [31] D. Chickering. The WinMine toolkit. Technical Report MSR-TR-2002-103, Microsoft, Redmond, WA, 2002.
- [32] D. Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2003.
- [33] D. Chickering, D. Heckerman, and C. Meek. Large-sample learning of Bayesian networks is NP-hard. *Journal of Machine Learning Research*, 5:1287–1330, 2004.
- [34] D. Chickering and C. Meek. On the incompatibility of faithfulness and monotone DAG faithfulness. *Journal of Artificial Intelligence*, 170(8):653–666, 2006.
- [35] C. Chow and C. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on the Information Theory*, 14(3):462–467, 1968.
- [36] R. Collins and A. Wragg. Maximum entropy histograms. *Journal of Physics A: Mathematical and General*, 10(9):1441–1464, 1977.
- [37] J. Comley and D. Dowe. Minimum message length and generalised Bayesian nets with asymmetric languages. In P. Grünwald, I. Myung, and M. Pitt, editors, *Advances in Minimum Description Length: Theory and Applications*, pages 265–294. MIT Press, Cambridge, MA, 2005.
- [38] G. Cooper. The computational complexity of probabilistic inference using Bayesian belief networks. *Journal of Artificial Intelligence*, 42(3–4):393–405, 1990.
- [39] G. Cooper. An overview of the representation and discovery of causal relationships using Bayesian networks. In C. Glymour and G. Cooper, editors, *Computation, causation, and discovery*, pages 3–62. MIT Press, Cambridge, MA, 1999.
- [40] G. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9(4):309–347, 1992.
- [41] R. Cowell. Conditions under which conditional independence and scoring methods lead to identical selection of Bayesian network models. In J. Breese and D. Koller, editors, *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence*, pages 91–97, San Francisco, CA, 2001. Morgan Kaufmann Publishers.

Bibliography

- [42] S. Curran and G. Murray. Matrix metalloproteinases: molecular aspects of their roles in tumour invasion and metastasis. *European Journal of Cancer*, 36(13):1621–1630, 2000.
- [43] S. Dasgupta. Learning mixtures of Gaussians. In *Proceedings of the 40th Annual IEEE Symposium on Foundations of Computer Science*, pages 634–644, Washington, DC, 1999. IEEE Computer Society.
- [44] D. Dash and M. Druzdzal. A hybrid anytime algorithm for the construction of causal models from sparse data. In *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, pages 142–149, San Francisco, CA, 1999. Morgan Kaufmann Publishers.
- [45] A. Datta, A. Choudhary, M. Bittner, and E. Dougherty. External control in Markovian genetic regulatory networks. *Machine Learning*, 52:169–191, 2003.
- [46] A. Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society B*, 41(1):1–31, 1979.
- [47] L. Devroye and G. Lugosi. *Combinatorial methods in density estimation*. Springer Verlag, New York, NY, 2001.
- [48] D. Dowson and A. Wragg. Maximum-entropy distributions having prescribed first and second moments. *IEEE Transactions on Information Theory*, 19(5):689–693, 1973.
- [49] B. Draper, C. Mello, B. Bowerman, J. Hardin, et al. MEX-3 is a KH domain protein that regulates blastomere identity in early *Caenorhabditis elegans* embryos. *Cell*, 87(2):205–216, 1996.
- [50] P. Drineas, R. Kannan, and M. Mahoney. Fast Monte Carlo algorithms for matrices I: Approximating matrix multiplication. Technical Report YALEU/DCS/TR-1269, Yale University, 2004.
- [51] R. Dudley. *Real analysis and probability*. Cambridge University Press, Cambridge, UK, 2nd edition, 2002.
- [52] D. Edwards. *Introduction to graphical modelling*. Springer Verlag, New York, NY, 2000.
- [53] T. Eiter and T. Lukasiewicz. Complexity results for structure-based causality. *Artificial Intelligence*, 142:53–89, 2002.
- [54] J. Feldman, R. O'Donnell, and R. Servedio. Learning mixtures of product distributions over discrete domains. In *Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science*, pages 501–510, 2005.
- [55] S. Fine and K. Scheinberg. Efficient SVM training using low-rank kernel representations. *Journal of Machine Learning Research*, 2:243–264, 2001.
- [56] R. Fisher. The distribution of the partial correlation coefficient. *Metron*, 3(3-4):329–332, 1924.

- [57] J. Florens, M. Mouchart, and J. Rolin. *Elements of Bayesian Statistics*. Marcel Dekker, New York, NY, 1990.
- [58] J. Fraumeni. Cigarette smoking and cancers of the urinary tract: Geographic variations in the United States. *Journal of the National Cancer Institute*, 41:1205–1211, 1970.
- [59] N. Friedman and D. Koller. Being Bayesian about network structure: A Bayesian approach to structure discovery in Bayesian networks. *Machine Learning*, 50(1–2):95–125, 2003.
- [60] K. Fukumizu, F. Bach, and A. Gretton. Statistical consistency of kernel canonical correlation analysis. *Journal of Machine Learning Research*, 8:361–383, 2007.
- [61] K. Fukumizu, F. Bach, and M. Jordan. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research*, 5:73–99, 2004.
- [62] K. Fukumizu, F. Bach, and M. Jordan. Kernel dimension reduction in regression. Technical Report 715, Department of Statistics, University of California, Berkeley, CA, 2006.
- [63] K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Proceedings of the 21th Neural Information Processing Systems Conference*, Cambridge, MA, 2007. MIT Press. 489–496.
- [64] L. García. Controlling the false discovery rate in ecological research. *TRENDS in Ecology and Evolution*, 18(11):553–554, 2003.
- [65] L. García. Escaping the Bonferroni iron claw in ecological studies. *OIKOS*, 105(3):657–663, 2004.
- [66] D. Geiger. *Graphoids: A qualitative framework for probabilistic inference*. PhD thesis, Cognitive Systems Laboratory, Department of Computer Science, University of California, Los Angeles, CA, 1990.
- [67] D. Geiger, T. Verma, and J. Pearl. Identifying independence in Bayesian networks. *Networks*, 20:507–534, 1990.
- [68] S. Gillispie and C. Lemieux. Enumerating Markov equivalence classes of acyclic digraph models. In *Proceedings of the 17th Annual Conference on Uncertainty in Artificial Intelligence*, pages 171–177, San Francisco, CA, 2001. Morgan Kaufmann Publishers.
- [69] F. Girosi. An equivalence between sparse approximation and support vector machines. *Neural Computation*, 10(6):1455–1480, 1998.
- [70] C. Glymour. *The mind’s arrows: Bayes nets and graphical causal models in psychology*. MIT press, Cambridge, MA, 2001.

Bibliography

- [71] P. Good. *Permutation tests: A practical guide to resampling methods for testing hypothesis*. Birkhäuser, Boston, MA, 1994.
- [72] C. Granger. Investigating causal relations by econometric and cross-spectral methods. *Econometrica*, 37:424–438, 1969.
- [73] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *Proceedings of the 16th Conference on Algorithmic Learning Theory*, pages 63–77, Berlin, 2005. Springer Verlag.
- [74] A. Gretton, K. Fukumizu, C. Teo, L. Song, et al. A kernel statistical test of independence. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Proceedings of the 21th Neural Information Processing Systems Conference*, Cambridge, MA, 2007. MIT Press. to appear.
- [75] A. Gretton, R. Herbrich, A. Smola, O. Bousquet, et al. Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6:2075–2129, 2005.
- [76] A. Gretton, A. Smola, O. Bousquet, R. Herbrich, et al. Kernel constrained covariance for dependence measurement. In R.G. Cowell and Z. Ghahramani, editors, *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics*, pages 112–119. Society for Artificial Intelligence and Statistics, 2005.
- [77] C. Groetsch. *The theory of Tikhonov regularization for Fredholm equations of the first kind*. Pitman Publishing Program, Boston, MA, 1984.
- [78] SOEP Group. The German Socio-Economic Panel (GSOEP) after more than 15 years - Overview. In E. Holst, D. Lillard, and T. DiPrete, editors, *Proceedings of the 2000 4th International Conference of German Socio-Economic Panel Study Users (GSOEP2000)*, *Vierteljahrshefte zur Wirtschaftsforschung*, volume 70(1), pages 7–14, Berlin, 2001. Duncker und Humblot.
- [79] C. Gu, D. Bates, Z. Chen, and G. Wahba. The computation of GCV function through householder tridiagonalization with application to the fitting of interaction spline models. *SIAM Journal on Matrix Analysis*, 10:457–480, 1990.
- [80] A. Gualtierotti. On cross-covariance operations. *SIAM Journal on Applied Mathematics*, 37(2):325–329, 1979.
- [81] I. Guyon, C. Aliferis, and A. Elisseeff. Causal feature selection. In H. Liu and H. Motoda, editors, *Computational methods of feature selection*, volume 2 of *Data Mining and Knowledge Discovery Series*. Chapman and Hall, London, 2007.
- [82] H. Halkin, L. Sheiner, C. Peck, and K. Melmon. Determinants of the renal clearance of digoxin. *Clinical Pharmacology and Therapeutics*, 17(4):385–394, 1975.
- [83] F. Harary and E. Palmer. *Graphical enumeration*. Academic Press, New York, NY, 1973.

- [84] D. Heckerman and J. Breese. Causal independence for probability assessment and inference using Bayesian networks. In *IEEE Transactions on Systems, Man, and Cybernetics*, 26, pages 826–831, 1996.
- [85] D. Heckerman, D. Geiger, and D. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. In *Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence*, pages 293–301, San Francisco, CA, 1994. Morgan Kaufmann Publishers.
- [86] D. Heckerman, C. Meek, and G. Cooper. A Bayesian approach to causal discovery. In C. Glymour and G. Cooper, editors, *Computation, Causation, and Discovery*, pages 141–165, Cambridge, MA, 1999. MIT Press.
- [87] M. Hénon. A two-dimensional mapping with a strange attractor. *Communications in Mathematical Physics*, 50:69–77, 1976.
- [88] M. Henrion. Some practical issues in constructing belief networks. In *Proceedings of the 3rd Conference on Uncertainty in Artificial Intelligence*, pages 161–173. Elsevier Science Publishers, 1987.
- [89] E. Herskovits. *Computer-based probabilistic network construction*. PhD thesis, Medical Information Sciences, Stanford University, Stanford, CA, 1991.
- [90] Y. Hochberg. More powerful procedures for multiple significance testing. *Statistics in Medicine*, 9(7):811–818, 1990.
- [91] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.
- [92] A. Hoerl and R. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.
- [93] D. Janzing. On causally asymmetric versions of Occam’s Razor and their relation to thermodynamics. <http://www.citebase.org/abstract?id=oai:arXiv.org:0708.3411>, 8 2007.
- [94] S. Jegelka and A. Gretton. Brisk kernel independent component analysis. In L. Bottou, O. Chapelle, D. DeCoste, and J. Weston, editors, *Large Scale Kernel Machines*, pages 225–250. MIT Press, Cambridge, MA, 2007.
- [95] R. Jelliffe and D. Blankenhorn. Improved method of digitalis therapy in patients with reduced renal function. *Circulation*, 35:11–150, 1967.
- [96] Y. Kano and S. Shimizu. Causal inference using nonnormality. In *Proceedings of the International Symposium on Science of Modeling, the 30th Anniversary of the Information Criterion*, pages 261–270, Tokyo, Japan, 2003.

Bibliography

- [97] S. Kim, H. Li, E. Dougherty, N. Chao, et al. Can Markov chain models mimic biological regulation? *Biological Systems*, 10(4):337–357, 2002.
- [98] S. Lauritzen. *Graphical models*. Oxford Statistical Science Series. Oxford University Press, Oxford, 1996.
- [99] S. Lauritzen, A. Dawid, and D. Spiegelhalter. Independence properties of directed Markov fields. *Networks*, 20(5):491–505, 1990.
- [100] S. Lauritzen and D. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 50(2):157–224, 1988.
- [101] P. Leray and O. Francois. BNT structure learning package: Documentation and experiments. Technical Report FRE CNRS 2645, Laboratoire PSI, Université et INSA de Rouen, 2004.
- [102] M. Lukić and J. Beder. Stochastic processes with sample paths in reproducing kernel Hilbert spaces. *Transactions of the American Mathematical Society*, 353(10):3945–3969, 2001.
- [103] W. Luo. Learning Bayesian networks in semi-deterministic systems. In *Proceedings of the 19th Conference of the Canadian Society for Computational Studies of Intelligence*, pages 230–241, Québec City, Québec, Canada, 2006.
- [104] M. Maduro. Endomesoderm specification in *Caenorhabditis elegans* and other nematodes. *Bioessays*, 28(10):1010–1022, 2006.
- [105] M. Maduro and J. Rothman. Making worm guts: The gene regulatory network of the *Caenorhabditis elegans* endoderm. *Developmental Biology*, 246:68–85, 2002.
- [106] K. Mardia, J. Kent, and J. Bibby. *Multivariate analysis*. Academic Press, New York, NY, 1979.
- [107] D. Margaritis. A Bayesian multiresolution independence test for continuous variables. In *Proceedings of the 17th conference on uncertainty in artificial intelligence*, pages 346–353, Pittsburgh, PA, 2001.
- [108] D. Margaritis. Distribution-free learning of graphical model structure in continuous domains. Technical Report TR-ISU-CS-04-06, Computer Science, Iowa State University, 2004.
- [109] D. Margaritis. Distribution-free learning of Bayesian network structure in continuous domains. In *Proceedings of the 20th National Conference on Artificial Intelligence*, pages 825–830, Seattle, WA, 2005.

- [110] E. Martín. Ignorable common information, null sets and Basu's first theorem. *The Indian Journal of Statistics*, 67(4):674–698, 2005.
- [111] B. McKay, G. Royle, I. Wanless, F. Oggier, et al. Acyclic digraphs and eigenvalues of (0,1)-matrices. *Journal of Integer Sequences*, 7:1–5, 2004. Article 04.3.3.
- [112] G. McLachlan and D. Peel. *Finite mixture models*. John Wiley and Sons Ltd., New York, NY, 2000.
- [113] C. Meek. Causal inference and causal explanation with background knowledge. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, pages 403–441, San Francisco, CA, 1995. Morgan Kaufmann.
- [114] A. Miller and M. Mihm. Mechanisms of disease: melanoma. *New England Journal of Medicine*, 355(1):51–65, 2006.
- [115] H. Minh, P. Niyogi, and Y. Yao. Mercer's theorem, feature maps, and smoothing. In G. Lugosi and H. Simon, editors, *Proceedings of the 19th Annual Conference on Learning Theory*, pages 154–168, Pittsburgh, PA, 2006.
- [116] D. Morrison. *Multivariate statistical methods*. McGraw-Hill, New York, NY, 1979.
- [117] K. Murphy. *Dynamic Bayesian networks: representation, inference and learning*. PhD thesis, Computer Science Division, University of California, Berkeley, CA, 2002.
- [118] R. Neapolitan. *Learning Bayesian networks*. Prentice Hall, Upper Saddle River, NJ, 2003.
- [119] D. Newman, S. Hettich, C. Blake, and C. Merz. UCI repository of machine learning databases. <http://www.ics.uci.edu/~mlern/MLRepository.html>, 1998.
- [120] K. Orii, T. Aoyama, K. Wakui, Y. Fukushima, et al. Genomic and mutational analysis of the mitochondrial trifunctional protein β -subunit (HADHB) gene in patients with trifunctional protein deficiency. *Human Molecular Genetics*, 6(8):1215–1224, 1997.
- [121] H. Pang, M. Bartlam, Q. Zeng, H. Miyatake, et al. Crystal structure of human pirin: an iron-binding nuclear protein and transcription cofactor. *Journal of Biological Chemistry*, 279(2):1491–1498, 2004.
- [122] E. Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.
- [123] A. Patton. Modelling asymmetric exchange rate dependence. *International Econometric Review*, 47(2):527–556, 2006.
- [124] J. Pearl. *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann, San Mateo, CA, 1988.

Bibliography

- [125] J. Pearl. *Causality: Models, reasoning, and inference*. Cambridge University Press, Cambridge, UK, 2000.
- [126] J. Pearl and A. Paz. A graph-based logic for reasoning about relevance relations. In B. du Boulay, D. Hogg, and L. Steels, editors, *Proceedings of the 7th European Conference on Artificial Intelligence*, pages 357–363, Brighton, UK, 1986.
- [127] G. Preston, J. Barrett, J. Biermann, and E. Murphy. Effects of alterations in calcium homeostasis on apoptosis during neoplastic progression. *Cancer Research*, 57:537–542, 1997.
- [128] T. Pukrop and C. Binder. The complex pathways of WNT5A in cancer progression. *Journal of Molecular Medicine*, 2007.
- [129] C. Rasmussen and Z. Ghahramani. Occam’s razor. In T. Leen, T. Dietterich, and V. Tresp, editors, *Proceedings of the 13th Neural Information Processing Systems Conference*, pages 294–300, Cambridge, MA, 2000. MIT Press.
- [130] M. Reed and B. Simon. *Functional analysis*. Academic Press, San Diego, CA, 1980.
- [131] H. Reichenbach. *The direction of time*. University of California Press, Berkeley, CA, 1956.
- [132] R. Renner and U. Maurer. About the mutual (conditional) information. In *Proceedings of IEEE International Symposium on Information Theory*, page 364, Lausanne, 2002.
- [133] T. Richardson and P. Spirtes. Ancestral graph Markov models. *The Annals of Statistics*, 30(4):962–1030, 2002.
- [134] R. Robinson. Counting labeled acyclic digraphs. In F. Harary, editor, *New directions in graph theory*, pages 239–273. Academic Press, New York, NY, 1973.
- [135] W. Salmon. *Scientific explanation and the causal structure of the world*. Princeton University Press, Princeton, NJ, 1984.
- [136] H. Scheffé. A useful convergence theorem for probability distributions. *The Annals of Mathematical Statistics*, 18(5):434–438, 1947.
- [137] B. Schölkopf and A. Smola. *Learning with kernels*. MIT Press, Cambridge, MA, 2002.
- [138] B. Schölkopf, A. Smola, and K. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1998.
- [139] T. Schweder. Plots of p-values to evaluate many tests simultaneously. *Biometrika*, 69(3):493–502, 1982.
- [140] W. Sewell and V. Shah. Social class, parental encouragement, and educational aspirations. *American Journal of Sociology*, 73:559–572, 1968.

- [141] R. Shachter. Probabilistic inference and influence diagrams. *Operations Research*, 36(4):589–604, 1988.
- [142] C. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423,623–656, 1948.
- [143] S. Shimizu, P. Hoyer, A. Hyvärinen, and A. Kerminen. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030, 2006.
- [144] R. Silva and R. Scheines. Generalized measurement models. Technical Report CMU-CALD-04-101, Carnegie Mellon University, 2005.
- [145] R. Silva, R. Scheines, C. Glymour, and P. Spirtes. Learning the structure of linear latent variable models. *Journal of Machine Learning Research*, 7:191–246, 2006.
- [146] B. Skyrms. *Causal necessity: A pragmatic investigation of the necessity of laws*. Yale University Press, London, 1980.
- [147] A. Smola, B. Schoelkopf, and K. Müller. The connection between regularization operators and support vector kernels. *Neural Networks*, 11:637–649, 1998.
- [148] R. Snijders, K. Sundberg, W. Holzgreve, G. Henry, et al. Maternal age- and gestation-specific risk for trisomy 21. *Ultrasound Obstet Gynecol*, 13(3):167–170, 1999.
- [149] L. Song, A. Smola, A. Gretton, and K. Borgwardt. A dependence maximization view of clustering. In Z. Ghahramani, editor, *Proceedings of the 24th International Conference on Machine Learning*, pages 815–822, Corvallis, OR, 2007.
- [150] J. Sosman, A. Weeraratna, and V. Sondak. When will Melanoma Vaccines be proven effective? *Journal of Clinical Oncology*, 20(3):387–389, 2004.
- [151] K. Spencer, V. Souter, N. Tul, R. Snijders, et al. A screening program for trisomy 21 at 10-14 weeks using fetal nuchal translucency, maternal serum free β -human chorionic gonadotropin and pregnancy-associated plasma protein-A. *Ultrasound Obstet Gynecol*, 13(3):231–237, 1999.
- [152] P. Spirtes and C. Glymour. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, 9(1):67–72, 1991.
- [153] P. Spirtes, C. Glymour, and R. Scheines. *Causation, prediction, and search (Lecture notes in statistics)*. Springer Verlag, New York, NY, 1993.
- [154] W. Spohn. Stochastic independence, causal independence, and shieldability. *Journal of Philosophical Logic*, 9(1):73–99, 1980.
- [155] W. Spohn. On the properties of conditional independence. In P. Humphreys, editor, *Probability and Probabilistic Causality*, volume 1 of *Patrik Suppes: Scientific Philosopher*, pages 173–196. Kluwer Academic Publishers, Dordrecht, Nederland, 1994.

Bibliography

- [156] R. Stanley. Acyclic orientation of graphs. *Discrete Mathematics*, 5:171–178, 1973.
- [157] H. Steck. On the use of skeletons when learning in Bayesian networks. In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, pages 558–565, San Francisco, CA, 2000. Morgan Kaufmann Publishers.
- [158] H. Steck and V. Tresp. Bayesian belief networks for data mining. In *Proceedings of the 2nd Workshop on Data Mining und Data Warehousing als Grundlage moderner entscheidungsunterstützender Systeme*, pages 145–154, Magdeburg, Germany, 1996.
- [159] I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, 2001.
- [160] J. Storey. The positive false discovery rate: A Bayesian interpretation and the q -value. *The Annals of Statistics*, 31(6):2013–2035, 2003.
- [161] E. Street and M. Carroll. Preliminary evaluation of a food product. In J. Tanur, editor, *Statistics: A guide to the unknown*, pages 220–238. Holden-Day, San Francisco, CA, 1972.
- [162] M. Studený. Semigraphoids and structures of probabilistic conditional independence. *Annals of Mathematics and Artificial Intelligence*, 21(1):71–98, 1997.
- [163] P. Suppes. *A probabilistic theory of causality*. North-Holland, Amsterdam, 1970.
- [164] A. Thomson and R. Randall-Maciver. *The ancient races of the Thebaid*. Oxford University Press, Oxford, 1905.
- [165] M. Tramo, W. Loftus, R. Green, T. Stukel, et al. Brain size, head size, and IQ in monozygotic twins. *Neurology*, 50:1246–1252, 1998.
- [166] A. W. van der Vaart. *Asymptotic statistics*. Cambridge University Press, Cambridge, UK, 1998.
- [167] W. Vandaele. Participation in illegitimate activities: Erlich revisited. In A. Blumstein, J. Cohen, and D. Nagin, editors, *Deterrence and incapacitation*, pages 270–335. National Academy of Sciences, Washinton, D.C., 1978.
- [168] K. Verhoeven, K. Simonsen, and L. McIntyre. Implementing false discovery rate control: increasing your power. *OIKOS*, 108(3):643–647, 2005.
- [169] T. Verma and J. Pearl. Equivalence and synthesis of causal models. In *Proceedings of the 6th Conference on Uncertainty in Artificial Intelligence*, pages 255–270, New York, NY, 1990. Elsevier Science Publishers.
- [170] T. Verma and J. Pearl. An algorithm for deciding if a set of observed independencies has a causal explanation. In *Proceedings of the 8th Conference on Uncertainty in Artificial Intelligence*, pages 323–330, San Francisco, CA, 1992. Morgan Kaufmann.

- [171] G. Wahba. *Spline models for observational data*. CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM, Philadelphia, PA, 1990.
- [172] A. Wald. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, 54(3):426–482, 1943.
- [173] A. Weeraratna, Y. Jiang, G. Hostetter, K. Rosenblatt, et al. WNT5A signalling directly affects cell motility and invasion of metastatic melanoma. *Cancer Cell*, 1(3):279–288, 2002.
- [174] M. Wellman and M. Henrion. Explaining “explaining away”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(3):287–292, 1993.
- [175] N. Wiener. The theory of prediction. In E. F. Beckenbach, editor, *Modern Mathematics for Engineers*, pages 165–190. McGraw-Hill, 1956.
- [176] J. Williamson. *Bayesian nets and causality*. Oxford University Press, Oxford, 2005.
- [177] R. Winkelmann. Health care reform and the number of doctor visits: an econometric analysis. *Journal of applied econometrics*, 19:455–472, 2004.
- [178] S. Yaramakala. Fast Markov blanket discovery. Master’s thesis, Computer Science, Iowa State University, Ames, IA, 2004.
- [179] R. Yeung. A new outlook on Shannon’s information measures. *IEEE Transactions on Information Theory*, 37(3):466–474, 1991.
- [180] Y. Yilmaz, E. Alpaydin, L. Akin, and T. Bilgiç. Handling of deterministic relationships in constraint-based causal discovery. In J. Gámez and A. Salmerón, editors, *Proceedings of the 1st European Workshop on Probabilistic Graphical Models*, Cuenca, Spain, 2002.
- [181] S. Yoon, S. Park, C. Yun, and A. Chung. Roles of matrix metalloproteinases in tumor metastasis and angiogenesis. *Journal of Biochemistry and Molecular biology*, 36(1):128–137, 2003.
- [182] J. Yu, V. Smith, P. Wang, A. Hartemink, et al. Advances to Bayesian network inference for generating causal networks from observational biological data. *Bioinformatics*, 20(18):3594–3603, 2004.
- [183] J. Zhang. Generalized do-calculus with testable causal assumptions. In M. Meila and X. Shen, editors, *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics*, pages 665–672, San Juan, Puerto Rico, 2007.
- [184] J. Zhu, T. Fukushige, J. McGhee, and J. Rothman. Reprogramming of early embryonic blastomeres into endodermal progenitors by a *Caenorhabditis elegans* GATA factor. *Genes & Development*, 12(24):3809–3814, 1998.

Acknowledgments

First of all, I would like to express my deep gratitude to my supervisor PD Dr. Dominik Janzing at the Department of Computer Science, University Karlsruhe for his consistent enthusiasm, his endless patience, his constructive criticism, and for all his help throughout my Ph.D. studies. He has taught me innumerable lessons and insights on the workings of academic research in general. Without him, I would not have dived into the intriguing field of causal learning. His technical and editorial advice was essential to the completion of this dissertation.

The last three years at Max Planck Institute for Biological Cybernetics, Tübingen have been a valuable experience. I would like to thank Prof. Dr. Bernhard Schölkopf, head of the Department of Empirical Inference for Machine Learning and Perception, for providing a working environment where an excellent research environment in all respects is provided. His support and encouragement were, in the end, what made this dissertation possible. Moreover, numerous extensive discussions with him and his advice were essential to develop the understanding of the topic.

I wish to express my particular appreciation to Dr. Kenji Fukumizu at Institute of Statistical Mathematics, Tokyo for his guidance, valuable advice and helpful discussions on the theory of kernel dependence measures. I am very grateful to Dr. Arthur Gretton at our department for discussions and comments on the Hilbert-Schmidt independence criterion. Many thanks are due to Mr. Siegfried Schloissnig at German Cancer Research Center, Heidelberg for comments on the experiments with biological data.

My thanks also go to Prof. Dr. D. Margaritis at Department of Computer Science, Iowa State University for providing his code of Bayesian multi-resolution independence test for continuous variables.

I gratefully acknowledge Mr. S. Norajitra, Mr. M. Beck, Mr. R. Heldele and Mr. B. Schumacher at Institute for Materials Research III, Research Center Karlsruhe for providing the ceramic surface data, Prof. Dr. G. P. McCabe, Department of Statistics, Purdue University for permission to use the Cheese data, Prof. Dr. M. S. Gazzaniga, Department of Psychology, University of California for permission to use the Brain Size and IQ data, Prof. Dr. A. Blumstein, H. John Heinz III School of Public Policy and Management, Carnegie Mellon University for permission to use the US Crime data, and Oxford University Press, University of Oxford for permission to use the Egyptian Skulls data. I thank also B. Janzing at the Meteorological Station Furtwangen for providing the temperature data.

Last but not least, I would like to thank my mother and my sister, having supported and encouraged me all the time.

Declaration of Academic Honesty

I hereby declare that I have written this thesis on my own, and used no other than the stated sources and aids.

Eidesstattliche Erklärung: Ich erkläre hiermit, dass ich diese Arbeit selbständig und nur mit den angegebenen Hilfsmitteln angefertigt habe.

Tübingen, April 2008

.....
(Xiaohai Sun)