NICTA

Introduction

Would cutting salt intake reduce heart attack ris
upping the minimum wage increase unemployme
problems differ from the traditional machine learn
in that they require us to predict the consequence
tervention, which may change the properties of th
tion from which our data is sampled. Without expe
testing the intervention or making assumptions to
its effect, such inference is impossible. There

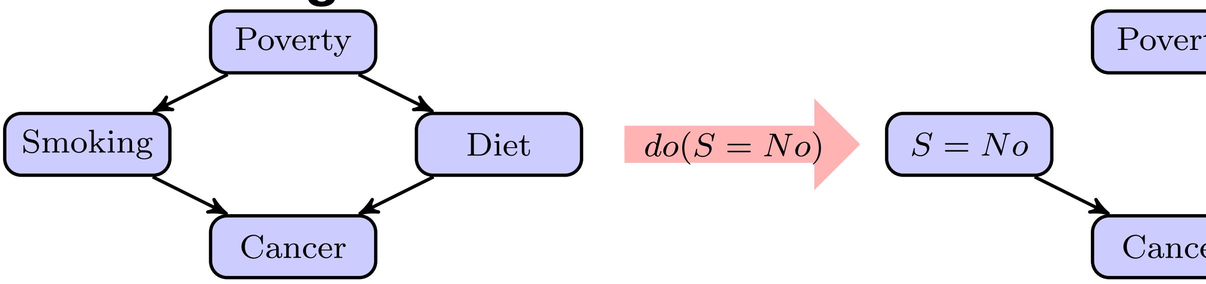
its effect, such inference is impossible. There
important questions where direct experimentation
sive, unethical or impossible. Recent research o
has clarified what assumptions allow causality to
mined and has demonstrated causal inference a
ery is possible with some very general assumptio

Causal Frameworks

Causal Directed Acyclic Graphs

- A causal DAG is a Bayesian network where defined to mean A causes B.
- Variables are independent of their non-effects direct causes (Causal Markov Property)
- An intervention that sets a subset of variables noted $do(\boldsymbol{X} = \boldsymbol{x})$, has a simple graphical repr in a causal DAG, G. All links entering intervene ables, $\boldsymbol{X}$, are deleted, resulting in the mutilate $G_{\overline{\boldsymbol{X}}}$ (figure 1). Thus, a causal DAG represents all possible interventional distributions over its v

Figure 1: *Intervention in a causal DAG*



(Causal) Structural Equation Models (S

- Represent each variable as a deterministic fun direct causes and a noise term, where the noise mutually independent.
- If the set of equations does not create a cycl

- If the set of equations does not create a cycl... Causal Markov Property holds and the SEM i... DAG, (but not visa-versa - SEMs can encode r... mation).

Counterfactuals

- Counterfactuals are statements about what wou... under alternate realities where some specified... fers. For example, consider people taking a me...

  For an individual, i, let: $\begin{cases} y_i^0 = & \text{outcome if } x_i = \\ y_i^1 = & \text{outcome if } x_i = \end{cases}$

- We can define a random variable Y^1, where P... distribution of outcome, Y, that would occur i... was treated. Similarly $P(Y^0)$ is the distribution o... if no-one was treated.
- If $(X \perp\!\!\!\perp Y^0|\boldsymbol{Z})$ & $(X \perp\!\!\!\perp Y^1|\boldsymbol{Z})$: $\longleftarrow$ Ignoreabilit... tion

  we can calculate counterfactual distributions... served ones:

  $P(Y^1|Z) = P(Y|X=1,Z)$ and $P(Y^0|Z) = P(Y$...

- Distributions over counterfactual variables that c... to interventions can be translated directly to th... tion $P(Y^1) = P(Y|do(X=1))$. However we c... queries with counterfactual variables that are no... tional. For example: *what is the proability tha... was not treated and died, would have recover... been treated?*. This query asks about the joint c... of $P(Y^0, Y^1)$.

Causal Infere

sk? Would
nt? Causal
ning setting
es of an in-
he distribu-
erimentally
o constrain

group	placebo	treatment	probability
1	die	die	$\alpha = P(Y^0 =$
2	die	recover	$\beta = P(Y^0 =$
3	recover	die	$\gamma = P(Y^0 =$
4	recover	recover	$\delta = P(Y^0 =$

- Counterfactuals can be defined in terms o
$f(X = 0, \epsilon_Y)$
- The ignorability assumption $(X \perp\!\!\!\perp Y^0|\boldsymbol{Z})$ &
is weaker than that implied by mutually in

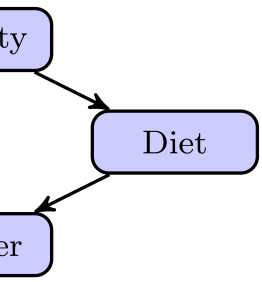
are many
n is expen-
n causality
o be deter-
and discov-
ns.

$A \rightarrow B$ is

given their

$\boldsymbol{X}$ to $\boldsymbol{x}$, de-
resentation
ed on vari-
ed network
s the set of
variables.

G



SEMs)

nction of its
e terms are

e then the

rors in a SEM, $(X \perp\!\!\!\perp Y^0, Y^1 | \boldsymbol{Z})$. The assum
made equivalent by utilizing SEMs with a
pendence of errors assumption [9].

Causal Inference

Consider the problem where the causal DAG
to theory or prior knowledge, and we wish t
come of an intervention of the form $P(\boldsymbol{Y} | do($
observational data.

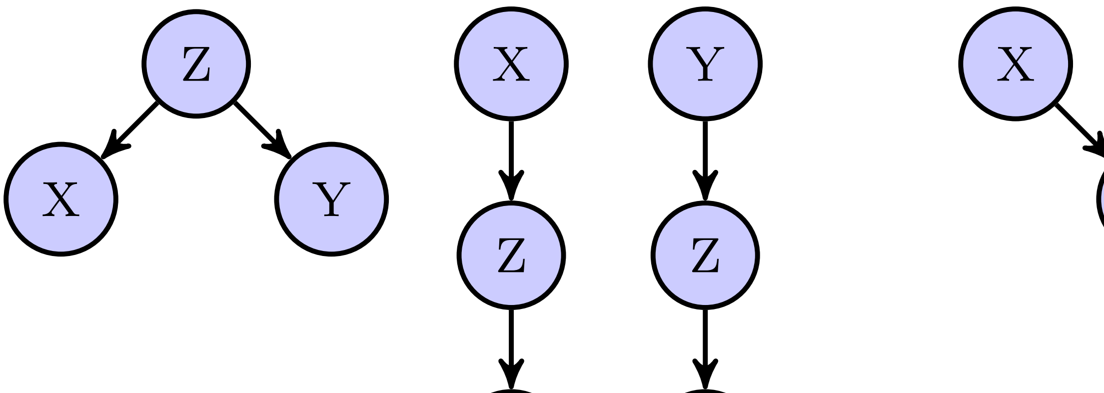
- If there are no latent variables, we can con
of any intervention by simply multiply the
mutliated network (figure 1).

The Do Calculus

The do-calculus consists of three rules, wh
d-separation in a causal DAG [7]. It is comp
effect is non-parametrically identifiable if and
terventional query can be reduced to an obs
via these rules [10].

Figure 2: *d-separation allows us to read co
pendences off a DAG. If a set of variables $\boldsymbol{Z}$
and $\boldsymbol{Y}$ in G then $(\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} | \boldsymbol{Z})$ in all distribution
with G.*

(a) $(X \perp\!\!\!\perp Y | Z)$ (b) v-struc



e then the
s a causal
more infor-

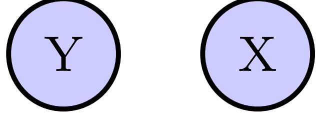
uld happen
d thing dif-
edical drug:
0 (not treated)
1 (treated)
(Y^1) is the
f everyone
of outcome

y Assump-

from ob-

$|X = 0, Z)$

correspond
ue do nota-
can phrase
ot interven-
t Joe, who
red had he
distribution



The Three Rules

1. A causal DAG remains a causal DAG after a
so d-separation still applies.

if $(\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{W}|\boldsymbol{X})$ in $G_{\overline{\boldsymbol{X}}}$

$$P(\boldsymbol{Y}|do(\boldsymbol{X}=\boldsymbol{x}), \boldsymbol{W}=\boldsymbol{w}) = P(\boldsymbol{Y}$$

2. If $\boldsymbol{Y}$ is independent of *how* variables $\boldsymbol{X}$ ta
then the effect on $\boldsymbol{Y}$ of setting $\boldsymbol{X}$ to some
alent to observing it take that value. If this r
the corresponding ignoreability assumptio
terfactual framework is satisfied.

if $(\boldsymbol{Y} \perp\!\!\!\perp \hat{\boldsymbol{X}}|\boldsymbol{X}, \boldsymbol{Z})$ in $G^{\dagger}$

$$P(\boldsymbol{Y}|do(\boldsymbol{X}=\boldsymbol{x}), \boldsymbol{Z}) = P$$

3. If there is no direct causal path from $\boldsymbol{X}$ to
vention on $\boldsymbol{X}$ does not change the distribu

if $(\boldsymbol{Y} \perp\!\!\!\perp \hat{\boldsymbol{X}}|\boldsymbol{Z})$ in $G^{\dagger}$

$$P(\boldsymbol{Y}|do(\boldsymbol{X}=\boldsymbol{x}), \boldsymbol{Z}) =$$

(For readability, this is a simplified version of the do-calculus that
on a single variable or cases where it is sufficient for identifiability
tion on all variables together. The fully general version is only s
see [7])

ence in Machi

Finnian Lattim

Australian National

finnlattimore@gm

of group

$0, Y^1 = 0)$
$0, Y^1 = 1)$
$1, Y^1 = 0)$
$1, Y^1 = 1)$

of SEMs $Y^0 \sim$

& $(X \perp\!\!\!\perp Y^1 | \boldsymbol{Z})$

ndependent er-

Causal Discovery

Causal discovery attempts to infer causal
based on more general assumptions.

Indepdendence Based Methods

Without Latent Variables

1. We assume our distribution P was

mptions can be
modified inde-

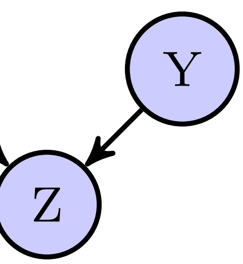
is known, due
o infer the out-
$\boldsymbol{X} = \boldsymbol{x}))$ using

mpute outcome
factors in the

ich result from
plete: a causal
d only if the in-
servational one

onditional inde-
d-separates $\boldsymbol{X}$
s P compatible

ture, $(X \perp\!\!\!\perp Y)$

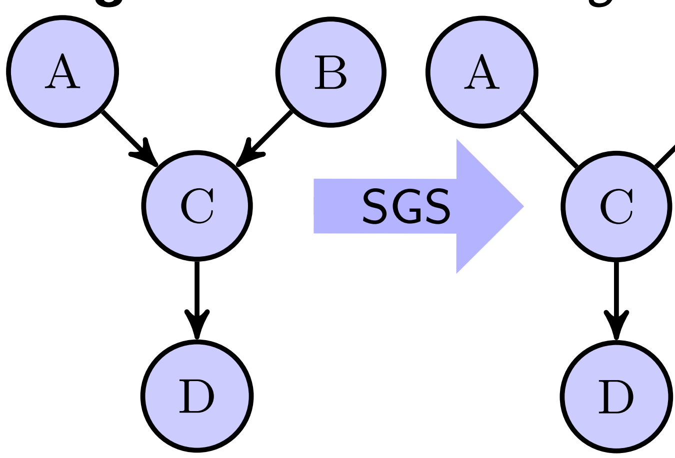


1. We assume our distribution P was g
(unknown) causal DAG over our c
(*causal sufficiency*)
2. We assume that all the conditional in
are implied by d-separation in the tr
(*faithfulness*)
3. Finding the causal structure equates
maps for P

Algorithm 1: *SGS or IC Algorith*
Input: A Distribution P over variables $\boldsymbol{V}$
Output: A partially directed network
Markov equivalence class for the genera

1. Create a complete undirected graph o
of variables $(a, b) \in V$ search for a set
such a set exists, delete the link $a - b$
2. For all pairs of unlinked-nodes $(\alpha, \beta$
neighbour c, if $c \notin S_{\alpha\beta}$ direct links towa
3. Recursively direct any remaining links
only one orientation that does not cre
additional v-structures ($\bullet \rightarrow \bullet \leftarrow \bullet$).

Figure 3: *The SGS Algor*



- The SGS algorithm is infeasible in p

an intervention

$|do(\boldsymbol{X} = \boldsymbol{x}))$

ke their values
value is equiv-
rule is satisfied
n in the coun-

$(\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Z})$

o $\boldsymbol{Y}$ then inter-
tion of $\boldsymbol{Y}$.

$P(\boldsymbol{Y}|\boldsymbol{Z})$

covers interventions
to consider interven-
lightly more complex

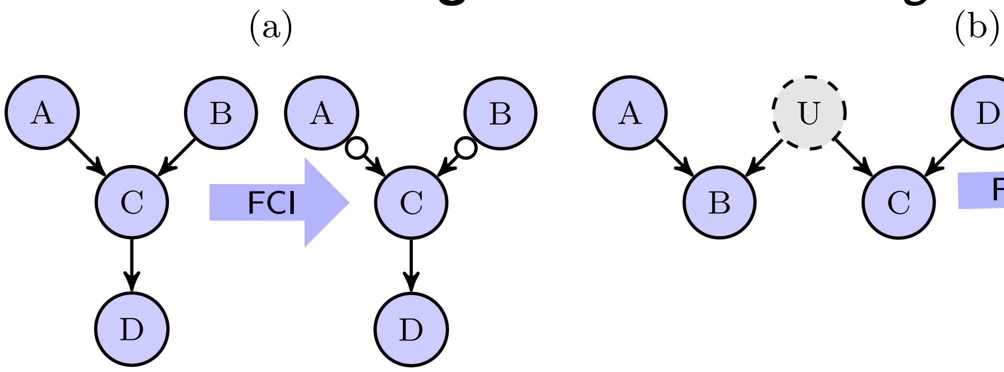
exponential number of (high order) co
dence tests it requires.

- The PC algorithm [11] modifies the SG
ploit any sparsity in the true network, le
ter average case performance.

With Latent Variables

- For every latent structure there is a d
lent structure in which every latent vari
with exactly two children [12]. This ke
rithm [11], which generalizes the PC a
latent and selection variables.
- The FCI algorithm returns an equivale
imal Ancestral Graphs (a generalizatio
DAGs are not closed under marginaliza

Figure 4: *The FCI Algori*



- The FCI algorithm discovers all aspec
ture identifiable from conditional indep
[13].

ne Learning:

nore
University
nail.com

structure from data

enerated by some

- It can be made to require a worst c
than exponential) number of con
tests for sparse graphs [1].
- Implementations of both the PC
available in the R package pcalg [5

Beyond independence

Independence based methods have f
require only very general, non-par
However they cannot distinguish b

enerated by some
bserved variables

dependences in P
ue causal network

to finding perfect

hm [11, 7].

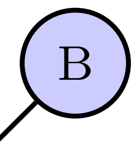
representing the
ting causal model.

ver V. For all pairs
S_{ab} s.t $a \perp\!\!\!\perp b|S_{ab}$. If

$)$ with a common
ards c.

for which there is
eate a cycle or any

ithm



ractise due to the

with equivalent dependency structure
$X \to Y$ and $X \leftarrow Y$.

Figure 5: *Figure from [3]. Additiv* *$f(X)+\epsilon$, are identifiable for most con* *but not in linear-gaussian case*

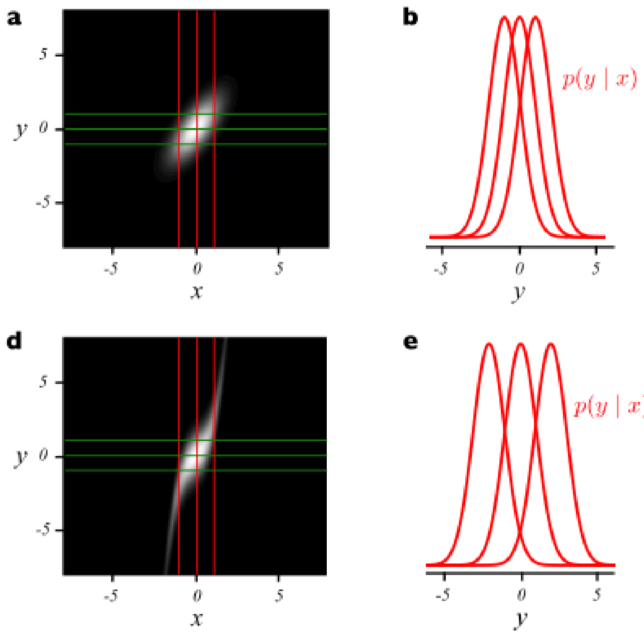
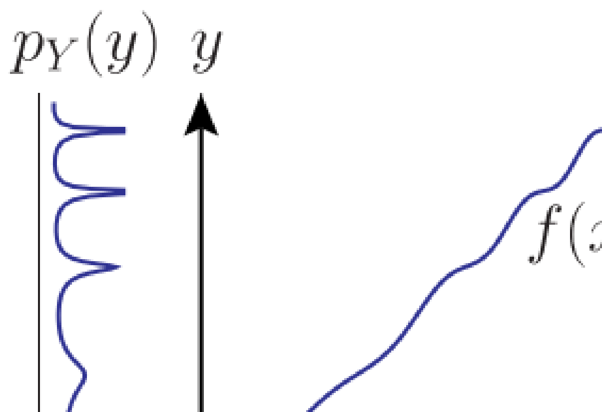


Figure 6: *Figure from [2]. The causa* *tifiable even where the relationship b* *terministic and invertible. Let $Y = f$* *most input distributions, $p_X(x)$, the c* *higher where f' is small and a large* *similar values of Y. If X causes Y (* *we would expect f and $p_X(x)$ to be* *should be correlated with f'.*



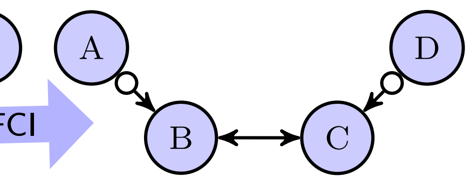
onditional indepen-

GS algorithm to ex-
eading to much bet-

ependency equiva-
iable is a root node
ey to the FCI algo-
algorithm to handle

ence class of Max-
on of DAGs) since
ation (figure 4b).

thm



cts of causal struc-
pendence relations

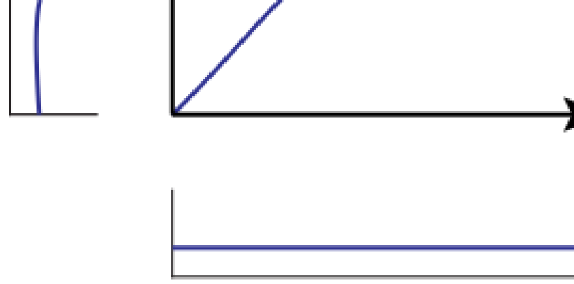
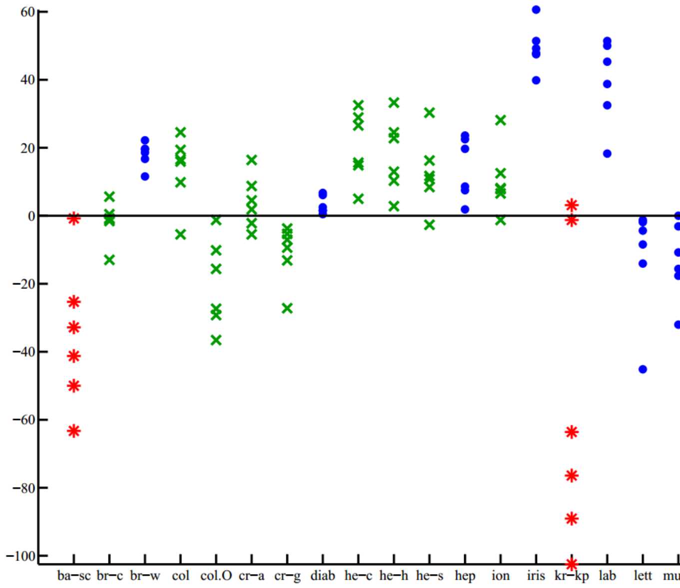


Figure 7: *Figure from [4]. The of mechanism and input can be g deterministic setting. If $X \to Y$ th should be independent, but $P(Y)$ Therefore, semi-supervised learning fit when trying to learn in the causa $P(Y|X)$) but could help when learnin rection.*



Learning what causality looks like

Suppose we had M different causal

$$D = \{\{x_j, y_j\}_{j=1}^{N_i}, l$$

A review

case polynomial (rather
nditional independence

and FCI algorithm are
5]

the advantage that they
rametric, assumptions.
etween causal graphs

where l_i is a binary label that inc
in dataset i.

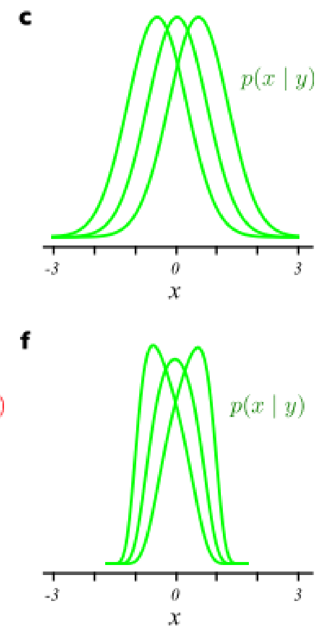
- Kernel mean embedding allow
and transform it to a point in so
- We expect there to be differen
tween $P(X)$, $P(Y)$ and $P(Y|X$

Algorithm

1. Let μ be a kernel mean embe

e; for example, between

ve noise models, $Y =$
nbinations of f and $P(\epsilon)$



al direction can be iden-
between X and Y is de-
$f(X) \Leftrightarrow X = g(Y)$. For
distribution $p_Y(y)$ will be
er region of X maps to
(but not if Y causes X)
independent and $p_Y(y)$

$x)$

tion P into some Hilbert space

2. For each data set $i = 1...M$,
   that approximates $\mu(P(X)), \mu($

3. Apply a standard classification
   Y or $Y \rightarrow X$

With more than two variables

- If you can come up with a
  $(P(Y), P(\epsilon), f)$, that guarante
  variate SEM $Y = f(X) + \epsilon$,
  to get the conditions under wh
  identifiable [8].

Referenc

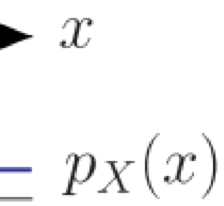
[1] Tom Claassen, Joris Mooij,
sparse causal models is not

[2] Povilas Daniusis, Dominik J
Zscheischler, Bastian Steud
hard Schölkopf. Inferring de
In *UAI*, 2010.

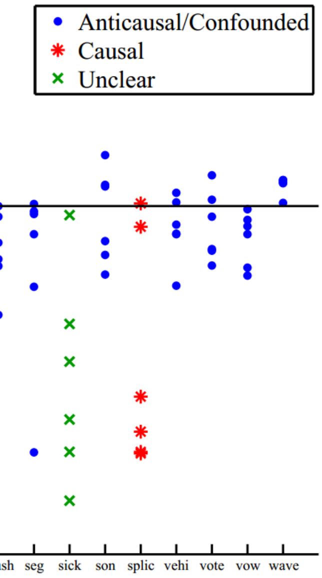
[3] Patrick Hoyer, Dominik Janz
linear causal discovery with
NIPS, 2009.

[4] Dominik Janzing, Jonas Pe
Zhang, Joris M. Mooij, and
causal and anticausal learni

[5] Markus Kalisch, Martin Mäc
loes H. Maathuis, and Pete
ence using graphical models



idea of independence eneralized to the non- hen $P(X)$ and $P(Y|X)$ and $P(X|Y)$ are not. should yield no bene- al direction, (estimating ng in the anti-causal di-



[6]

pairs data sets.

$$i\}_{i=1}^{M}$$

JSS, VV(Ii), 2012.

[6] David Lopez-Paz, Krikamo Recht. The Randomized C *Prepr. arXiv1409.4366*, Sept

[7] Judea Pearl. *Causality: m ence*. MIT Press, Cambridg

[8] Jonas Peters, Joris Mooij, D hard Schölkopf. Causal disc tive noise models. *JMLR*, 15

[9] Thomas S. Richardson and world intervention graphs (S counterfactual and graphica Center for the Statistics and versity of Washington Series

[10] Ilya Shpitser and Judea Pea methods for the causal hiera 2008.

[11] Peter Spirtes, Clark N Glym *Causation, prediction, and* press, 2000.

[12] TS Verma. Graphical aspec nical report, 1993.

[13] Jiji Zhang. On the comple for causal discovery in the founders and selection bia 172(16-17):1873–1896, Nov

dicates if $X \to Y$ or $Y \to X$

s us to take a distribution P
ome Hilbert space.

nces in the relationships be-
$)$ for $X \to Y$ and $Y \to X$

m 2:

edding that maps a distribu-e.

construct a feature vector $(P(Y)), \mu(P(X,Y))$

n algorithm to learn if $X \rightarrow$

a condition, on the triple es identifiability for the bi- you can extend that result hich the multivariate case is

ces

and Tom Heskes. Learning NP-hard. In *UAI*, 2013.

Janzing, Joris Mooij, Jakob del, Kun Zhang, and Bern- terministic causal relations.

zing, and Joris Mooij. Non- additive noise models. In

eters, Eleni Sgouritsa, Kun d Bernhard Schölkopf. On ng. In *ICML*, 2012.

chler, Diego Colombo, Mar- r Bühlmann. Causal infer- s with the R package pcalg.

s with the R package pcalg.

l Muandet, and Benjamin
ausation Coefficient. *arXiv*
tember 2014.

odels, reasoning and infer-
e, 2000.

Dominik Janzing, and Bern-
overy with continuous addi-
5:2009–2053, 2014.

James M. Robins. Single
SWIGs): a unification of the
al approaches to causality.
d the Social Sciences, Uni-
. Working Paper 128, 2013.

arl. Complete identification
archy. *JMLR*, 9:1941–1979,

our, and Richard Scheines.
search, volume 81. MIT

cts of causal models. Tech-

eteness of orientation rules
e presence of latent con-
as. *Artificial Intelligence*,
vember 2008.