Recovering from Selection Bias in Discrete Causal Models.

Robin J. Evans
University of Oxford

and

Vanessa Didelez
University of Bristol

UAI Causal Workshop
16th July 2015

Outline

Outline

Selection Bias

Selection bias is perennial in statistics.

Examples:

- case-control studies;
- studies with dropout;
- survey response bias;
- polling;
- after dinner speakers (survivor bias);
- ...

Selection Bias

Selection bias is perennial in statistics.

Examples:

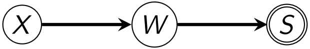
- case-control studies;
- studies with dropout;
- survey response bias;
- polling;
- after dinner speakers (survivor bias);
- ...

Possible remedies:

- re-weighting with extra information;
- bias modelling;
- sensitivity analysis;
- use the odds-ratio.

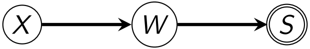
Case-Control Study Example

- binary exposure X;
- binary outcome W (e.g. disease presence);
- selection indicator S;
  case-control, so selection ($S = 1$) depends upon W.



Case-Control Study Example

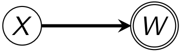
- binary exposure X;
- binary outcome W (e.g. disease presence);
- selection indicator S;
  case-control, so selection ($S = 1$) depends upon W.



We **observe** data from $p(x, w \mid s = 1) = p(x \mid w)p(w \mid s = 1)$.

Case-Control Study Example

- binary exposure X;
- binary outcome W (e.g. disease presence);
- selection indicator S;
  case-control, so selection ($S = 1$) depends upon W.

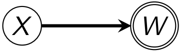


We **observe** data from $p(x, w \mid s = 1) = p(x \mid w)p(w \mid s = 1)$.

Equivalent to the conditional $p(x \mid w)$ with $p(w)$ unknown.

Case-Control Study Example

- binary exposure X;
- binary outcome W (e.g. disease presence);
- selection indicator S;
  case-control, so selection $(S = 1)$ depends upon W.



We **observe** data from $p(x, w \mid s = 1) = p(x \mid w)p(w \mid s = 1)$.

Equivalent to the conditional $p(x \mid w)$ with $p(w)$ unknown.

Without further assumptions we cannot recover $p(w)$ nor therefore $p(w \mid x) = p(w \mid do(x))$.

Well known that we can recover and use the causal odds-ratio.

Structural Information

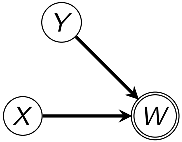
However, with background information we might be able to do better.

Structural Information

However, with background information we might be able to do better.

Suppose that there is a covariate Y, known to be independent of X.

Example: X gene, Y background environmental effect, W disease. (Moerkerke et al., 2010)

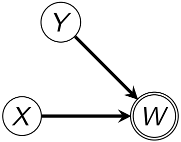


Structural Information

However, with background information we might be able to do better.

Suppose that there is a covariate Y, known to be independent of X.

Example: X gene, Y background environmental effect, W disease. (Moerkerke et al., 2010)



$X \perp\!\!\!\perp Y$ but generally $X \not\perp\!\!\!\perp Y \mid W$ due to 'explaining away'.

So **true** weighting $p(w)$ of $p(x, y \mid w)$ tables gives $X \perp\!\!\!\perp Y$:

$$\sum_w p(w) \cdot p(x, y \mid w) = f(x) \cdot g(y).$$

Concrete Example

Suppose we observe $+$ve correlation under $W = 0$, $-$ve given $W = 1$.

$W = 0$	0	1
0	0.4	0.1
1	0.1	0.4

$W = 1$	0	1
0	0.2	0.3
1	0.3	0.2

Concrete Example

Suppose we observe +ve correlation under $W = 0$, −ve given $W = 1$.

$W = 0$	0	1
0	0.4	0.1
1	0.1	0.4

$W = 1$	0	1
0	0.2	0.3
1	0.3	0.2

True marginal table $p(x, y) = \alpha p(x, y \mid w = 0) + (1 - \alpha) p(x, y \mid w = 1)$ some unknown α.

Concrete Example

Suppose we observe +ve correlation under $W = 0$, $-$ve given $W = 1$.

$W = 0$	0	1
0	0.4	0.1
1	0.1	0.4

$W = 1$	0	1
0	0.2	0.3
1	0.3	0.2

True marginal table $p(x, y) = \alpha p(x, y \mid w = 0) + (1 - \alpha) p(x, y \mid w = 1)$ some unknown α.

Mixture is:

	0	1
0	$0.2 + 0.2\alpha$	$0.3 - 0.2\alpha$
1	$0.3 - 0.2\alpha$	$0.2 + 0.2\alpha$

Independence means $(0.2 + 0.2\alpha)^2 - (0.3 - 0.2\alpha)^2 = 0$.

Concrete Example

Suppose we observe +ve correlation under $W = 0$, −ve given $W = 1$.

$W = 0$	0	1
0	0.4	0.1
1	0.1	0.4

$W = 1$	0	1
0	0.2	0.3
1	0.3	0.2

True marginal table $p(x, y) = \alpha p(x, y \mid w = 0) + (1 - \alpha)p(x, y \mid w = 1)$ some unknown α.

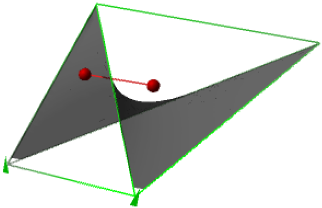
Mixture is:

	0	1
0	0.25	0.25
1	0.25	0.25

Only value giving independence in this case: $\alpha = 0.25$.

Geometric Picture

Surface of independence in 2×2 probability simplex:



Idea

It's common to use background information to augment studies: e.g. particular re-weightings for groups in a survey.

e.g.:
Bowden and Vansteelandt (2010)
Borboudakis and Tsamardinos (2015).

Idea

It's common to use background information to augment studies: e.g. particular re-weightings for groups in a survey.

e.g.:
Bowden and Vansteelandt (2010)
Borboudakis and Tsamardinos (2015).

Can we use **structural information** to recover a joint distribution, rather than particular numbers?

Outline

Identifiability

Let $p_\theta : \Theta \to \mathcal{M}$ be a map from a parameter space Θ to a collection of probability distributions $\mathcal{M}$.

Identifiability

Let $p_\theta : \Theta \to \mathcal{M}$ be a map from a parameter space Θ to a collection of probability distributions $\mathcal{M}$.

Say that θ is **generically k-identifiable** if the fibers

$$F(\theta) = \{\theta' : p_\theta = p_{\theta'}\}, \qquad\qquad \forall \theta \in \Theta \setminus \mathcal{O}$$

have cardinality at most $k \in \mathbb{N}$ for some $\mathcal{O}$ of measure zero.

Identifiability

Let $p_\theta : \Theta \to \mathcal{M}$ be a map from a parameter space Θ to a collection of probability distributions $\mathcal{M}$.

Say that θ is **generically k-identifiable** if the fibers

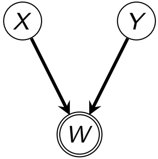
$$F(\theta) = \{\theta' : p_\theta = p_{\theta'}\}, \qquad\qquad \forall \theta \in \Theta \setminus \mathcal{O}$$

have cardinality at most $k \in \mathbb{N}$ for some $\mathcal{O}$ of measure zero.

So 'almost everywhere' at most a k-to-one map.

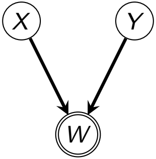
Identifiability

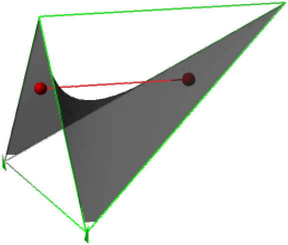
In the marginal independence example, surface of independence is quadratic so at most 2 solutions.



Identifiability

In the marginal independence example, surface of independence is quadratic so at most 2 solutions.

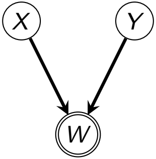


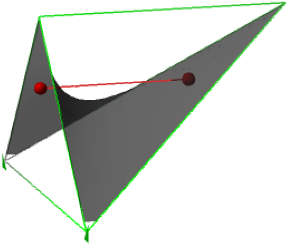


Cases with two solutions are manifestations of Simpson's paradox.

Identifiability

In the marginal independence example, surface of independence is quadratic so at most 2 solutions.





Cases with two solutions are manifestations of Simpson's paradox.

If either $X \perp\!\!\!\perp W \mid Y$ or $Y \perp\!\!\!\perp W \mid X$ then lose identifiability (X and Y are analogous to instruments).

Overall: generically 2-identifiable.

Main Result

Discrete random variables $\boldsymbol{X}, W$, with d_x, d_w states.

$$\begin{aligned} p(\boldsymbol{x}, w) &= (p(\boldsymbol{x}), p(w \mid \boldsymbol{x})) \\ &\in \mathcal{M}_X \times \mathcal{M}_{W|X}. \end{aligned}$$

Main Result

Discrete random variables $\boldsymbol{X}, W$, with d_x, d_w states.

$$p(\boldsymbol{x}, w) = (p(\boldsymbol{x}), p(w \mid \boldsymbol{x})) \\ \in \mathcal{M}_X \times \mathcal{M}_{W|X}.$$

Separate marginal model for $\boldsymbol{X}$ and conditional model for $W \mid \boldsymbol{X}$.

Main Result

Discrete random variables $\boldsymbol{X}, W$, with d_x, d_w states.

$$p(\boldsymbol{x}, w) = (p(\boldsymbol{x}), p(w \mid \boldsymbol{x})) \\ \in \mathcal{M}_X \times \mathcal{M}_{W|X}.$$

Separate marginal model for $\boldsymbol{X}$ and conditional model for $W \mid \boldsymbol{X}$.

Want conditions on $\mathcal{M}_X$ and $\mathcal{M}_{W|X}$ that lead to (generic) k-identifiability of $p(\boldsymbol{x}, w)$ from $p(\boldsymbol{x} \mid w)$.

Main Result

Discrete random variables $\boldsymbol{X}, W$, with d_x, d_w states.

$$p(\boldsymbol{x}, w) = (p(\boldsymbol{x}), p(w \mid \boldsymbol{x})) \\ \in \mathcal{M}_X \times \mathcal{M}_{W|X}.$$

Separate marginal model for $\boldsymbol{X}$ and conditional model for $W \mid \boldsymbol{X}$.

Want conditions on $\mathcal{M}_X$ and $\mathcal{M}_{W|X}$ that lead to (generic) k-identifiability of $p(\boldsymbol{x}, w)$ from $p(\boldsymbol{x} \mid w)$.

Theorem

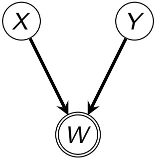
Suppose

- $\mathcal{M}_{W|X}$ unrestricted;
- $\mathcal{M}_X$ has codimension ℓ.

Then $p(w)$ generically k-identifiable from $p(x \mid w)$ if and only if $d_w - 1 \leq \ell$.

i.e. $d_w - 1$ unknowns, ℓ constraints.

Example: Marginal Independence



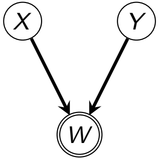
Marginal independence case:

independence is $(d_x - 1)(d_y - 1)$ constraints;

so works iff

$$(d_x - 1)(d_y - 1) \geq d_w - 1.$$

Example: Marginal Independence



Marginal independence case:

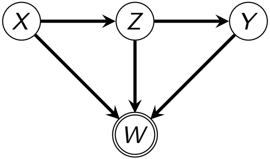
independence is $(d_x - 1)(d_y - 1)$ constraints;

so works iff

$$(d_x - 1)(d_y - 1) \geq d_w - 1.$$

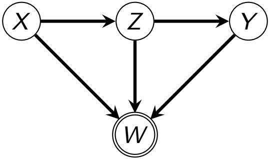
All binary case: 1 constraint, 1 unknown, so this is **just identified** (generically up to 2 solutions).

Example: Conditional Independence



In this case marginal model $X \perp\!\!\!\perp Y \mid Z$, but we observe $p(x, y, z \mid w)$.

Example: Conditional Independence

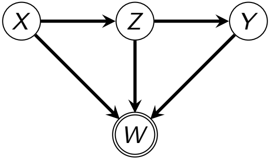


In this case marginal model $X \perp\!\!\!\perp Y \mid Z$, but we observe $p(x, y, z \mid w)$.

This model implies $(d_x - 1)(d_y - 1)d_z$ constraints, $d_w - 1$ unknowns.

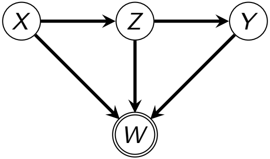
In the all binary case for example, we have generic 1-identifiability.

Example: Conditional Independence



But more is true!

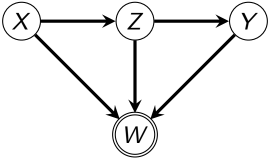
Example: Conditional Independence



But more is true!

$$p(z) \cdot p(x, y, z) - p(x, z) \cdot p(y, z) = 0, \qquad \forall x, y, z.$$

Example: Conditional Independence

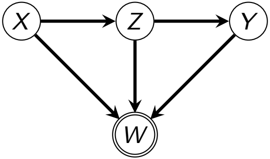


But more is true!

$$p(z) \cdot p(x, y, z) - p(x, z) \cdot p(y, z) = 0, \qquad \forall x, y, z.$$

Replace $p(x, y, z) = \sum_w p(x, y, z \mid w)\alpha(w)$, to get series of quadratic equations in $\alpha(w)$.

Example: Conditional Independence



But more is true!

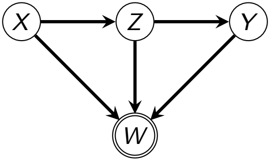
$$p(z) \cdot p(x, y, z) - p(x, z) \cdot p(y, z) = 0, \qquad \forall x, y, z.$$

Replace $p(x, y, z) = \sum_w p(x, y, z \mid w)\alpha(w)$, to get series of quadratic equations in $\alpha(w)$.

All binary case gives **two** independent quadratics for one unknown. For distributions not in model, generically these don't have common solutions.

$\implies$ we have a degree of freedom to test this model.

Example: Conditional Independence

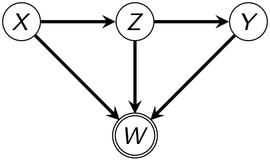


Fitting: given counts can just maximize the conditional log-likelihood:

$$\sum_{\boldsymbol{x},w} n(\boldsymbol{x},w)\log p(\boldsymbol{x}\mid w) = \sum_{\boldsymbol{x},w} n(\boldsymbol{x},w)\log p(\boldsymbol{x},w) - \sum_{w} n(w)\log p(w),$$

use a likelihood ratio test.

Example: Conditional Independence



Fitting: given counts can just maximize the conditional log-likelihood:

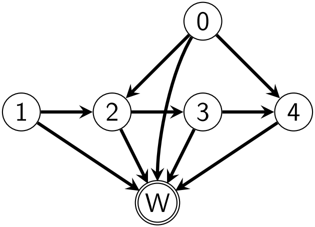
$$\sum_{\boldsymbol{x},w} n(\boldsymbol{x},w)\log p(\boldsymbol{x}\mid w) = \sum_{\boldsymbol{x},w} n(\boldsymbol{x},w)\log p(\boldsymbol{x},w) - \sum_{w} n(w)\log p(w),$$

use a likelihood ratio test.

Model is irregular and behaves like a latent variable model.

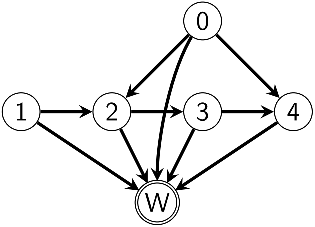
Example: Bayesian Network

Any Bayesian network (or ancestral graph, nested model, ...) such that all other variables are parents of W is potentially identifiable:



Example: Bayesian Network

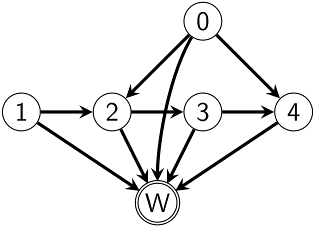
Any Bayesian network (or ancestral graph, nested model, ...) such that all other variables are parents of W is potentially identifiable:



For binary variables this $\mathcal{M}_X$ has codimension 19.

Example: Bayesian Network

Any Bayesian network (or ancestral graph, nested model, ...) such that all other variables are parents of W is potentially identifiable:



For binary variables this $\mathcal{M}_X$ has codimension 19.

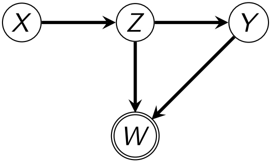
Of course, could then recover appropriate causal effects from the joint.

This may appear to contradict Bareinboim and Tian (2015), but they require *strict* identifiability.

Outline

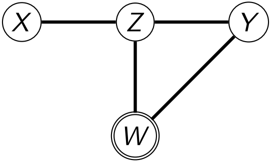
Variation Independence

Beware additional independences!



Variation Independence

Beware additional independences!

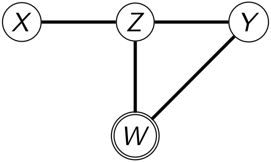


In this case

$$p(x, y, z, w) = p(w) \cdot p(y, z \mid w) \cdot p(x \mid z)$$

Variation Independence

Beware additional independences!



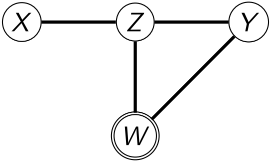
In this case

$$p(x, y, z, w) = p(w) \cdot p(y, z \mid w) \cdot p(x \mid z)$$
$$p(x, y, z \mid w) = \qquad p(y, z \mid w) \cdot p(x \mid z).$$

Note that $p(x, y, z \mid w)$ is in this model if and only if this factorization holds, **regardless of the value of** $p(w)$.

Variation Independence

Beware additional independences!



In this case

$$p(x, y, z, w) = p(w) \cdot p(y, z \mid w) \cdot p(x \mid z)$$
$$p(x, y, z \mid w) = \qquad p(y, z \mid w) \cdot p(x \mid z).$$

Note that $p(x, y, z \mid w)$ is in this model if and only if this factorization holds, **regardless of the value of** $p(w)$.

Therefore $p(w)$ is clearly unidentifiable.

Lessons

1. These results are all **generic**.
   There are areas of the joint distribution which need to be avoided (think of these as faithfulness conditions).

Lessons

1. These results are all **generic**.
   There are areas of the joint distribution which need to be avoided (think of these as faithfulness conditions).
2. In particular: we can't just 'weaken' our assumptions to make life easier (e.g. adding extra edges on the graph).

Lessons

1. These results are all **generic**.
   There are areas of the joint distribution which need to be avoided (think of these as faithfulness conditions).

2. In particular: we can't just 'weaken' our assumptions to make life easier (e.g. adding extra edges on the graph).
3. The constraint was exhibited directly in the observed distribution

$$p(x, y, z \mid w) = p(y, z \mid w) \cdot p(x \mid z).$$

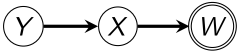
   so:

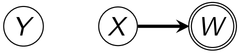
   - the model can still be tested (more easily than in the non-degenerate case);
   - we can 'see' when the procedure fails.

Parameter Cuts

Proposition

Suppose $p(x \mid y, w)$ is variation independent of $p(y, w)$ in $\mathcal{M}$.
Then $p(x, y, w)$ identifiable from $p(x, y \mid w)$ if and only if
$p(y, w)$ identifiable from $p(y \mid w)$





Parameter Cuts

Proposition

Suppose $p(x \mid y, w)$ is variation independent of $p(y, w)$ in $\mathcal{M}$.
Then $p(x, y, w)$ identifiable from $p(x, y \mid w)$ if and only if
$p(y, w)$ identifiable from $p(y \mid w)$

In other words, $p(x \mid y, w)$ doesn't help us to identify $p(w)$.

Parameter Cuts

Proposition

Suppose $p(x \mid y, w)$ is variation independent of $p(y, w)$ in $\mathcal{M}$.
Then $p(x, y, w)$ identifiable from $p(x, y \mid w)$ if and only if
$p(y, w)$ identifiable from $p(y \mid w)$

In other words, $p(x \mid y, w)$ doesn't help us to identify $p(w)$.

This sort of variation independence is also called a **parameter cut** between (Y, W) and $X|Y, W$.

Parameter Cuts

Proposition

Suppose $p(x \mid y, w)$ is variation independent of $p(y, w)$ in $\mathcal{M}$.
Then $p(x, y, w)$ identifiable from $p(x, y \mid w)$ if and only if
$p(y, w)$ identifiable from $p(y \mid w)$

In other words, $p(x \mid y, w)$ doesn't help us to identify $p(w)$.

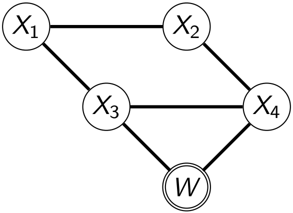
This sort of variation independence is also called a **parameter cut** between (Y, W) and $X|Y, W$.

Corollary

If $p(x \mid w)$ is variation independent of $p(w)$, then $p(x, w)$ is **not** identifiable from $p(x \mid w)$.

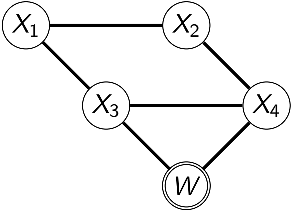
Example

Any undirected (in fact hierarchical) model is therefore not identified:



Example

Any undirected (in fact hierarchical) model is therefore not identified:

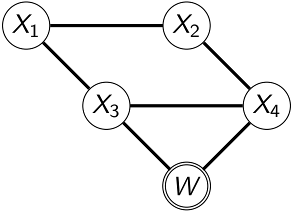


$$p(x_1, x_2, x_3, x_4, w) = \psi_{12}(x_1, x_2) \cdot \psi_{24}(x_2, x_4) \cdot \psi_{34w}(x_3, x_4, w) \cdot \psi_{13}(x_1, x_3).$$

Note that if I multiply by $1/p(w)$, the structure of the RHS is preserved.

Example

Any undirected (in fact hierarchical) model is therefore not identified:



$$p(x_1, x_2, x_3, x_4, w) = \psi_{12}(x_1, x_2) \cdot \psi_{24}(x_2, x_4) \cdot \psi_{34w}(x_3, x_4, w) \cdot \psi_{13}(x_1, x_3).$$

Note that if I multiply by $1/p(w)$, the structure of the RHS is preserved.
So no 'destroyed' structure to try to recover!

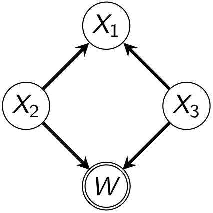
Bayesian Networks

Lemma

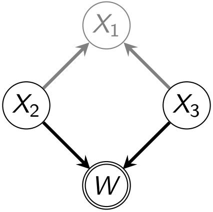
Let $\mathcal{M}(\mathcal{G})$ be a Bayesian network model over a DAG $\mathcal{G}$ with vertex w. Then $p(\boldsymbol{x}_V, x_w)$ is identifiable from $p(\boldsymbol{x}_v \mid x_w)$ if and only if it is identifiable from $p(\boldsymbol{x}_{\mathrm{an}(w)} \mid x_w)$.

That is, we can ignore any non-ancestors of w (in any member of the Markov equivalence class).

Example

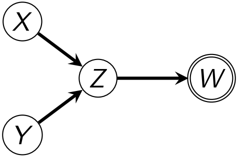


Example



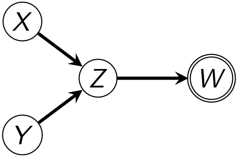
Reduces to the marginal independence model.

Choke Points



Suppose W has three states, but Z only two.

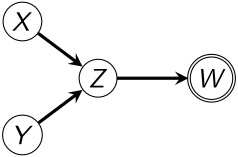
Choke Points



Suppose W has three states, but Z only two.

$$p(x, y, z \mid w) = p(x, y \mid z)p(z \mid w).$$

Choke Points



Suppose W has three states, but Z only two.

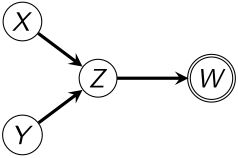
$$p(x, y, z \mid w) = p(x, y \mid z)p(z \mid w).$$

Now, $X \perp\!\!\!\perp Y$ can be used to determine $p(z)$ as before, but

$$\left\{ \alpha(w) : \sum_{w} \alpha(w)p(z \mid w) = p(z) \right\}$$

is an under-determinied linear system.

Choke Points



Suppose W has three states, but Z only two.

$$p(x, y, z \mid w) = p(x, y \mid z)p(z \mid w).$$

Now, $X \perp\!\!\!\perp Y$ can be used to determine $p(z)$ as before, but

$$\left\{\alpha(w) : \sum_{w} \alpha(w)p(z \mid w) = p(z)\right\}$$

is an under-determinied linear system. So $p(w)$ unidentifiable.

This is a more subtle kind of 'unfaithfulness'.

Outline

Causal Learning

Schölkopf et al. (2013) look at semi-supervised learning:
few samples from $p(x, y)$, many from $p(x)$.

Causal Learning

Schölkopf et al. (2013) look at semi-supervised learning:
few samples from $p(x, y)$, many from $p(x)$.

Their conclusions:

Causal	Anti-Causal
$X \rightarrow Y$	$X \leftarrow Y$
poor performance	good performance

Causal Learning

Schölkopf et al. (2013) look at semi-supervised learning:
few samples from $p(x, y)$, many from $p(x)$.

Their conclusions:

Causal	Anti-Causal
$X \to Y$	$X \leftarrow Y$
poor performance	good performance

separation of input and causal mechanisms:

parameter cut X, $Y|X$ parameter cut Y, $X|Y$

Note that parameter cut X, $Y|X$ means $p(x)$ gives no information about $p(y \mid x)$.

Summary

- Detection of and recovery from selection bias is possible in causal models.

Summary

- Detection of and recovery from selection bias is possible in causal models.
- Could in principle be used for causal discovery.

Summary

- Detection of and recovery from selection bias is possible in causal models.
- Could in principle be used for causal discovery.

Some limitations:

Summary

- Detection of and recovery from selection bias is possible in causal models.
- Could in principle be used for causal discovery.

Some limitations:

- Sample size needed may be quite large if selection is dramatic.

Summary

- Detection of and recovery from selection bias is possible in causal models.
- Could in principle be used for causal discovery.

Some limitations:

- Sample size needed may be quite large if selection is dramatic.
- Constraints are hard to characterize;

Summary

- Detection of and recovery from selection bias is possible in causal models.
- Could in principle be used for causal discovery.

Some limitations:

- Sample size needed may be quite large if selection is dramatic.
- Constraints are hard to characterize;
- Model is irregular, and likelihood seems hard to maximize in practice.

References

Bareinboim, Tian and Pearl. Recovering from selection bias in causal and statistical inference, *AAAI-14*, 2014.

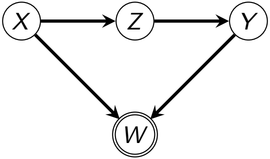
Borboudakis and Tsamardinos. Bayesian Network Learning with Discrete Case-Control Data, *UAI* 2015.

Bowden and Vansteelandt. Mendelian randomization analysis of case-control data using structural mean models. *Stats in Medicine*, 2010.

Moerkerke, Vansteelandt and Lange. A doubly robust test for gene–environment interaction in family-based studies of affected offspring. *Biostatistics*, (2010).

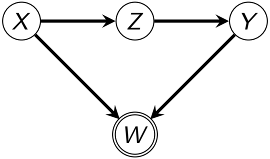
Schölkopf et al. Semi-supervised Learning in Causal and Anticausal Settings. *Empirical Inference*, Springer, 2013.

Degenerate Conditional Independence



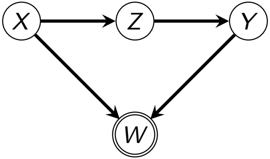
This case is not covered by the other results directly.

Degenerate Conditional Independence



This case is not covered by the other results directly.
Can reduce to marginal independence case by considering $p(x, y \mid z, w)$ for fixed levels of $Z = z$.

Degenerate Conditional Independence



This case is not covered by the other results directly.
Can reduce to marginal independence case by considering $p(x, y \mid z, w)$ for fixed levels of $Z = z$.

In fact: each level of Z gives the same equations, so this is equivalent to case of marginal independence $X \perp\!\!\!\perp Y$.