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

1. Introduction
2. Formalities
3. Parameter Cuts
4. Conclusions
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.
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 $W$.

\[
X \xrightarrow{} W \xrightarrow{} S
\]
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 $p(w \mid x) = p(w \mid \text{do}(x))$. Well known that we can recover and use the causal 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 $W$.

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 $W$.

$$X \rightarrow W$$

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.
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)
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)

\[ X \perp Y \] but generally $X \not\perp Y \mid W$ due to ‘explaining away’.

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

\[
\sum_w p(w) \cdot p(x, y \mid w) = f(x) \cdot g(y).
\]
Concrete Example

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

<table>
<thead>
<tr>
<th>$W = 0$</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.4</td>
<td>0.1</td>
</tr>
<tr>
<td>1</td>
<td>0.1</td>
<td>0.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$W = 1$</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.2</td>
<td>0.3</td>
</tr>
<tr>
<td>1</td>
<td>0.3</td>
<td>0.2</td>
</tr>
</tbody>
</table>
Concrete Example

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

<table>
<thead>
<tr>
<th>$W = 0$</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.4</td>
<td>0.1</td>
</tr>
<tr>
<td>1</td>
<td>0.1</td>
<td>0.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$W = 1$</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.2</td>
<td>0.3</td>
</tr>
<tr>
<td>1</td>
<td>0.3</td>
<td>0.2</td>
</tr>
</tbody>
</table>

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

some unknown $\alpha$. 
Concrete Example
Suppose we observe +ve correlation under $W = 0$, −ve given $W = 1$.

\[
\begin{array}{c|cc}
W = 0 & 0 & 1 \\
\hline
0 & 0.4 & 0.1 \\
1 & 0.1 & 0.4 \\
\end{array}
\quad
\begin{array}{c|cc}
W = 1 & 0 & 1 \\
\hline
0 & 0.2 & 0.3 \\
1 & 0.3 & 0.2 \\
\end{array}
\]

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

Mixture is:

\[
\begin{array}{|c|cc}
| & 0 & 1 \\
\hline
0 & 0.2 + 0.2\alpha & 0.3 - 0.2\alpha \\
1 & 0.3 - 0.2\alpha & 0.2 + 0.2\alpha \\
\end{array}
\]

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 $W = 1$.

<table>
<thead>
<tr>
<th>$W = 0$</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.4</td>
<td>0.1</td>
</tr>
<tr>
<td>1</td>
<td>0.1</td>
<td>0.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$W = 1$</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.2</td>
<td>0.3</td>
</tr>
<tr>
<td>1</td>
<td>0.3</td>
<td>0.2</td>
</tr>
</tbody>
</table>

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

Mixture is:

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.25</td>
<td>0.25</td>
</tr>
<tr>
<td>1</td>
<td>0.25</td>
<td>0.25</td>
</tr>
</tbody>
</table>

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

Surface of independence in $2 \times 2$ 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)
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)

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

1 Introduction
2 Formalities
3 Parameter Cuts
4 Conclusions
Identifiability

Let $p_\theta : \Theta \rightarrow \mathcal{M}$ be a map from a parameter space $\Theta$ to a collection of probability distributions $\mathcal{M}$. So 'almost everywhere' a $k$-to-one map.
Identifiability

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

Say that $\theta$ is **generically $k$-identifiable** if the fibers

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

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

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

Say that $\theta$ is **generically $k$-identifiable** if the fibers

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

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

Overall: generically 2-identifiable.
Main Result

Discrete random variables $X, W$, with $d_x, d_w$ states.

$$p(x, w) = (p(x), p(w \mid x))$$

$$\in \mathcal{M}_X \times \mathcal{M}_{W \mid X}.$$
Main Result

Discrete random variables $X, W$, with $d_x, d_w$ states.

$$p(x, w) = (p(x), p(w | x))$$

$$\in \mathcal{M}_X \times \mathcal{M}_{W|X}.$$ 

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

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

Theorem

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

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, $\ell$ constraints.
Main Result

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

$$p(x, w) = (p(x), p(w | x))$$

$$\in M_X \times M_{W|X}.$$

Separate marginal model for $\mathbf{X}$ and conditional model for $\mathbf{W} | \mathbf{X}$.

Want conditions on $M_X$ and $M_{W|X}$ that lead to (generic) $k$-identifiability of $p(x, w)$ from $p(x | w)$.
Main Result

Discrete random variables $X, W$, with $d_x, d_w$ states.

$$p(x, w) = (p(x), p(w | x)) \in M_X \times M_{W|X}.$$ 

Separate marginal model for $X$ and conditional model for $W | X$.

Want conditions on $M_X$ and $M_{W|X}$ that lead to (generic) $k$-identifiability of $p(x, w)$ from $p(x | w)$.

Theorem

Suppose

- $M_{W|X}$ unrestricted;
- $M_X$ has codimension $\ell$.

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

i.e. $d_w - 1$ unknowns, $\ell$ 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 \textbf{just identified}
(generically up to 2 solutions).
Example: Conditional Independence

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

In this case marginal model $X \perp Y \mid Z$, 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.
But more is true!
Example: Conditional Independence

But more is true!

\[ p(z) \cdot p(x, y, z) - p(x, z) \cdot p(y, z) = 0, \quad \forall x, y, z. \]
Example: Conditional Independence

But more is true!

\[ p(z) \cdot p(x, y, z) - p(x, z) \cdot p(y, z) = 0, \quad \forall x, y, z. \]

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

![Diagram of random variables X, Z, Y, W]

But more is true!

\[ p(z) \cdot p(x, y, z) - p(x, z) \cdot p(y, z) = 0, \quad \forall x, y, z. \]

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.

\[ \implies \text{we have a degree of freedom to test this model}. \]
Example: Conditional Independence

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

\[
\sum_{x, w} n(x, w) \log p(x | w) = \sum_{x, w} n(x, w) \log p(x, w) - \sum_{w} n(w) \log p(w),
\]

use a likelihood ratio test.
Example: Conditional Independence

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

$$\sum_{x,w} n(x, w) \log p(x | w) = \sum_{x,w} n(x, w) \log p(x, w) - \sum_w n(w) \log p(w),$$

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:

![Bayesian Network Diagram]

For binary variables this has codimension 19. Of course, we could then recover appropriate causal effects from the joint. This may appear to contradict Bareinboim and Tian (2015), but they require strict identifiability.
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

1 Introduction

2 Formalities

3 Parameter Cuts

4 Conclusions
Variation Independence

Beware additional independences!

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

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.
Variation Independence

Beware additional independences!

In this case

\[ p(x, y, z, w) = p(w) \cdot p(y, z | w) \cdot p(x | z) \]
Variation Independence

Beware additional independences!

In this case

\[ p(x, y, z, w) = p(w) \cdot p(y, z \mid w) \cdot p(x \mid z) \]
\[ p(x, y, z \mid w) = \quad p(y, z \mid w) \cdot p(x \mid z). \]

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!

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

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

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

Therefore \(p(w)\) is clearly unidentifiable.
These results are all **generic**.
There are areas of the joint distribution which need to be avoided
(think of these as faithfulness conditions).
These results are all **generic**. There are areas of the joint distribution which need to be avoided (think of these as faithfulness conditions).

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 \( p(y \mid w) \)

\[ \text{Y} \xrightarrow{} \text{X} \xrightarrow{} \text{W} \quad \text{Y} \xrightarrow{} \text{X} \xrightarrow{} \text{W} \]
**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 $p(w)$. 
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 $X \mid Y, W$. 

Parameter Cuts
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 \( p(y \mid w) \).

\[
\begin{array}{c}
Y \\
\rightarrow \\
X \\
\rightarrow \\
W
\end{array}
\quad
\begin{array}{c}
Y \\
X \\
\rightarrow \\
W
\end{array}
\]

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 \(X \mid Y, W\).

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:

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

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

Let $\mathcal{M}(\mathcal{G})$ be a Bayesian network model over a DAG $\mathcal{G}$ with vertex $w$. Then $p(x_V, x_w)$ is identifiable from $p(x_V \mid x_w)$ if and only if it is identifiable from $p(x_{\text{an}(w)} \mid x_w)$. That is, we can ignore any non-ancestors of $w$ (in any member of the Markov equivalence class).
Example

Reduces to the marginal independence model.
Example

Reduces to the marginal independence model.
Suppose $W$ has three states, but $Z$ only two.
Suppose \( W \) has three states, but \( Z \) only two.

\[
p(x, y, z \mid w) = p(x, y \mid z)p(z \mid w).
\]
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, $X \perp Y$ 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-determined linear system.
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-determined linear system. So $p(w)$ unidentifiable.

This is a more subtle kind of ‘unfaithfulness’.
Outline

1 Introduction

2 Formalities

3 Parameter Cuts

4 Conclusions
Causal Learning

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

<table>
<thead>
<tr>
<th>Causal</th>
<th>Anti-Causal</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X \rightarrow Y$</td>
<td>$X \leftarrow Y$</td>
</tr>
</tbody>
</table>

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

28 / 31
Causal Learning

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

Their conclusions:

- **Causal**
  - $X \rightarrow Y$
  - poor performance

- **Anti-Causal**
  - $X \leftarrow Y$
  - 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:

<table>
<thead>
<tr>
<th>Causal</th>
<th>Anti-Causal</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X \rightarrow Y$</td>
<td>$X \leftarrow Y$</td>
</tr>
<tr>
<td>poor performance</td>
<td>good performance</td>
</tr>
</tbody>
</table>

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 $p(y|x)$. 
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.

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.


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 $X \perp\!\!\!\!\!\!\!\!\perp Y$. 