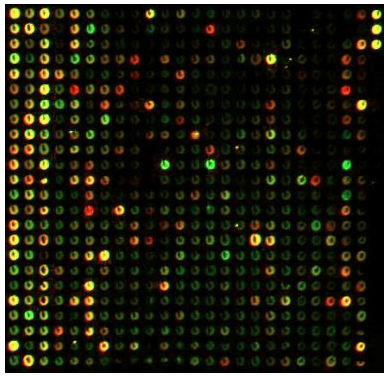
Conditional independence:

Graphical models, causal inference and double machine learning

Rajen D. Shah (Statistical Laboratory, University of Cambridge)

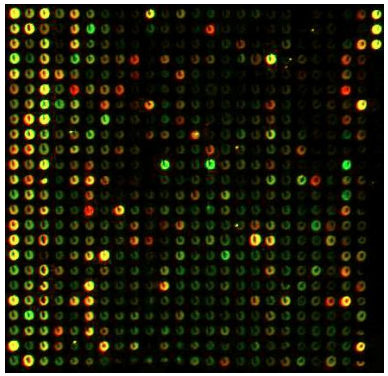
CUSO Workshop
2–5 February 2025

Analysis of multivariate data



- When faced with multivariate data, it is of interest to understand which variables are 'related' to one another.

Analysis of multivariate data



- When faced with multivariate data, it is of interest to understand which variables are 'related' to one another.
- One approach: regard variables as related if they are *dependent*. Recall X and Y are *independent*, written $X \perp\!\!\!\perp Y$ if

$$\mathbb{P}(X \in B_X, Y \in B_Y) = \mathbb{P}(X \in B_X)\,\mathbb{P}(Y \in B_Y)$$

for all measurable sets B_X, B_Y.

Correlation

- Recall that when $(X, Y) \in \mathbb{R} \times \mathbb{R}$ are jointly Gaussian, then $X \perp\!\!\!\perp Y$ if and only if $\mathrm{Cov}(X, Y) = 0$.
- Convenient measure of dependence is the *correlation*:

$$\rho := \mathrm{Corr}(X, Y) := \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$
$$= \frac{\mathbb{E}[\{X - \mathbb{E}(X)\}\{Y - \mathbb{E}(Y)\}]}{\sqrt{\mathbb{E}\{X - \mathbb{E}(X^2)\}\mathbb{E}\{Y - \mathbb{E}(Y^2)\}}} \in [-1, 1].$$

$$\text{Always have} \qquad X \perp\!\!\!\perp Y \implies \rho = 0.$$
$$\text{When } (X, Y) \text{ are Gaussian,} \qquad X \perp\!\!\!\perp Y \iff \rho = 0.$$

- Sample version $\hat{\rho}$ replaces each expectation with an empirical average. This coincides with the MLE in a Gaussian model.

Independence tests

If $(X_1, Y_1), \ldots, (X_n, Y_n)$ are i.i.d. copies of (X, Y) and $X \perp\!\!\!\perp Y$, then for any permutation $\pi : [n] \to [n]$,

$$\begin{pmatrix} X_1, Y_1 \\ X_2, Y_2 \\ \vdots \\ X_n, Y_n \end{pmatrix} \stackrel{d}{=} \begin{pmatrix} X_{\pi(1)}, Y_1 \\ X_{\pi(2)}, Y_2 \\ \vdots \\ X_{\pi(n)}, Y_n \end{pmatrix}.$$

Independence tests

If $(X_1, Y_1), \ldots, (X_n, Y_n)$ are i.i.d. copies of (X, Y) and $X \perp\!\!\!\perp Y$, then for any permutation $\pi : [n] \to [n]$,

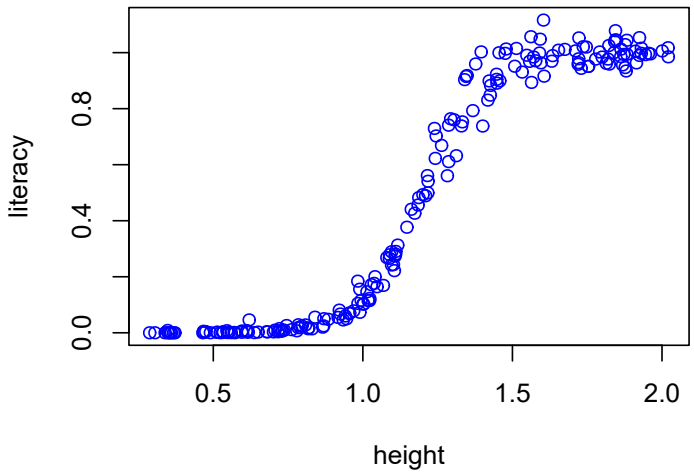
$$\begin{pmatrix} X_1, Y_1 \\ X_2, Y_2 \\ \vdots \\ X_n, Y_n \end{pmatrix} \stackrel{d}{=} \begin{pmatrix} X_{\pi(1)}, Y_1 \\ X_{\pi(2)}, Y_2 \\ \vdots \\ X_{\pi(n)}, Y_n \end{pmatrix}.$$

Permutation test: Let T be some test statistic (e.g. $|\hat{\rho}|$), and let $T_1, \ldots, T_B$ be test statistics calculated on permuted data. If $U \sim U[0, 1]$, then

$$\frac{U + \sum_{b=1}^{B} \mathbb{1}_{\{T_b \geq T\}}}{B + 1} \sim U[0, 1].$$

Conditional independence

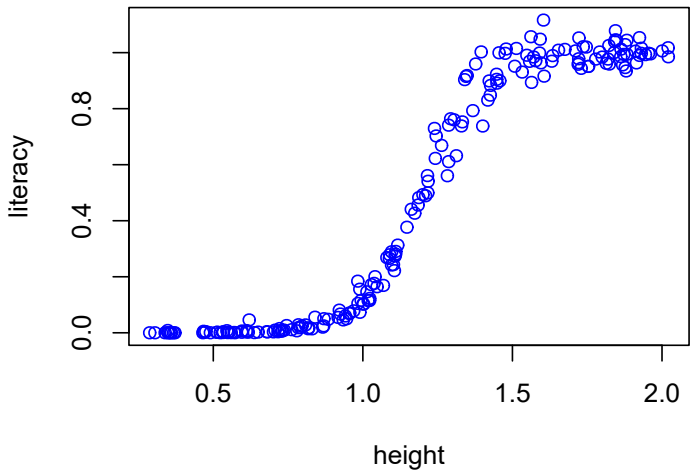
Statistical formalisms of (un)correlatedness or (in)dependence can be less useful.



E.g., 'Height' and 'literacy' would not be independent...

Conditional independence

Statistical formalisms of (un)correlatedness or (in)dependence can be less useful.



E.g., 'Height' and 'literacy' would not be independent...

...but they would expected to be *conditionally independent* given 'age'.

Outline

- Fundamental properties
- Undirected graphical models
- Structural causal models and directed graphical models
- Estimation and testing

Definitions and fundamental properties

Definition

We say random variables X and Y are *conditionally independent* given a random variable Z, and write

$$X \perp\!\!\!\perp Y \,|\, Z,$$

if for all measurable sets B_X, B_Y we have

$$\mathbb{P}(X \in B_X, Y \in B_Y \,|\, Z) = \mathbb{P}(X \in B_X \,|\, Z)\, \mathbb{P}(Y \in B_Y \,|\, Z).$$

Definitions and fundamental properties

Definition

We say random variables X and Y are *conditionally independent* given a random variable Z, and write

$$X \perp\!\!\!\perp Y \,|\, Z,$$

if for all measurable sets B_X, B_Y we have

$$\mathbb{P}(X \in B_X, Y \in B_Y \,|\, Z) = \mathbb{P}(X \in B_X \,|\, Z)\, \mathbb{P}(Y \in B_Y \,|\, Z).$$

If X and Y are not conditionally independent given Z, then they are *conditionally dependent*, which we denote by $X \not\perp\!\!\!\perp Y \,|\, Z$.

Definitions and fundamental properties

Definition

We say random variables X and Y are *conditionally independent* given a random variable Z, and write

$$X \perp\!\!\!\perp Y \,|\, Z,$$

if for all measurable sets B_X, B_Y we have

$$\mathbb{P}(X \in B_X, Y \in B_Y \,|\, Z) = \mathbb{P}(X \in B_X \,|\, Z)\, \mathbb{P}(Y \in B_Y \,|\, Z).$$

If X and Y are not conditionally independent given Z, then they are *conditionally dependent*, which we denote by $X \not\perp\!\!\!\perp Y \,|\, Z$.

Proposition

$X \perp\!\!\!\perp Y \,|\, Z$ if and only if for all measurable B

$$\mathbb{P}(X \in B \,|\, Y, Z) = \mathbb{P}(X \in B \,|\, Z).$$

Fundamental properties

Informally: 'Knowing Z renders Y irrelevant for learning about X'.

Fundamental properties

Informally: 'Knowing Z renders Y irrelevant for learning about X'.

If X, Y and Z have a joint density with respect to Lebesgue measure, then

$$\begin{aligned} X \perp\!\!\!\perp Y \mid Z &\iff f_{XY|Z}(x, y|z) = f_{X|Z}(x|z) f_{Y|Z}(y|z) \\ &\iff f_{X|YZ}(x|y, z) = f_{X|Z}(x|z). \end{aligned}$$

New conditional independencies from old ones

- **Weak union:**

$$X \perp\!\!\!\perp Y, Z \implies \begin{cases} X \perp\!\!\!\perp Y \mid Z \\ X \perp\!\!\!\perp Z \mid Y. \end{cases}$$

'If (Y, Z) are irrelevant for learning X, then Y remains unhelpful for learning X even after knowing Z'.

New conditional independencies from old ones

- **Weak union:**

$$X \perp\!\!\!\perp Y, Z \implies \begin{cases} X \perp\!\!\!\perp Y \mid Z \\ X \perp\!\!\!\perp Z \mid Y. \end{cases}$$

'If (Y, Z) are irrelevant for learning X, then Y remains unhelpful for learning X even after knowing Z'.

- **Contraction:**

$$\left.\begin{array}{l} X \perp\!\!\!\perp Z \\ X \perp\!\!\!\perp Y \mid Z \end{array}\right\} \implies X \perp\!\!\!\perp Y, Z$$

'In seeking to learn X, if Z is irrelevant, and if Y is irrelevant after knowing Z, then (Y, Z) must have been irrelevant to begin with.'

New conditional independencies from old ones

- **Weak union:**

$$X \perp\!\!\!\perp Y, Z \mid W \implies \begin{cases} X \perp\!\!\!\perp Y \mid Z, W \\ X \perp\!\!\!\perp Z \mid Y, W. \end{cases}$$

'If (Y, Z) are irrelevant for learning X, then Y remains unhelpful for learning X even after knowing Z'.

- **Contraction:**

$$\left.\begin{array}{l} X \perp\!\!\!\perp Z \mid W \\ X \perp\!\!\!\perp Y \mid Z, W \end{array}\right\} \implies X \perp\!\!\!\perp Y, Z \mid W$$

'In seeking to learn X, if Z is irrelevant, and if Y is irrelevant after knowing Z, then (Y, Z) must have been irrelevant to begin with.'

New conditional independencies from old ones

- **Weak union:**

$$X \perp\!\!\!\perp Y, Z \mid W \Longrightarrow \begin{cases} X \perp\!\!\!\perp Y \mid Z, W \\ X \perp\!\!\!\perp Z \mid Y, W. \end{cases}$$

'If (Y, Z) are irrelevant for learning X, then Y remains unhelpful for learning X even after knowing Z'.

- **Contraction:**

$$\left.\begin{array}{l} X \perp\!\!\!\perp Z \mid W \\ X \perp\!\!\!\perp Y \mid Z, W \end{array}\right\} \Longrightarrow X \perp\!\!\!\perp Y, Z \mid W$$

'In seeking to learn X, if Z is irrelevant, and if Y is irrelevant after knowing Z, then (Y, Z) must have been irrelevant to begin with.'

- **Intersection:**

$$\left.\begin{array}{l} X \perp\!\!\!\perp Y \mid Z, W \\ X \perp\!\!\!\perp Z \mid Y, W \end{array}\right\} \Longrightarrow X \perp\!\!\!\perp Y, Z \mid W.$$

This holds when X, Y, Z, W has a joint density and the marginal density f_{YZW} is positive everywhere.

Graphs

A *graph* is a pair $\mathcal{G} = (V, E)$ where

- $V = [p]$ is a set of *vertices* or *nodes*
- $E \subseteq V \times V$ with $(v, v) \notin E$ for any $v \in V$ is a set of *edges*.

Graphs

A *graph* is a pair $\mathcal{G} = (V, E)$ where

- $V = [p]$ is a set of *vertices* or *nodes*
- $E \subseteq V \times V$ with $(v, v) \notin E$ for any $v \in V$ is a set of *edges*.

Let $j, k \in V$.

- We say there is an *edge* between j and k and that j and k are *adjacent* if either $(j, k) \in E$ or $(k, j) \in E$.
- An edge (j, k) is *undirected* if also $(k, j) \in E$ and we write $j - k$ to indicate this.
- Otherwise it is *directed* and we write $j \to k$ to represent this. We say j is a *parent* of k and k is a *child* of j. We write $\text{pa}(k)$ for the set of parents of k.

Graphs

A *graph* is a pair $\mathcal{G} = (V, E)$ where

- $V = [p]$ is a set of *vertices* or *nodes*
- $E \subseteq V \times V$ with $(v, v) \notin E$ for any $v \in V$ is a set of *edges*.

Let $j, k \in V$.

- We say there is an *edge* between j and k and that j and k are *adjacent* if either $(j, k) \in E$ or $(k, j) \in E$.
- An edge (j, k) is *undirected* if also $(k, j) \in E$ and we write $j - k$ to indicate this.
- Otherwise it is *directed* and we write $j \to k$ to represent this. We say j is a *parent* of k and k is a *child* of j. We write $\text{pa}(k)$ for the set of parents of k.
- If all edges in the graph are (un)directed we call it an *(un)directed graph*.

Graphs

- A *path* from j to k is a sequence $j = j_1, j_2, \ldots, j_m = k$ of (at least two) distinct vertices such that j_l and j_{l+1} are adjacent.

Graphs

- A *path* from j to k is a sequence $j = j_1, j_2, \ldots, j_m = k$ of (at least two) distinct vertices such that j_l and j_{l+1} are adjacent.
- It is a *directed path* if $j_l \to j_{l+1}$ for all l. We then call k a *descendant* of j. The set of descendants of j will be denoted $\mathrm{de}(j)$.

Graphs

- A *path* from j to k is a sequence $j = j_1, j_2, \ldots, j_m = k$ of (at least two) distinct vertices such that j_l and j_{l+1} are adjacent.
- It is a *directed path* if $j_l \to j_{l+1}$ for all l. We then call k a *descendant* of j. The set of descendants of j will be denoted $\mathrm{de}(j)$.
- A *directed cycle* is (almost) a directed path but with the start and end points the same.
- A *directed acyclic graph (DAG)* is a directed graph containing no directed cycles.

Undirected graphical models

Notation:

- For $v \in \mathbb{R}^p$ and $S \subseteq [p]$, $v_S \in \mathbb{R}^{|S|}$ is the subvector of v whose components are indexed by S.

Undirected graphical models

Notation:

- For $v \in \mathbb{R}^p$ and $S \subseteq [p]$, $v_S \in \mathbb{R}^{|S|}$ is the subvector of v whose components are indexed by S.
- $-j$ and $-jk$ in subscripts are shorthand for $[p] \setminus \{j\}$ and $[p] \setminus \{j, k\}$ respectively.

Undirected graphical models

Notation:

- For $v \in \mathbb{R}^p$ and $S \subseteq [p]$, $v_S \in \mathbb{R}^{|S|}$ is the subvector of v whose components are indexed by S.
- $-j$ and $-jk$ in subscripts are shorthand for $[p] \setminus \{j\}$ and $[p] \setminus \{j, k\}$ respectively.
- When $A, B \subseteq [p]$ are disjoint and $S = \emptyset$, for a random vector $Z \in \mathbb{R}^p$, we interpret $Z_A \perp\!\!\!\perp Z_B \,|\, Z_S$ as the (unconditional) independence relationship $Z_A \perp\!\!\!\perp Z_B$.

Undirected graphical models

Notation:

- For $v \in \mathbb{R}^p$ and $S \subseteq [p]$, $v_S \in \mathbb{R}^{|S|}$ is the subvector of v whose components are indexed by S.
- $-j$ and $-jk$ in subscripts are shorthand for $[p] \setminus \{j\}$ and $[p] \setminus \{j, k\}$ respectively.
- When $A, B \subseteq [p]$ are disjoint and $S = \emptyset$, for a random vector $Z \in \mathbb{R}^p$, we interpret $Z_A \perp\!\!\!\perp Z_B \,|\, Z_S$ as the (unconditional) independence relationship $Z_A \perp\!\!\!\perp Z_B$.

Definition

The *conditional independence graph* of a distribution P on $\mathbb{R}^p$ with $p \geq 2$ is the undirected graph where given $Z \sim P$, we have for all $j, k \in [p]$ that

$$j - k \iff Z_j \not\perp\!\!\!\perp Z_k \,|\, Z_{-jk}.$$

Undirected graphical models

A second approach relating graphs and conditional independencies asks for the distribution to reflect further aspects of the structure of the graph through the notion of graph separation:

Definition

Given disjoint sets A and B of vertices in a graph, we say a set of vertices S *separates* A and B if every path between A and B contains a node in S.

Definition

A distribution P on $\mathbb{R}^p$ is *global Markov* with respect to an undirected graph $\mathcal{G}$ with p vertices if whenever $Z \sim P$ and A, B and S are disjoint sets of vertices such that S separates A from B, we have $Z_A \perp\!\!\!\perp Z_B \,|\, Z_S$.

Undirected graphical models

Theorem

If a distribution P on $\mathbb{R}^p$ is such that $Z \sim P$ satisfies the intersection property, then it is global Markov with respect to its conditional independence graph.

Undirected graphical models

Theorem

If a distribution P on $\mathbb{R}^p$ is such that $Z \sim P$ satisfies the intersection property, then it is global Markov with respect to its conditional independence graph.

Consider a regression setting with response–predictor pair (Y, X) and suppose their joint distribution satisfies the intersection property.

$$S := \{j : Y \not\perp\!\!\!\perp X_j \,|\, X_{-j}\}$$

is the *Markov blanket* of Y satisfying

$$Y \perp\!\!\!\perp X_{S^c} \,|\, X_S.$$

'X_S contains all the information relevant for learning Y'.

In this way, conditional independence gives a model-free way of formalising variable significance.

Structural causal models

Definition

A structural causal model (SCM) is a system of equations

$$Z_j := h_j(Z_{P_j}, \varepsilon_j)$$

where

- $\varepsilon_1, \ldots, \varepsilon_p$ are independent noise variables
- $P_j \subseteq [p] \setminus \{j\}$ are such that the graph with $P_j = \mathrm{pa}(j)$ is a DAG.

Structural causal models

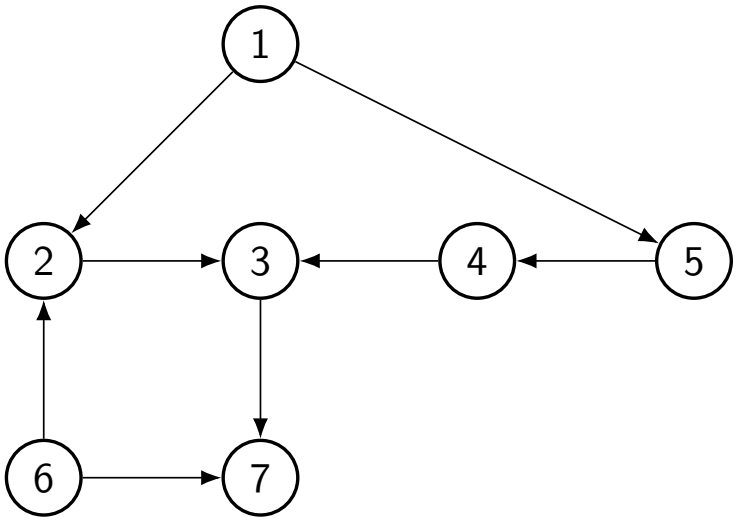
Definition

A structural causal model (SCM) is a system of equations

$$Z_j := h_j(Z_{P_j}, \varepsilon_j)$$

where

- $\varepsilon_1, \ldots, \varepsilon_p$ are independent noise variables
- $P_j \subseteq [p] \setminus \{j\}$ are such that the graph with $P_j = \mathrm{pa}(j)$ is a DAG.



Every DAG has a permutation $\pi : [p] \to [p]$ known as a *topological ordering* where $k \in \mathrm{de}(j) \implies \pi(k) > \pi(j)$.

Structural causal models

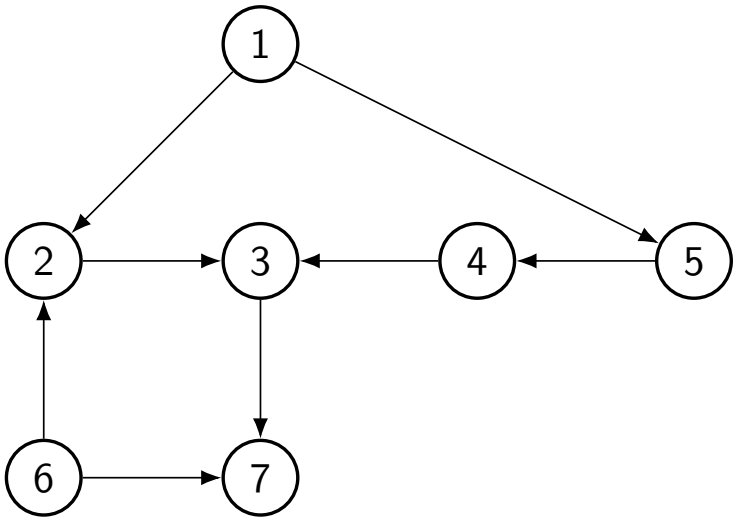
Definition

A structural causal model (SCM) is a system of equations

$$Z_j := h_j(Z_{P_j}, \varepsilon_j)$$

where

- $\varepsilon_1, \ldots, \varepsilon_p$ are independent noise variables
- $P_j \subseteq [p] \setminus \{j\}$ are such that the graph with $P_j = \mathrm{pa}(j)$ is a DAG.



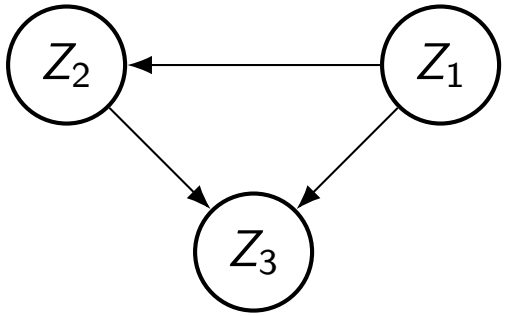
Every DAG has a permutation $\pi : [p] \to [p]$ known as a *topological ordering* where $k \in \mathrm{de}(j) \implies \pi(k) > \pi(j)$.

Then Z_j is a function of $\varepsilon_{\pi^{-1}(1)}, \varepsilon_{\pi^{-1}(2)}, \ldots, \varepsilon_{\pi^{-1}(j)}$, so the SCM gives a recipe for generating Z.

Interventions

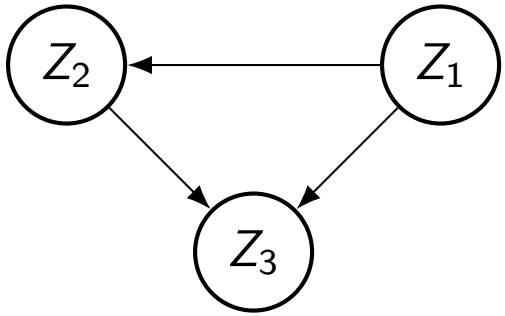
E.g. $Z_1 =$ large kidney stones,
$Z_2 =$ large incision, $Z_3 =$ success.

$$Z_1 = \varepsilon_1, \qquad \varepsilon_1 \sim \text{Bern}(1/2),$$
$$Z_2 = \mathbb{1}_{\{\varepsilon_2(1+2Z_1)>3/4\}}, \qquad \varepsilon_2 \sim U[0,1],$$
$$Z_3 = \mathbb{1}_{\{\varepsilon_3(2-Z_1+7Z_2/4-3Z_1Z_2/2)>1/4\}}, \qquad \varepsilon_3 \sim U[0,1].$$



Interventions

E.g. $Z_1 =$ large kidney stones,
$Z_2 =$ large incision, $Z_3 =$ success.



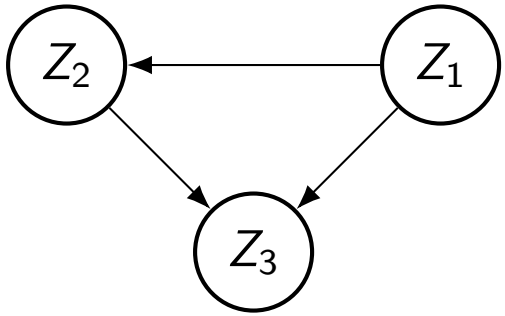
$$
\begin{aligned}
Z_1 &= \varepsilon_1, & \varepsilon_1 &\sim \text{Bern}(1/2),\\
Z_2 &= \mathbb{1}_{\{\varepsilon_2(1+2Z_1)>3/4\}}, & \varepsilon_2 &\sim U[0,1],\\
Z_3 &= \mathbb{1}_{\{\varepsilon_3(2-Z_1+7Z_2/4-3Z_1Z_2/2)>1/4\}}, & \varepsilon_3 &\sim U[0,1].
\end{aligned}
$$

An SCM gives us a way of reasoning about how a joint distribution may change after an *intervention*.

If we intervene by setting e.g. variable j to $a \in \mathbb{R}$, then replacing the assignment for Z_j with $Z_j := a$ gives a new SCM, which determines a new joint distribution.

Interventions

E.g. $Z_1 =$ large kidney stones,
$Z_2 =$ large incision, $Z_3 =$ success.



$$Z_1 = \varepsilon_1, \qquad \varepsilon_1 \sim \text{Bern}(1/2),$$
$$Z_2 = \mathbb{1}_{\{\varepsilon_2(1+2Z_1)>3/4\}}, \qquad \varepsilon_2 \sim U[0,1],$$
$$Z_3 = \mathbb{1}_{\{\varepsilon_3(2-Z_1+7Z_2/4-3Z_1Z_2/2)>1/4\}}, \qquad \varepsilon_3 \sim U[0,1].$$

An SCM gives us a way of reasoning about how a joint distribution may change after an *intervention*.

If we intervene by setting e.g. variable j to $a \in \mathbb{R}$, then replacing the assignment for Z_j with $Z_j := a$ gives a new SCM, which determines a new joint distribution.

We denote expectations / probabilities w.r.t. the new distribution by adding $| \, do(Z_j = a)$ e.g. $\mathbb{E}(\cdot \,|\, do(Z_j = a))$.

Interventions

$$
\begin{aligned}
Z_1 &= \varepsilon_1, \\
Z_2 &= \mathbb{1}_{\{\varepsilon_2(1+2Z_1)>3/4\}}, \\
Z_3 &= \mathbb{1}_{\{\varepsilon_3(2-Z_1+7Z_2/4-3Z_1Z_2/2)>1/4\}}.
\end{aligned}
$$

$$
\begin{aligned}
&do(Z_2 = 1) \\
Z_1 &= \varepsilon_1, \\
Z_2 &= 1, \\
Z_3 &= \mathbb{1}_{\{\varepsilon_3(15/4-5Z_1/2)>1/4\}},
\end{aligned}
$$

Interventions

$$
\begin{aligned}
Z_1 &= \varepsilon_1, \\
Z_2 &= \mathbb{1}_{\{\varepsilon_2(1+2Z_1)>3/4\}}, \\
Z_3 &= \mathbb{1}_{\{\varepsilon_3(2-Z_1+7Z_2/4-3Z_1Z_2/2)>1/4\}}.
\end{aligned}
\qquad
\begin{aligned}
&do(Z_2 = 1) \\
Z_1 &= \varepsilon_1, \\
Z_2 &= 1, \\
Z_3 &= \mathbb{1}_{\{\varepsilon_3(15/4-5Z_1/2)>1/4\}},
\end{aligned}
$$

$\mathbb{P}\big(Z_3 = 1 \,|\, do(Z_2 = 1)\big) > \mathbb{P}\big(Z_3 = 1 \,|\, do(Z_2 = 0)\big) \Longrightarrow$ 'big incisions are preferred'

Interventions

$$
\begin{aligned}
Z_1 &= \varepsilon_1, \\
Z_2 &= \mathbb{1}_{\{\varepsilon_2(1+2Z_1)>3/4\}}, \\
Z_3 &= \mathbb{1}_{\{\varepsilon_3(2-Z_1+7Z_2/4-3Z_1Z_2/2)>1/4\}}.
\end{aligned}
\qquad
\begin{aligned}
&do(Z_2 = 1) \\
Z_1 &= \varepsilon_1, \\
Z_2 &= 1, \\
Z_3 &= \mathbb{1}_{\{\varepsilon_3(15/4-5Z_1/2)>1/4\}},
\end{aligned}
$$

$\mathbb{P}\big(Z_3 = 1 \,|\, do(Z_2 = 1)\big) > \mathbb{P}\big(Z_3 = 1 \,|\, do(Z_2 = 0)\big) \Longrightarrow$ 'big incisions are preferred'

but naive conditioning 'suggests' the opposite conclusion

$$\mathbb{P}(Z_3 = 1 \,|\, Z_2 = 1) < \mathbb{P}(Z_3 = 1 \,|\, Z_2 = 0).$$

Interventions

$$
\begin{aligned}
Z_1 &= \varepsilon_1, \\
Z_2 &= \mathbb{1}_{\{\varepsilon_2(1+2Z_1)>3/4\}}, \\
Z_3 &= \mathbb{1}_{\{\varepsilon_3(2-Z_1+7Z_2/4-3Z_1Z_2/2)>1/4\}}.
\end{aligned}
\qquad
\begin{aligned}
&do(Z_2 = 1) \\
Z_1 &= \varepsilon_1, \\
Z_2 &= 1, \\
Z_3 &= \mathbb{1}_{\{\varepsilon_3(15/4-5Z_1/2)>1/4\}},
\end{aligned}
$$

$\mathbb{P}\big(Z_3 = 1 \,|\, do(Z_2 = 1)\big) > \mathbb{P}\big(Z_3 = 1 \,|\, do(Z_2 = 0)\big) \Longrightarrow$ 'big incisions are preferred'

but naive conditioning 'suggests' the opposite conclusion

$$\mathbb{P}(Z_3 = 1 \,|\, Z_2 = 1) < \mathbb{P}(Z_3 = 1 \,|\, Z_2 = 0).$$

Why do we have this discrepancy here? The presence of large kidney stones increases the chance of open surgery being performed, but also decreases the chance of success of either treatment.

- The example illustrates that the conclusions that may be drawn from an SCM go far beyond those of its associated joint distribution.
- The crucial additional piece of information included in the SCM is the causal structure encoded by the associated DAG.

Hidden confounders

We have assumed:

(i) an intervention in the real world corresponds to the mathematical process of 'do calculus';

(ii) the true data generating mechanism corresponds to a given SCM

These are **strong assumptions**.

Hidden confounders

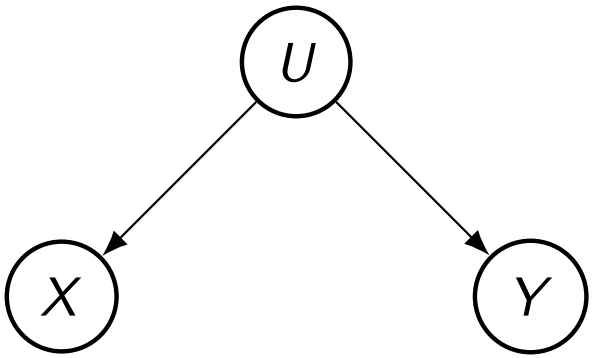
We have assumed:

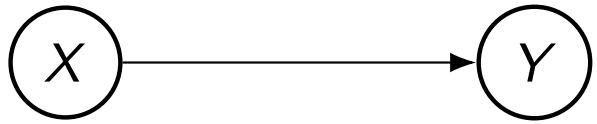
(i) an intervention in the real world corresponds to the mathematical process of 'do calculus';

(ii) the true data generating mechanism corresponds to a given SCM

These are **strong assumptions**.

A classical example of where (ii) would fail is when there are unobserved variables:





If U were unobserved, we might incorrectly postulate the causal DAG on the right.

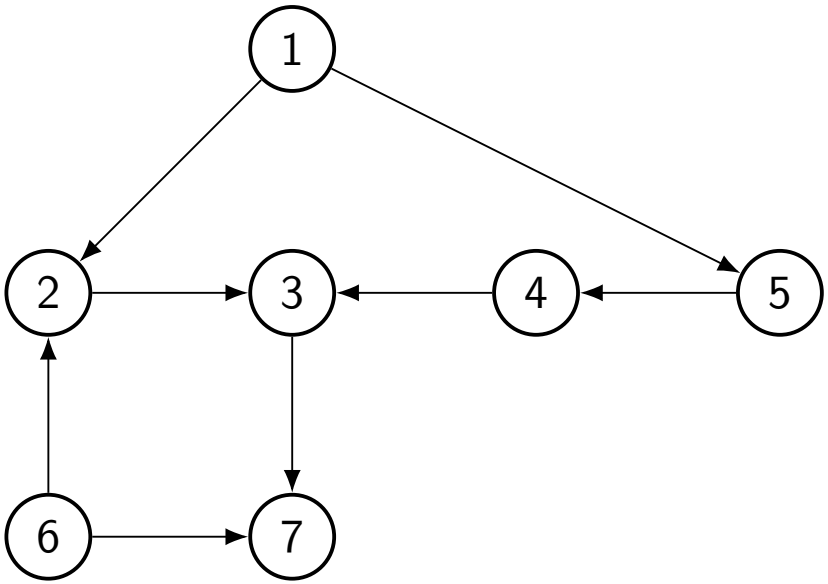
The intervention distributions would then not be indicative of what could be expected in the real world, where intervening on X should have no effect on Y.

d-separation

Given a (not necessarily directed) path $j_1 j_2 \dots j_{m-1} j_m$, we say j_ℓ is a *collider* (relative to the path) if $j_{\ell-1} \to j_\ell \leftarrow j_{\ell+1}$.

The path is *blocked* by a set of nodes S with $j_1, j_m \notin S$ if for some $\ell \in \{2, \dots, m-1\}$ either:

- j_ℓ is not a collider and $j_\ell \in S$, or
- j_ℓ is a collider and neither j_ℓ nor any of its descendants are in S.



Given disjoint subsets of nodes A, B and S, we say A and B are *d-separated* by S, and write

$$A \perp\!\!\!\perp_{\mathcal{G}} B \mid S,$$

if every path between A and B is blocked by S.

Global Markov property for DAGs

Definition

We say a distribution P on $\mathbb{R}^p$ is *global Markov* with respect to a DAG $\mathcal{G}$ on p vertices if whenever $Z \sim P$,

$$A \perp\!\!\!\perp_{\mathcal{G}} B \,|\, S \implies Z_A \perp\!\!\!\perp Z_B \,|\, Z_S$$

for all disjoint sets of vertices A, B and S.

Theorem

If a distribution P is generated by a structural causal model with DAG $\mathcal{G}$, then P is global Markov with respect to $\mathcal{G}$.

Causal structure learning

Edges in a conditional independence graph therefore have the following causal interpretation:

Proposition

Let $Z \sim P$. If $Z_j \not\perp\!\!\!\perp Z_k \,|\, Z_{-jk}$ for $j, k \in [p]$, then any causal DAG that generates P must either have $j \to k$, $j \leftarrow k$ or $j \to \ell \leftarrow k$ for some $\ell \in [p]$.

Causal structure learning

Edges in a conditional independence graph therefore have the following causal interpretation:

Proposition

Let $Z \sim P$. If $Z_j \not\perp\!\!\!\perp Z_k \mid Z_{-jk}$ for $j, k \in [p]$, then any causal DAG that generates P must either have $j \to k$, $j \leftarrow k$ or $j \to \ell \leftarrow k$ for some $\ell \in [p]$.

Proposition

If nodes j and k in a DAG $\mathcal{G}$ are not adjacent, then they are d-separated by $pa(j)$ or $pa(k)$.

Consequently, if $Z_j \not\perp\!\!\!\perp Z_k \mid Z_A$ for every $A \subseteq [p] \setminus \{j, k\}$, then either $Z_j \to Z_k$ or $Z_j \leftarrow Z_k$.

Learning the whole DAG?

Given a complete list of conditional dependencies and independencies, is it possible to pin down the exact causal DAG?

Learning the whole DAG?

Given a complete list of conditional dependencies and independencies, is it possible to pin down the exact causal DAG?

Problem 1: A distribution P for a random pair (X, Y) with $X \perp\!\!\!\perp Y$ could be generated via the SCM with equations

$$X = \varepsilon_1, \qquad Y = 0 \times X + \varepsilon_2,$$

where $\varepsilon_1 \perp\!\!\!\perp \varepsilon_2$, with DAG $X \to Y$.

Learning the whole DAG?

Given a complete list of conditional dependencies and independencies, is it possible to pin down the exact causal DAG?

Problem 1: A distribution P for a random pair (X, Y) with $X \perp\!\!\!\perp Y$ could be generated via the SCM with equations

$$X = \varepsilon_1, \qquad Y = 0 \times X + \varepsilon_2,$$

where $\varepsilon_1 \perp\!\!\!\perp \varepsilon_2$, with DAG $X \to Y$.

Solution 1: A distribution P satisfies *causal minimality* with respect to $\mathcal{G}$ if it is global Markov with respect to $\mathcal{G}$ but not with respect to any proper subgraph of $\mathcal{G}$ with the same nodes.

Learning the whole DAG?

Given a complete list of conditional dependencies and independencies, is it possible to pin down the exact causal DAG?

Problem 1: A distribution P for a random pair (X, Y) with $X \perp\!\!\!\perp Y$ could be generated via the SCM with equations

$$X = \varepsilon_1, \qquad Y = 0 \times X + \varepsilon_2,$$

where $\varepsilon_1 \perp\!\!\!\perp \varepsilon_2$, with DAG $X \to Y$.

Solution 1: A distribution P satisfies *causal minimality* with respect to $\mathcal{G}$ if it is global Markov with respect to $\mathcal{G}$ but not with respect to any proper subgraph of $\mathcal{G}$ with the same nodes.

Problem 2: Two DAGs sharing the same *skeleton* and *v-structures* ($j \to \ell \leftarrow k$ with j, k not adjacent) generate the same list of conditional independencies.

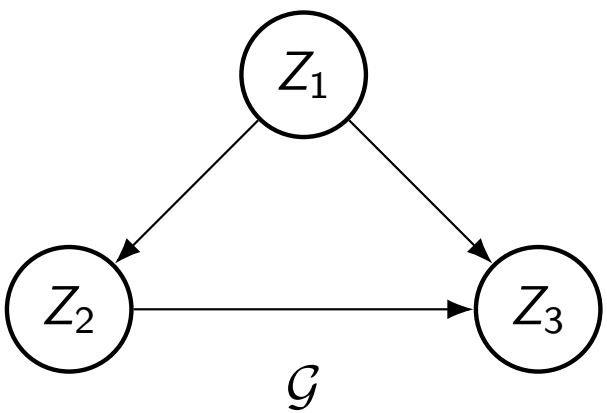
Solution 2: Only seek to learn the the skeleton and v-structures.

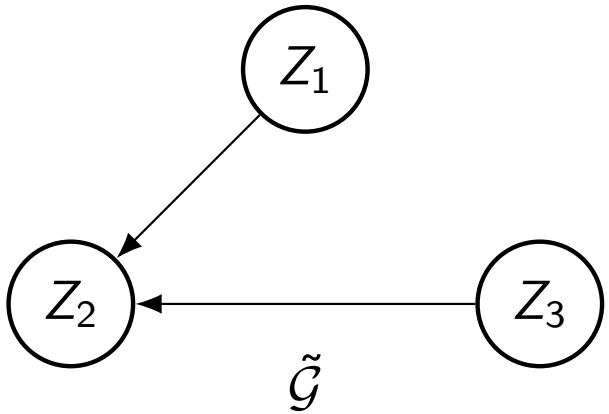
Faithfulness

Consider the SCM with $\varepsilon \sim \mathcal{N}_3(0, I)$ and

$$\begin{aligned} Z_1 &= \varepsilon_1 \\ Z_2 &= \alpha Z_1 + \varepsilon_2 \\ Z_3 &= \beta Z_1 + \gamma Z_2 + \varepsilon_3, \end{aligned}$$

If $\beta = -1$, $\alpha = \gamma = 1$, then $\beta + \alpha\gamma = 0$, so $Z_1 \perp\!\!\!\perp Z_3$.





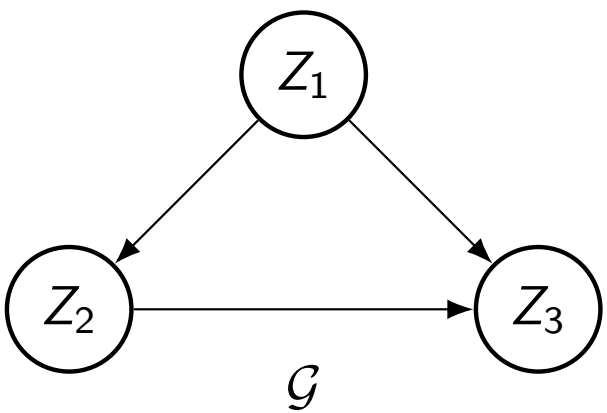
The SCM satisfies causal minimality w.r.t. both $\mathcal{G}$ and $\tilde{\mathcal{G}}$.

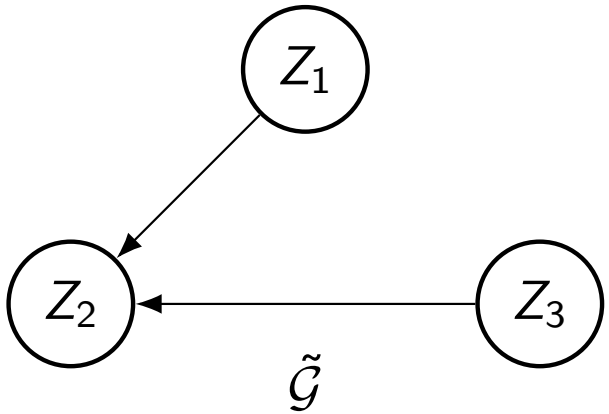
Faithfulness

Consider the SCM with $\varepsilon \sim \mathcal{N}_3(0, I)$ and

$$\begin{aligned} Z_1 &= \varepsilon_1 \\ Z_2 &= \alpha Z_1 + \varepsilon_2 \\ Z_3 &= \beta Z_1 + \gamma Z_2 + \varepsilon_3, \end{aligned}$$

If $\beta = -1$, $\alpha = \gamma = 1$, then $\beta + \alpha\gamma = 0$, so $Z_1 \perp\!\!\!\perp Z_3$.





The SCM satisfies causal minimality w.r.t. both $\mathcal{G}$ and $\tilde{\mathcal{G}}$.

A distribution on $\mathbb{R}^p$ is *faithful* to a DAG $\mathcal{G}$ if $Z_A \perp\!\!\!\perp Z_B \,|\, Z_S \iff A \perp\!\!\!\perp_{\mathcal{G}} B \,|\, S$.

Testing conditional independence

We have seen how conditional independence provides a compelling way to formalise relatedness between variables. We would therefore like to test for conditional dependence given data.

Testing conditional independence

We have seen how conditional independence provides a compelling way to formalise relatedness between variables. We would therefore like to test for conditional dependence given data.

- Null hypothesis $\mathcal{P}$: the collection of distributions for (X, Y, Z) absolutely continuous with respect to Lebesgue measure where $X \perp\!\!\!\perp Y \mid Z$.
- Alternative hypothesis $\mathcal{Q}$: as above but with $X \not\perp\!\!\!\perp Y \mid Z$.
- Data $(x_i, y_i, z_i)_{i=1}^n$ i.i.d. copies of (X, Y, Z).
- ψ_n a potentially randomised test with $\psi_n = 1$ meaning "reject".
- α: significance level.

A "good" test ψ_n should have

$$\sup_{P \in \mathcal{P}} \mathbb{P}_P(\psi_n = 1) \leq \alpha \quad \text{and} \quad \mathbb{P}_Q(\psi_n = 1) \gg \alpha \text{ for many } Q \in \mathcal{Q}$$

Testing conditional independence

Recall the total variation distance is given by

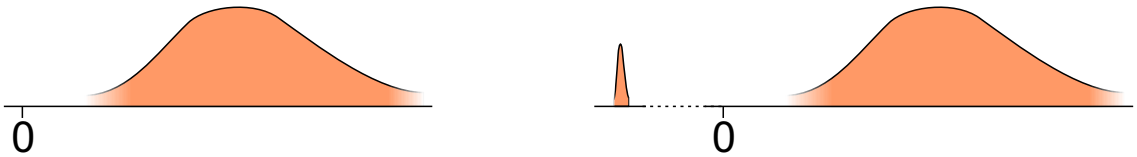
$$\|Q - P\|_{TV} := \sup_{\text{measurable } A} |\mathbb{P}_P(W \in A) - \mathbb{P}_Q(W \in A)|$$
$$= \sup_{\text{test } \psi_1} |\mathbb{P}_P(\psi_1 = 1) - \mathbb{P}_Q(\psi_1 = 1)|.$$

Testing conditional independence

Recall the total variation distance is given by

$$\|Q - P\|_{TV} := \sup_{\text{measurable } A} |\mathbb{P}_P(W \in A) - \mathbb{P}_Q(W \in A)|$$
$$= \sup_{\text{test } \psi_1} |\mathbb{P}_P(\psi_1 = 1) - \mathbb{P}_Q(\psi_1 = 1)|.$$

It is impossible to test whether the mean of a distribution is non-zero:

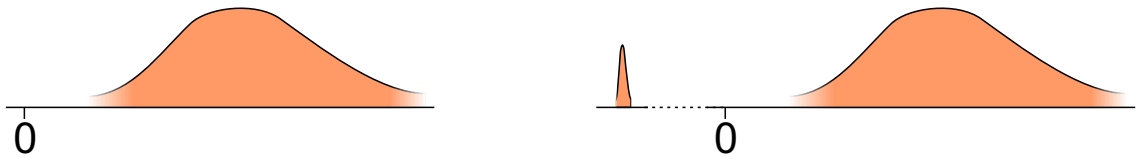


Testing conditional independence

Recall the total variation distance is given by

$$\|Q - P\|_{TV} := \sup_{\text{measurable } A} |\mathbb{P}_P(W \in A) - \mathbb{P}_Q(W \in A)|$$
$$= \sup_{\text{test } \psi_1} |\mathbb{P}_P(\psi_1 = 1) - \mathbb{P}_Q(\psi_1 = 1)|.$$

It is impossible to test whether the mean of a distribution is non-zero:



Setting of CI testing: There exists $Q \in \mathcal{Q}$ such that

$$\inf_{P \in \mathcal{P}} \|Q - P\|_{TV} \geq 1/24.$$

Null and alternative hypotheses are "separated" in terms of KL divergence:

$$\sup_{Q \in \mathcal{Q}} \inf_{P \in \mathcal{P}} \mathrm{KL}(P||Q) = \infty = \sup_{Q \in \mathcal{Q}} \inf_{P \in \mathcal{P}} \mathrm{KL}(Q||P).$$

Hardness of conditional independence testing

Theorem (With great power comes... great Type I error [SP20])

Suppose ψ_n has size α. Then the power at each alternative $Q \in \mathcal{Q}$ is at most α.

Hardness of conditional independence testing

Theorem (With great power comes... great Type I error [SP20])

Suppose ψ_n has size α. Then the power at each alternative $Q \in \mathcal{Q}$ is at most α.

Suppose ψ_n has power β at an alternative $Q \in \mathcal{Q}$. Then there exists null distribution $P \in \mathcal{P}$ such that $\mathbb{P}_P(\psi_n = 1) \geq \beta$.

Hardness of conditional independence testing

Theorem (With great power comes... great Type I error [SP20])

Suppose ψ_n has size α. Then the power at each alternative $Q \in \mathcal{Q}$ is at most α.

Suppose ψ_n has power β at an alternative $Q \in \mathcal{Q}$. Then there exists null distribution $P \in \mathcal{P}$ such that $\mathbb{P}_P(\psi_n = 1) \geq \beta$.

We can only hope to have type I error control over a strict subset of the null of conditional independence.

Modelling assumptions must be imposed in order to restrict the null of conditional independence.

Hardness of conditional independence testing

Theorem (With great power comes... great Type I error [SP20])

Suppose ψ_n has size α. Then the power at each alternative $Q \in \mathcal{Q}$ is at most α.

Suppose ψ_n has power β at an alternative $Q \in \mathcal{Q}$. Then there exists null distribution $P \in \mathcal{P}$ such that $\mathbb{P}_P(\psi_n = 1) \geq \beta$.

We can only hope to have type I error control over a strict subset of the null of conditional independence.

Modelling assumptions must be imposed in order to restrict the null of conditional independence.

We will focus on two strategies for restricting the null:

1. Imposing a parametric model (Gaussianity);
2. Requiring conditional expectations to be estimated sufficiently well.

Normal conditionals

When considering conditional independencies, Gaussianity is a natural choice, not just because of the ubiquity of Gaussianity throughout Statistics, but also because it offers very simple characterisations of conditional independence.

Theorem

Let $Z \sim \mathcal{N}_p(\mu, \Sigma)$ with $\Sigma \in \mathbb{R}^{p \times p}$ positive definite, and let $A, B \subseteq [p]$ be non-empty. The conditional distribution of Z_A given $Z_B = z_B$ is

$$\mathcal{N}_{|A|}\left(\mu_A + \Sigma_{A,B}\Sigma_{B,B}^{-1}(z_B - \mu_B),\ \Sigma_{A,A} - \Sigma_{A,B}\Sigma_{B,B}^{-1}\Sigma_{B,A}\right).$$

Notation: $\Sigma_{A,B}$ is the submatrix of Σ consisting of those rows and columns indexed by A and B respectively.

- Dependence on Z_B is only in the conditional mean.
- Conditional distribution is Gaussian. (Recall that in Gaussian distributions, we have independence $\iff$ uncorrelatedness.)

Nodewise regressions

Specialising to $A = \{k\}$, $B = [p] \setminus \{k\}$:

$$\mathbb{E}(Z_A \,|\, Z_B) = \mu_A + \Sigma_{A,B}\Sigma_{B,B}^{-1}(Z_B - \mu_B).$$

Equivalently

$$Z_k = \underbrace{\mu_k - \Sigma_{k,-k}\Sigma_{-k,-k}^{-1}\mu_{-k}}_{\text{intercept}} + \underbrace{(\Sigma_{-k,-k}^{-1}\Sigma_{-k,k})^\top}_{\text{coef. vector}=:\beta} Z_{-k} + \underbrace{\varepsilon}_{\text{error}}$$

where $\varepsilon \,|\, Z_{-k} \sim \mathcal{N}(0, \Sigma_{k,k} - \Sigma_{k,-k}\Sigma_{-k,-k}^{-1}\Sigma_{-k,k})$.

Nodewise regressions

Specialising to $A = \{k\}$, $B = [p] \setminus \{k\}$:

$$\mathbb{E}(Z_A \,|\, Z_B) = \mu_A + \Sigma_{A,B}\Sigma_{B,B}^{-1}(Z_B - \mu_B).$$

Equivalently

$$Z_k = \underbrace{\mu_k - \Sigma_{k,-k}\Sigma_{-k,-k}^{-1}\mu_{-k}}_{\text{intercept}} + \underbrace{(\Sigma_{-k,-k}^{-1}\Sigma_{-k,k})^\top}_{\text{coef. vector}=:\beta} Z_{-k} + \underbrace{\varepsilon}_{\text{error}}$$

where $\varepsilon \,|\, Z_{-k} \sim \mathcal{N}(0, \Sigma_{k,k} - \Sigma_{k,-k}\Sigma_{-k,-k}^{-1}\Sigma_{-k,k})$.

Key point: If a component of β is zero, then the conditional distribution of Z_k given Z_{-k} will not depend on that component of Z_{-k}. I.e.

$$\text{zeroes in } \beta \iff \text{Conditional independencies}$$

Nodewise regressions

Specialising to $A = \{k\}$, $B = [p] \setminus \{k\}$:

$$\mathbb{E}(Z_A \,|\, Z_B) = \mu_A + \Sigma_{A,B}\Sigma_{B,B}^{-1}(Z_B - \mu_B).$$

Equivalently

$$Z_k = \underbrace{\mu_k - \Sigma_{k,-k}\Sigma_{-k,-k}^{-1}\mu_{-k}}_{\text{intercept}} + \underbrace{(\Sigma_{-k,-k}^{-1}\Sigma_{-k,k})^\top}_{\text{coef. vector}=:\beta} Z_{-k} + \underbrace{\varepsilon}_{\text{error}}$$

where $\varepsilon \,|\, Z_{-k} \sim \mathcal{N}(0, \Sigma_{k,k} - \Sigma_{k,-k}\Sigma_{-k,-k}^{-1}\Sigma_{-k,k})$.

Key point: If a component of β is zero, then the conditional distribution of Z_k given Z_{-k} will not depend on that component of Z_{-k}. I.e.

$$\text{zeroes in } \beta \iff \text{Conditional independencies}$$

$\rightarrow$ We seek a way to estimate the zeros in a sparse linear model.

Digression: Ridge and Lasso

Consider a regression setting where $\boldsymbol{Y} \in \mathbb{R}^n$ is a response vector and $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ a matrix of predictors.

If p is large, then the OLS solution $\hat{\beta} := (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}$ can have high variance.

In a linear model $\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon$ where $\mathrm{Cov}(\varepsilon) = \sigma^2 I$, we have

$$\frac{1}{n}\mathbb{E}\|\boldsymbol{Y} - \boldsymbol{X}\hat{\beta}\|_2^2 = \frac{\sigma^2 p}{n}.$$

Digression: Ridge and Lasso

Consider a regression setting where $\boldsymbol{Y} \in \mathbb{R}^n$ is a response vector and $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ a matrix of predictors.

If p is large, then the OLS solution $\hat{\beta} := (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}$ can have high variance.

In a linear model $\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon$ where $\mathrm{Cov}(\varepsilon) = \sigma^2 I$, we have

$$\frac{1}{n}\mathbb{E}\|\boldsymbol{Y} - \boldsymbol{X}\hat{\beta}\|_2^2 = \frac{\sigma^2 p}{n}.$$

Moreover when $p \gg n$, $\boldsymbol{X}$ will not have full column rank and OLS is not unique.

Digression: Ridge and Lasso

Consider a regression setting where $\boldsymbol{Y} \in \mathbb{R}^n$ is a response vector and $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ a matrix of predictors.

If p is large, then the OLS solution $\hat{\beta} := (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}$ can have high variance.

In a linear model $\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon$ where $\mathrm{Cov}(\varepsilon) = \sigma^2 I$, we have

$$\frac{1}{n}\mathbb{E}\|\boldsymbol{Y} - \boldsymbol{X}\hat{\beta}\|_2^2 = \frac{\sigma^2 p}{n}.$$

Moreover when $p \gg n$, $\boldsymbol{X}$ will not have full column rank and OLS is not unique.

Ridge regression: $\hat{\beta}^{\mathrm{R}}_\lambda := \underset{\beta \in \mathbb{R}^p}{\arg\min} \{\|\boldsymbol{Y} - \boldsymbol{X}\beta\|_2^2 + \lambda \|\beta\|_2^2\}$.

Constrained form: If $\|\hat{\beta}^{\mathrm{R}}_\lambda\|_2 = s$, then $\hat{\beta}^{\mathrm{R}}_\lambda$ also minimises

$$\|\boldsymbol{Y} - \boldsymbol{X}\beta\|_2^2 \qquad \text{subject to } \|\beta\|_2 \leq s.$$

Digression: Ridge and Lasso

Consider a regression setting where $\boldsymbol{Y} \in \mathbb{R}^n$ is a response vector and $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ a matrix of predictors.

If p is large, then the OLS solution $\hat{\beta} := (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}$ can have high variance.
In a linear model $\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon$ where $\mathrm{Cov}(\varepsilon) = \sigma^2 I$, we have

$$\frac{1}{n}\mathbb{E}\|\boldsymbol{Y} - \boldsymbol{X}\hat{\beta}\|_2^2 = \frac{\sigma^2 p}{n}.$$

Moreover when $p \gg n$, $\boldsymbol{X}$ will not have full column rank and OLS is not unique.

Ridge regression: $\hat{\beta}^{\mathrm{R}}_\lambda := \underset{\beta \in \mathbb{R}^p}{\arg\min} \{\|\boldsymbol{Y} - \boldsymbol{X}\beta\|_2^2 + \lambda \|\beta\|_2^2\}$.

Constrained form: If $\|\hat{\beta}^{\mathrm{R}}_\lambda\|_2 = s$, then $\hat{\beta}^{\mathrm{R}}_\lambda$ also minimises

$$\|\boldsymbol{Y} - \boldsymbol{X}\beta\|_2^2 \qquad \text{subject to } \|\beta\|_2 \leq s.$$

Lasso: $\hat{\beta}^{\mathrm{L}}_\lambda := \underset{\beta \in \mathbb{R}^p}{\arg\min} \{\frac{1}{2n}\|\boldsymbol{Y} - \boldsymbol{X}\beta\|_2^2 + \lambda \|\beta\|_1\}$.

Digression: Ridge and Lasso

Consider a regression setting where $\boldsymbol{Y} \in \mathbb{R}^n$ is a response vector and $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ a matrix of predictors.

If p is large, then the OLS solution $\hat{\beta} := (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}$ can have high variance.
In a linear model $\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon$ where $\mathrm{Cov}(\varepsilon) = \sigma^2 I$, we have

$$\frac{1}{n}\mathbb{E}\|\boldsymbol{X}\beta - \boldsymbol{X}\hat{\beta}\|_2^2 = \frac{\sigma^2 p}{n}.$$

Moreover when $p \gg n$, $\boldsymbol{X}$ will not have full column rank and OLS is not unique.

Ridge regression: $\hat{\beta}^{\mathrm{R}}_\lambda := \underset{\beta \in \mathbb{R}^p}{\arg\min} \{\|\boldsymbol{Y} - \boldsymbol{X}\beta\|_2^2 + \lambda \|\beta\|_2^2\}.$

Constrained form: If $\|\hat{\beta}^{\mathrm{R}}_\lambda\|_2 = s$, then $\hat{\beta}^{\mathrm{R}}_\lambda$ also minimises

$$\|\boldsymbol{Y} - \boldsymbol{X}\beta\|_2^2 \qquad \text{subject to } \|\beta\|_2 \leq s.$$

Square root Lasso: $\hat{\beta}^{\mathrm{sq}}_\lambda := \underset{\beta \in \mathbb{R}^p}{\arg\min} \{\frac{1}{\sqrt{n}}\|\boldsymbol{Y} - \boldsymbol{X}\beta\|_2 + \lambda \|\beta\|_1\}.$

Lasso coefficients are sparse

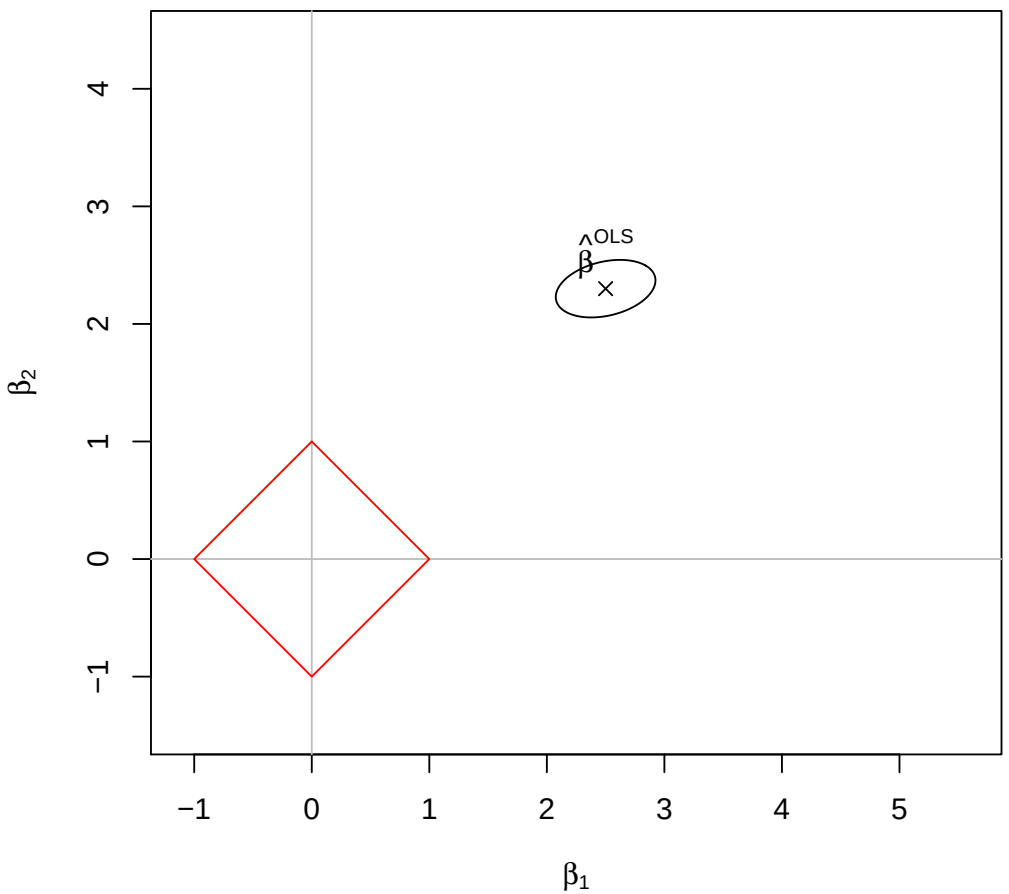


Figure: Contours of $\|\boldsymbol{Y} - \boldsymbol{X}\beta\|_2^2$ are ellipses centred at $\hat{\beta}$.

Lasso coefficients are sparse

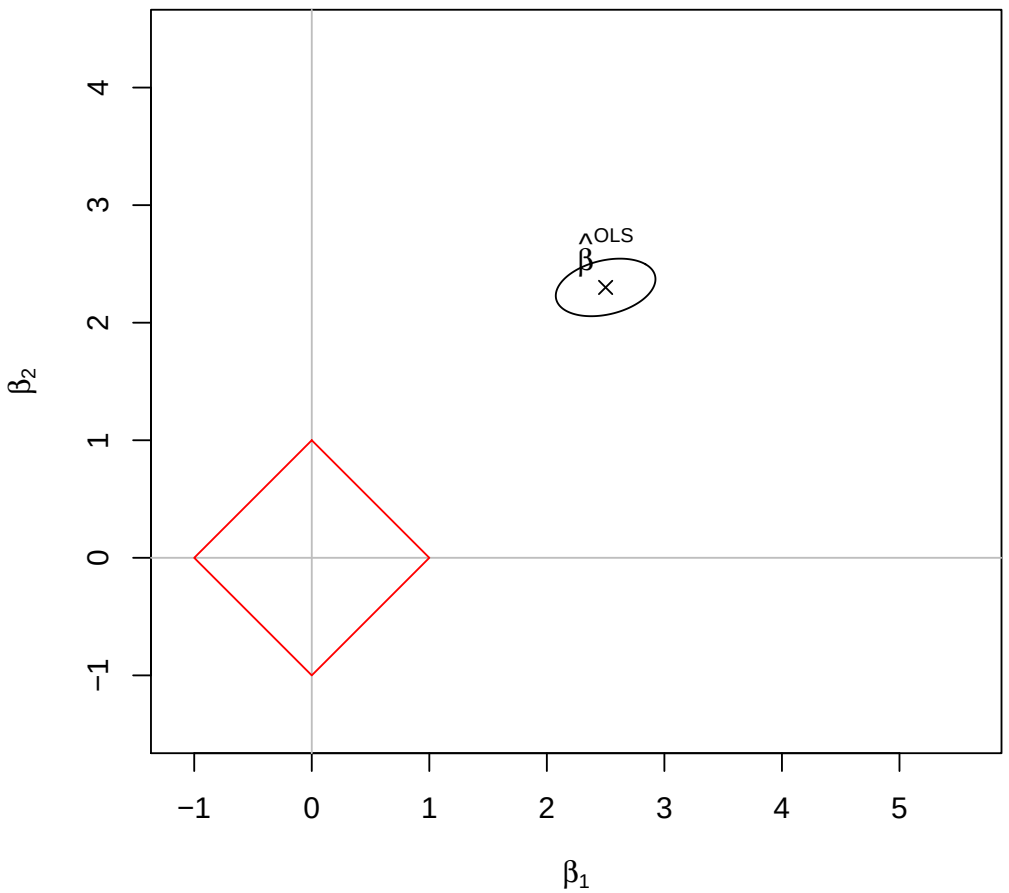
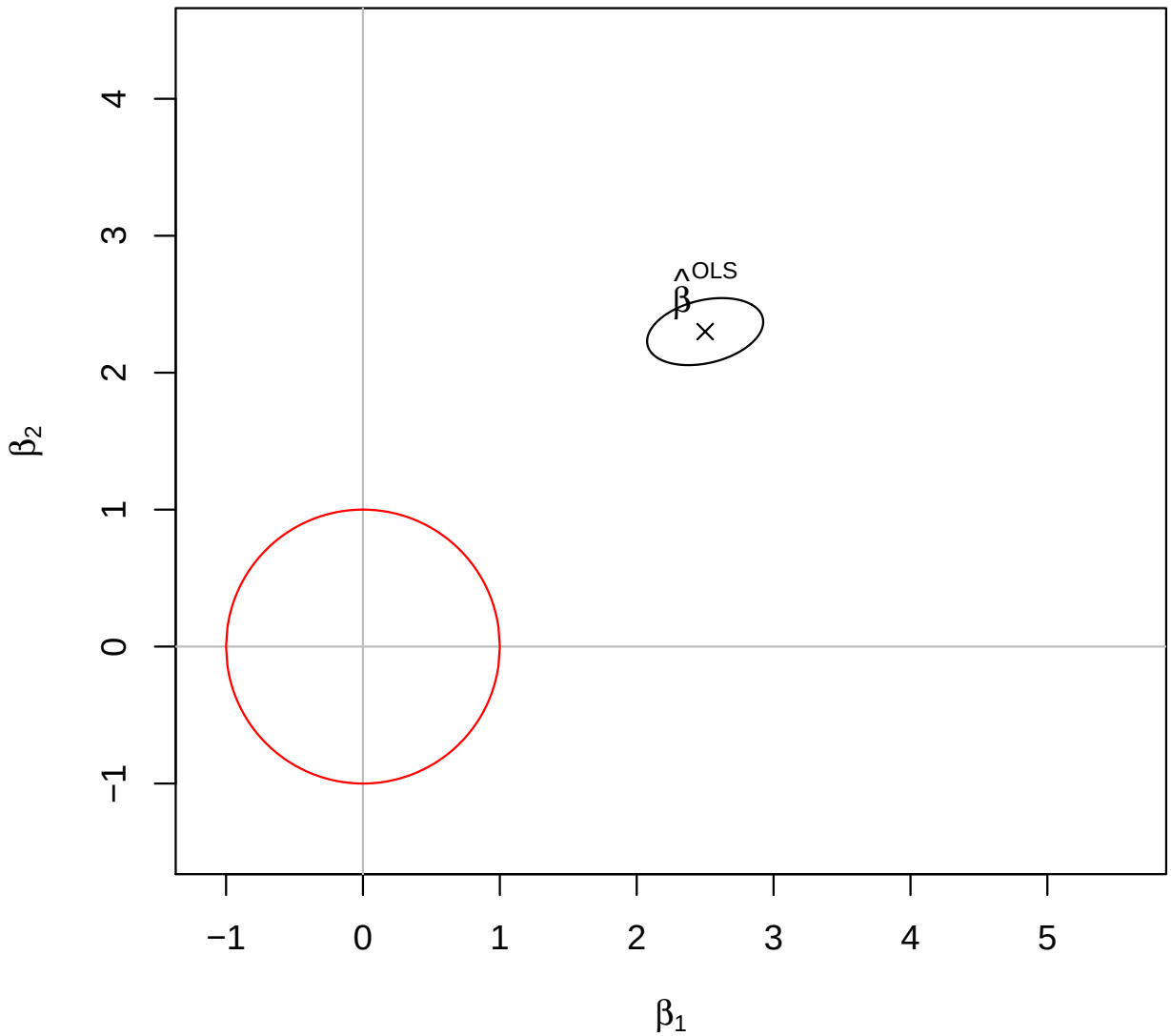


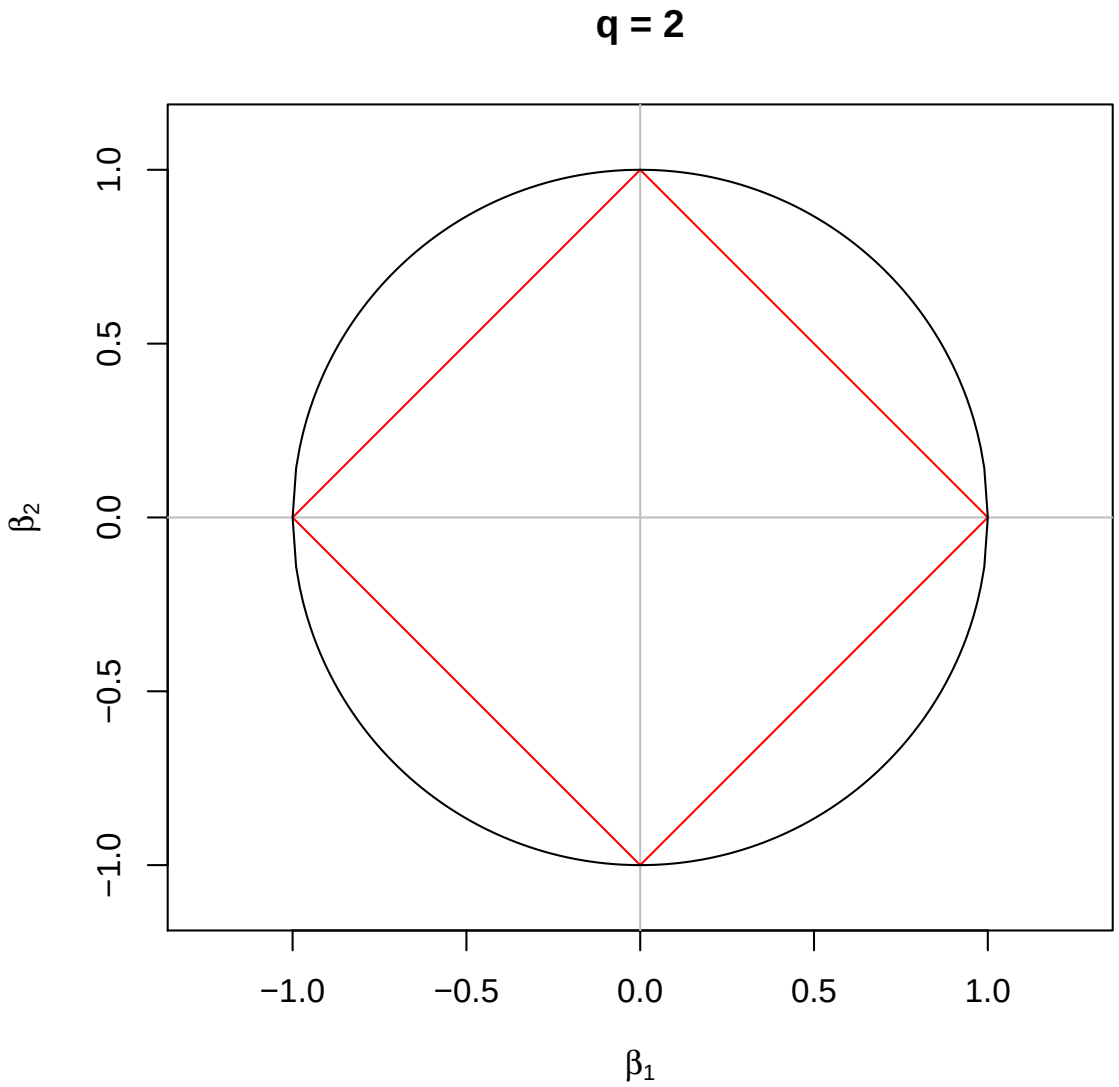
Figure: Contours of $\|\boldsymbol{Y} - \boldsymbol{X}\beta\|_2^2$ are ellipses centred at $\hat{\beta}$.

Ridge regression coefficients are (almost) always non-zero



ℓ_q balls

Consider penalty functions $\propto \|\beta\|_q = \left(\sum_{k=1}^{p} \beta_k^q\right)^{1/q}$ and $p = 2$.



Properties of the Lasso

Consider a (sparse) linear model

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{\varepsilon}$$

where only $s \ll p$ of the components of β are non-zero and $\boldsymbol{\varepsilon} \sim \mathcal{N}_p(0, \sigma^2 I)$.

- While Lasso coefficients are sparse, they require relatively strong conditions on $\boldsymbol{X}$ in order to select the correct coefficients with high probability.

Properties of the Lasso

Consider a (sparse) linear model

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon$$

where only $s \ll p$ of the components of β are non-zero and $\varepsilon \sim \mathcal{N}_p(0, \sigma^2 I)$.

- While Lasso coefficients are sparse, they require relatively strong conditions on $\boldsymbol{X}$ in order to select the correct coefficients with high probability.
- Recall that if we applied OLS only on the variables with non-zero coefficients (setting all others to zero), we would have

$$\frac{1}{n}\mathbb{E}\|\boldsymbol{X}(\beta - \hat{\beta})\|_2^2 = \frac{\sigma^2 s}{n}.$$

- Under relatively mild conditions on $\boldsymbol{X}$, the (square-root) Lasso achieves when $\lambda \asymp \sqrt{(\log p)/n}$:

$$\frac{1}{n}\|\boldsymbol{X}(\beta - \hat{\beta}^{\text{sq}}_{\lambda})\|_2^2 \leq \text{const.} \times \frac{\sigma^2 s \log p}{n}$$

$$\|\hat{\beta}^{\text{sq}}_{\lambda} - \beta\|_1 \leq \text{const.} \times \frac{\sigma s \sqrt{\log p}}{\sqrt{n}}.$$

Back to nodewise regressions

Given $Z_1, \ldots, Z_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}_p(\mu, \Sigma)$, set

$$\boldsymbol{Z} := \begin{pmatrix} Z_1^\top \\ \vdots \\ Z_n^\top \end{pmatrix} := \left(\boldsymbol{Z}_1 \,|\, \cdots \,|\, \boldsymbol{Z}_p \right).$$

Recall:

Zero coefficients from a linear regression of $\boldsymbol{Z}_j$ on $\boldsymbol{Z}_{-j}$

$$\Longleftrightarrow$$

conditional independencies.

Back to nodewise regressions

Given $Z_1, \ldots, Z_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}_p(\mu, \Sigma)$, set

$$\boldsymbol{Z} := \begin{pmatrix} Z_1^\top \\ \vdots \\ Z_n^\top \end{pmatrix} := \left(\boldsymbol{Z}_1 \,|\, \cdots \,|\, \boldsymbol{Z}_p\right).$$

Recall:

Zero coefficients from a linear regression of $\boldsymbol{Z}_j$ on $\boldsymbol{Z}_{-j}$

$$\Longleftrightarrow$$

conditional independencies.

- Perform Lasso regression of $\boldsymbol{Z}_j$ onto $\boldsymbol{Z}_{-j}$ and obtain estimated indices of non-zero coefficients $\hat{S}_j$.
- $$\left.\begin{array}{l} k \in \hat{S}_j \text{ AND } j \in \hat{S}_k \\ j \in \hat{S}_k \text{ OR } j \in \hat{S}_k \end{array}\right\} \rightarrow \text{ edge } j - k \text{ in CIG.}$$

Precision matrix and conditional independence

One way to resolve the ambiguity in using 'AND' or 'OR' rules involves using a characterisation of conditional independence using the precision matrix $\Omega := \Sigma^{-1}$.

Recall that $\mathbb{E}(Z_k \mid Z_{-k}) = \text{const.} + (\Sigma_{-k,-k}^{-1}\Sigma_{-k,k})^\top Z_{-k}$.

$$\Sigma = \begin{pmatrix} \Sigma_{k,k} & \Sigma_{k,-k} \\ \Sigma_{-k,k} & \Sigma_{-k,-k} \end{pmatrix} \qquad \Omega = \begin{pmatrix} \Omega_{k,k} & -\Omega_{k,k}\Sigma_{k,-k}\Sigma_{-k,-k}^{-1} \\ \underbrace{-\Omega_{k,k}\Sigma_{-k,-k}^{-1}\Sigma_{-k,k}}_{=\Omega_{-k,k}} & \Omega_{-k,-k} \end{pmatrix}$$

Precision matrix and conditional independence

One way to resolve the ambiguity in using 'AND' or 'OR' rules involves using a characterisation of conditional independence using the precision matrix $\Omega := \Sigma^{-1}$.

Recall that $\mathbb{E}(Z_k \mid Z_{-k}) = \text{const.} + (\Sigma_{-k,-k}^{-1}\Sigma_{-k,k})^\top Z_{-k}$.

$$\Sigma = \begin{pmatrix} \Sigma_{k,k} & \Sigma_{k,-k} \\ \Sigma_{-k,k} & \Sigma_{-k,-k} \end{pmatrix} \qquad \Omega = \begin{pmatrix} \Omega_{k,k} & -\Omega_{k,k}\Sigma_{k,-k}\Sigma_{-k,-k}^{-1} \\ \underbrace{-\Omega_{k,k}\Sigma_{-k,-k}^{-1}\Sigma_{-k,k}}_{=\Omega_{-k,k}} & \Omega_{-k,-k} \end{pmatrix}$$

$$\Omega_{j,k} = 0 \iff Z_j \perp\!\!\!\perp Z_k \mid Z_{-jk}.$$

$\rightarrow$ Seek to estimate zeroes in Ω.

Graphical Lasso

Let $\ell(\mu, \Omega)$ be the log-likelihood of $Z_1, \ldots, Z_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}_p(\mu, \Omega^{-1})$.

The *graphical Lasso* performs

$$(\hat{\mu}, \hat{\Omega}) = \underset{\mu \in \mathbb{R}^p, \Omega \succ 0}{\arg\min} \; -\ell(\mu, \Omega) + \lambda \sum_{j,k} |\Omega_{j,k}|.$$

Can show $\hat{\mu} = \bar{Z}$ and

$$\hat{\Omega} = \underset{\Omega : \Omega \succ 0}{\arg\min} \left\{ -\log\det\Omega + \mathrm{tr}(S\Omega) + \lambda \sum_{j,k} |\Omega_{j,k}| \right\}$$

where S is the empirical covariance matrix.

Often we omit the diagonal terms from the penalty as these are irrelevant for the CIG.

Partial correlation

Since $Z_k, Z_j \mid Z_{-jk}$ has a Gaussian distribution, we have

$$Z_j \perp\!\!\!\perp Z_k \mid Z_{-jk} \iff \mathrm{Corr}(Z_j, Z_k \mid Z_{-jk}) = 0.$$

Consider $(X, Y, Z) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^p$ jointly Gaussian.
Recall writing $\xi := X - \mathbb{E}(X \mid Z)$ and $\varepsilon := Y - \mathbb{E}(Y \mid Z)$, each of the conditional expectations is linear and

$$\mathrm{Corr}(X, Y \mid Z) = \frac{\mathbb{E}(\varepsilon\xi)}{\sqrt{\mathbb{E}(\varepsilon^2)\mathbb{E}(\xi^2)}} =: \rho.$$

Empirical version $\hat{\rho}$ replaces population residuals with residuals from linear regressions.

Testing using partial correlation

One can test $\rho = 0$ by comparing $\hat{\rho}$ to its distribution under the null.

Testing using partial correlation

One can test $\rho = 0$ by comparing $\hat{\rho}$ to its distribution under the null.

In fact we have

$$T_{\text{OLS}} = \sqrt{n - p - 1} \frac{\hat{\rho}}{\sqrt{1 - \hat{\rho}^2}},$$

where T_{OLS} is the t-statistic for testing the significance of X in a linear model of Y on (X, Z) (and an intercept term).

Testing using partial correlation

One can test $\rho = 0$ by comparing $\hat{\rho}$ to its distribution under the null.

In fact we have

$$T_{\text{OLS}} = \sqrt{n - p - 1} \frac{\hat{\rho}}{\sqrt{1 - \hat{\rho}^2}},$$

where T_{OLS} is the t-statistic for testing the significance of X in a linear model of Y on (X, Z) (and an intercept term).

An interesting consequence is that T_{OLS} is "double estimation-friendly" for testing $X \perp\!\!\!\perp Y \,|\, Z$ in the sense that $T_{\text{OLS}} \xrightarrow{d} \mathcal{N}(0, 1)$ provided either

- $Y = Z^\top \beta + \varepsilon$ with $\varepsilon \perp\!\!\!\perp (X, Z)$ or
- $X = Z^\top \theta + \xi$ with $\xi \perp\!\!\!\perp (Y, Z)$.

Regularised partial correlation

What about testing conditional independence in the high-dimensional setting where $p \gg n$?

Regularised partial correlation

What about testing conditional independence in the high-dimensional setting where $p \gg n$?

We can replace the each of the OLS regression involved in $\hat{\rho}$ with (square-root) Lasso regressions to obtain a *regularised partial correlation* $\tilde{\rho}$. [RSZZ15, SB23]

Regularised partial correlation

What about testing conditional independence in the high-dimensional setting where $p \gg n$?

We can replace the each of the OLS regression involved in $\hat{\rho}$ with (square-root) Lasso regressions to obtain a *regularised partial correlation* $\tilde{\rho}$. [RSZZ15, SB23]

Suppose $X \perp\!\!\!\perp Y \,|\, Z$. Have $Y = \beta^\top Z + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ for some $\sigma^2 > 0$.

Let s_β be the number of non-zeroes in $\beta \in \mathbb{R}^p$.

Provided $s_\beta \log(p)/\sqrt{n} \ll 1$, we will have

$$\sqrt{n}\tilde{\rho} \approx \mathcal{N}(0, 1).$$

Regularised partial correlation

What about testing conditional independence in the high-dimensional setting where $p \gg n$?

We can replace the each of the OLS regression involved in $\hat{\rho}$ with (square-root) Lasso regressions to obtain a *regularised partial correlation* $\tilde{\rho}$. [RSZZ15, SB23]

Suppose $X \perp\!\!\!\perp Y \,|\, Z$. Have $Y = \beta^\top Z + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ for some $\sigma^2 > 0$.

Let s_β be the number of non-zeroes in $\beta \in \mathbb{R}^p$.

Provided $s_\beta \log(p)/\sqrt{n} \ll 1$, we will have

$$\sqrt{n}\tilde{\rho} \approx \mathcal{N}(0, 1).$$

Symmetry of $\tilde{\rho}$ entails that we will have the same result if instead we have a sparse linear model of X on Z.

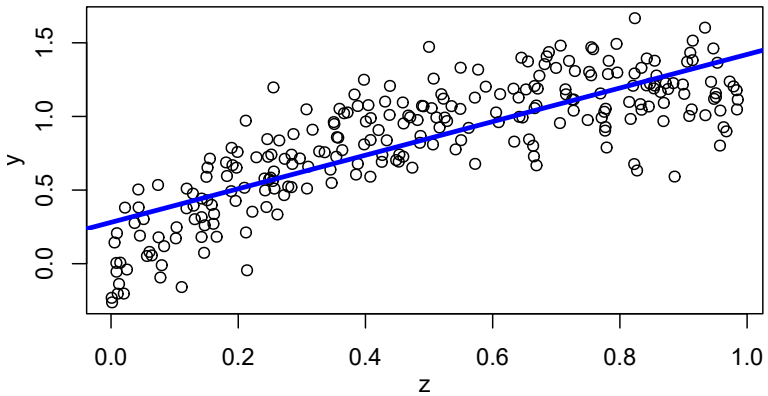
Beyond Gaussianity

The assumption that (X, Y, Z) are jointly Gaussian is convenient, but may be hard to defend for larger sample sizes.

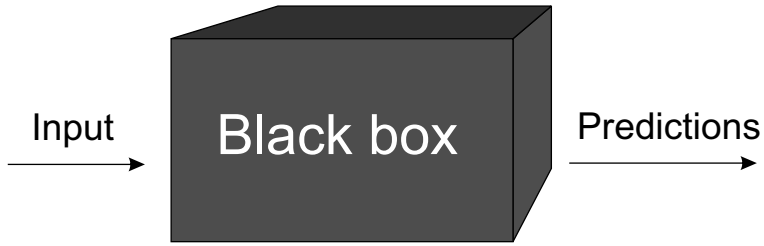
Nevertheless, we have seen that some assumption is required as otherwise $X \perp\!\!\!\perp Y \,|\, Z$ is untestable.

In fact regression is famously 'impossible': there are no regression procedures with risk converging uniformly to zero across all data generating processes [GKKW06].

However, we have many successful methods: Neural networks, random forests, boosted trees, kernel ridge regression,...



One way of restricting the null is via models...



...or we can phrase the subset of the null where we want to control size as one where user-chosen regression methods perform sufficiently well.

Generalised Covariance Measure

Recall we always have $X \perp\!\!\!\perp Y \mid Z \implies \mathbb{E}\mathrm{Cov}(X, Y \mid Z) = 0$, where

$$\mathbb{E}\mathrm{Cov}(X, Y \mid Z) = \mathbb{E}[\{X - \underbrace{\mathbb{E}(X \mid Z)}_{=:f(Z)}\}\{Y - \underbrace{\mathbb{E}(Y \mid Z)}_{=:g(Z)}\}].$$

Let (X_i, Y_i, Z_i) be i.i.d. copies of (X, Y, Z).
Empirical version:

$$\frac{1}{n}\sum_{i=1}^{n} \underbrace{\{X_i - \hat{f}(Z_i)\}\{Y_i - \hat{g}(Z_i)\}}_{=:L_i}.$$

We expect $\mathbb{E}(L_i) \approx 0$ under the null, which suggests the *generalised covariance measure (GCM)* [SP20]:

$$T_{\mathrm{GCM}} := \sqrt{n}\frac{\frac{1}{n}\sum_{i=1}^{n} L_i}{\frac{1}{n}\sum_{i=1}^{n}(L_i - \bar{L})^2}.$$

‘Single machine learning’

In fact, we also have $\mathbb{E}\mathrm{Cov}(X, Y, \mid Z) = \mathbb{E}[X\{Y - \mathbb{E}(Y \mid Z)\}]$.

This suggests a simpler approach involving

$$T_{\mathsf{simp}} := \underbrace{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i\{Y_i - g(Z_i)\}}_{\xrightarrow{d} \mathcal{N}(0,v)} + \underbrace{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i\{g(Z_i) - \hat{g}(Z_i)\}}_{\Delta \text{ (small?)}} .$$

‘Single machine learning’

In fact, we also have $\mathbb{E}\mathrm{Cov}(X, Y, \mid Z) = \mathbb{E}[X\{Y - \mathbb{E}(Y \mid Z)\}]$.

This suggests a simpler approach involving

$$T_{\mathsf{simp}} := \underbrace{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i\{Y_i - g(Z_i)\}}_{\xrightarrow{d} \mathcal{N}(0,v)} + \underbrace{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i\{g(Z_i) - \hat{g}(Z_i)\}}_{\Delta \text{ (small?)}}.$$

$$\Delta^2 \overset{\text{C–S}}{\leq} \left(\frac{1}{n} \sum_{i=1}^{n} X_i^2 \right) \left(\sum_{i=1}^{n} \{g(Z_i) - \hat{g}(Z_i)\}^2 \right).$$

However, typically Δ does not decay to zero.

Consequently T_{simp} will not be asymptotically mean-zero under the null.

Size control

Theorem

Let $E_f := \frac{1}{n}\sum_{i=1}^{n}\{f(Z_i) - \hat{f}(Z_i)\}^2$, $E_g := \frac{1}{n}\sum_{i=1}^{n}\{g(Z_i) - \hat{g}(Z_i)\}^2$, *and suppose*

1. $n E_f E_g \xrightarrow{p} 0$ *and* $E_f \xrightarrow{p} 0$, $E_g \xrightarrow{p} 0$;
2. $0 < \mathbb{E}(\varepsilon^2\xi^2) < \infty$ *and there exists* $\sigma^2 > 0$ *such that* $\mathrm{Var}(X|Z), \mathrm{Var}(Y|Z) < \sigma^2$.

Then

$$\sup_{t\in\mathbb{R}} |\mathbb{P}(T_{GCM} \leq t) - \Phi(t)| \to 0.$$

Using two regressions (double machine learning [CCD+18]) gives product of biases.

Size control

Theorem

Let $E_f := \frac{1}{n}\sum_{i=1}^{n}\{f(Z_i) - \hat{f}(Z_i)\}^2$, $E_g := \frac{1}{n}\sum_{i=1}^{n}\{g(Z_i) - \hat{g}(Z_i)\}^2$, *and suppose*

1. $n E_f E_g \xrightarrow{p} 0$ *and* $E_f \xrightarrow{p} 0$, $E_g \xrightarrow{p} 0$;
2. $0 < \mathbb{E}(\varepsilon^2\xi^2) < \infty$ *and there exists* $\sigma^2 > 0$ *such that* $\mathrm{Var}(X|Z), \mathrm{Var}(Y|Z) < \sigma^2$.

Then

$$\sup_{t\in\mathbb{R}} |\mathbb{P}(T_{GCM} \leq t) - \Phi(t)| \to 0.$$

Using two regressions (double machine learning [CCD+18]) gives product of biases.

Key condition $n E_f E_g \xrightarrow{p} 0$ satisfied in many nonparametric / high-dimensional settings. E.g.

- high-dimensional linear models with $s_X s_Y \log(p)^2/n \to 0$ ($Z \in \mathbb{R}^p$)

Size control

Theorem

Let $E_f := \frac{1}{n}\sum_{i=1}^{n}\{f(Z_i) - \hat{f}(Z_i)\}^2$, $E_g := \frac{1}{n}\sum_{i=1}^{n}\{g(Z_i) - \hat{g}(Z_i)\}^2$, *and suppose*

1. $n E_f E_g \xrightarrow{p} 0$ *and* $E_f \xrightarrow{p} 0$, $E_g \xrightarrow{p} 0$;
2. $0 < \mathbb{E}(\varepsilon^2\xi^2) < \infty$ *and there exists* $\sigma^2 > 0$ *such that* $\mathrm{Var}(X|Z), \mathrm{Var}(Y|Z) < \sigma^2$.

Then

$$\sup_{t\in\mathbb{R}} |\mathbb{P}(T_{GCM} \leq t) - \Phi(t)| \to 0.$$

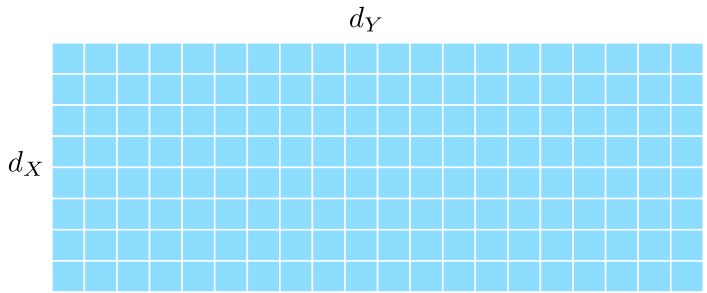
Using two regressions (double machine learning [CCD+18]) gives product of biases.

Key condition $n E_f E_g \xrightarrow{p} 0$ satisfied in many nonparametric / high-dimensional settings. E.g.

- high-dimensional linear models with $s_X s_Y \log(p)^2/n \to 0$ ($Z \in \mathbb{R}^p$) (stronger than $\min(s_Y^2, s_X^2)\log(p)^2/n \to 0$.)
- f, g lie in an RKHS and have bounded RKHS norm and reproducing kernel k admits a Mercer decomposition.

Multivariate case

- Consider now $X \in \mathbb{R}^{d_X}$, $Y \in \mathbb{R}^{d_Y}$.
- Let $\boldsymbol{L}_{jk} \in \mathbb{R}^n$ be the vector of products of residuals from regressing $\boldsymbol{X}_j$ and $\boldsymbol{Y}_k$ on to $\boldsymbol{Z}$ respectively.
- Can form GCM test statistic T_{jk} for each $j = 1, \dots, d_X$, $k = 1, \dots, d_Y$ based on $\boldsymbol{L}_{jk}$.



Power

Can show the GCM has power against alternatives where $|\mathbb{E}\mathrm{Cov}(X, Y \mid Z)| \geq \text{const.} \times n^{-1/2}$.

Power

Can show the GCM has power against alternatives where $|\mathbb{E}\mathrm{Cov}(X, Y \mid Z)| \geq \text{const.} \times n^{-1/2}$.

But there are alternatives that it has no power against:

$$(X, Y, \varepsilon) \sim \mathcal{N}_3(0, I) \qquad Y = X^2 + \varepsilon.$$

Then $\mathrm{Cov}(X, Y \mid Z) = 0$ while $X \not\perp\!\!\!\perp Y \mid Z$.

Power

Can show the GCM has power against alternatives where $|\mathbb{E}\mathrm{Cov}(X, Y \mid Z)| \geq \text{const.} \times n^{-1/2}$.

But there are alternatives that it has no power against:

$$(X, Y, \varepsilon) \sim \mathcal{N}_3(0, I) \qquad Y = X^2 + \varepsilon.$$

Then $\mathrm{Cov}(X, Y \mid Z) = 0$ while $X \not\perp\!\!\!\perp Y \mid Z$. One can apply the GCM to X^2, Y, Z to get power.

In general can pre-transform by replacing X by any function of (X, Z) and similarly for Y.

Power

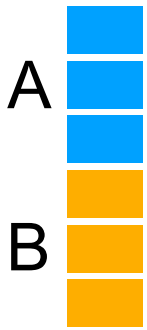
Can show the GCM has power against alternatives where $|\mathbb{E}\mathrm{Cov}(X, Y \mid Z)| \geq \text{const.} \times n^{-1/2}$.

But there are alternatives that it has no power against:

$$(X, Y, \varepsilon) \sim \mathcal{N}_3(0, I) \qquad Y = X^2 + \varepsilon.$$

Then $\mathrm{Cov}(X, Y \mid Z) = 0$ while $X \not\perp\!\!\!\perp Y \mid Z$. One can apply the GCM to X^2, Y, Z to get power.

In general can pre-transform by replacing X by any function of (X, Z) and similarly for Y.



Hunt-and-test:

1. **'Hunt'**: Use Part A to determine which transformation of (X, Z) to use
2. **Test**: Apply the GCM test to Part B using the transformed data.

Looking beyond

More broadly, conditional independence testing and related topics remain highly active research areas.

Looking beyond

More broadly, conditional independence testing and related topics remain highly active research areas. Things I did not talk about:

- Hidden variables; instrumental variables...
- Relationship to invariance and stability...
- Causal effect estimation....
- Semiparametric statistics...

Thank you for listening.

Selected references

[CCD+18] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins, *Double/debiased machine learning for treatment and structural parameters*, 2018.

[GKKW06] László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk, *A distribution-free theory of nonparametric regression*, Springer Science & Business Media, 2006.

[RSZZ15] Zhao Ren, Tingni Sun, Cun-Hui Zhang, and Harrison H Zhou, *Asymptotic normality and optimalities in estimation of large gaussian graphical models*, The Annals of Statistics **43** (2015), no. 3.

[SB23] Rajen D Shah and Peter Bühlmann, *Double-estimation-friendly inference for high-dimensional misspecified models*, Statistical Science **38** (2023), no. 1, 68–91.

[SP20] Rajen D Shah and Jonas Peters, *The hardness of conditional independence testing and the generalised covariance measure*, The Annals of Statistics (2020).