Coarsening at random

Anders Munch

June 2, 2023

Ideal experiment and observed data

Consider an "ideal" statistical problem $(\mathcal{Q}, \theta)$,

θ scientifically meaningful parameter

$Q \in \mathcal{Q}$ distribution from which we wished we had data

Ideal experiment and observed data

Consider an "ideal" statistical problem $(\mathcal{Q}, \theta)$,

$$\begin{aligned} \theta &\quad \text{scientifically meaningful parameter} \\ Q \in \mathcal{Q} &\quad \text{distribution from which we wished we had data} \end{aligned}$$

Example (ideal data)

- $(X, Y) \sim Q$ and $\theta(Q) = \mathbb{E}_Q[Y]$
- $T \sim Q$ and $\theta(Q) = \mathbb{E}_Q[\mathbb{1}\{T > t\}]$
- $(W, Y(0), Y(1)) \sim Q$ and $\theta(Q) = \mathbb{E}_Q[Y(0) - Y(1)]$

Ideal experiment and observed data

Consider an "ideal" statistical problem $(\mathcal{Q}, \theta)$,

θ scientifically meaningful parameter

$Q \in \mathcal{Q}$ distribution from which we wished we had data

Example (ideal data)

- $(X, Y) \sim Q$ and $\theta(Q) = \mathbb{E}_Q[Y]$
- $T \sim Q$ and $\theta(Q) = \mathbb{E}_Q[\mathbb{1}\{T > t\}]$
- $(W, Y(0), Y(1)) \sim Q$ and $\theta(Q) = \mathbb{E}_Q[Y(0) - Y(1)]$

Example (observed data)

Unfortunately, we only have data available from a "corrupted sample":

- (X, R, RY) where R is a binary indicator of missing data
- $(\tilde{T}, \Delta)$ where $\tilde{T} = T \wedge C$ and $\mathbb{1}\{T \leq C\}$ for a censoring time C
- (W, A, Y) where $Y = AY(1) + (1 - A)Y(0)$

Coarsened data

Data with loss of information can in many cases be describe as a *coarsened* version of ideal or full data. We imagine that the full data is drawn from some unknown $Q \in \mathcal{Q}$ and then some (unknown) coarsening mechanism $G \in \mathcal{G}$ determines what we get to see.

Coarsened data

Data with loss of information can in many cases be describe as a *coarsened* version of ideal or full data. We imagine that the full data is drawn from some unknown $Q \in \mathcal{Q}$ and then some (unknown) coarsening mechanism $G \in \mathcal{G}$ determines what we get to see.

$$Z \sim Q \text{ and } \mathcal{C} \sim G \longmapsto O \sim P_{Q,G}$$

Example (coarsening)

- Draw $(X, Y) \sim Q$ and $R \sim G \longmapsto (X, R, RY) \sim P_{Q,G}$
- Draw $T \sim Q$ and $C \sim G \longmapsto (\tilde{T}, \Delta) \sim P_{Q,G}$
- Draw $(W, Y(0), Y(1)) \sim Q$ and $A \sim G \longmapsto (W, A, Y) \sim P_{Q,G}$

Coarsened data

Data with loss of information can in many cases be describe as a *coarsened* version of ideal or full data. We imagine that the full data is drawn from some unknown $Q \in \mathcal{Q}$ and then some (unknown) coarsening mechanism $G \in \mathcal{G}$ determines what we get to see.

$$Z \sim Q \text{ and } \mathcal{C} \sim G \longmapsto O \sim P_{Q,G}$$

Example (coarsening)

- Draw $(X, Y) \sim Q$ and $R \sim G \longmapsto (X, R, RY) \sim P_{Q,G}$
- Draw $T \sim Q$ and $C \sim G \longmapsto (\tilde{T}, \Delta) \sim P_{Q,G}$
- Draw $(W, Y(0), Y(1)) \sim Q$ and $A \sim G \longmapsto (W, A, Y) \sim P_{Q,G}$

The term *coarsening* refers to that we only get to see a "coarse-grained" version of the data which is less informative than the original "fine-grained" data.

Target parameter – in the target population!

Target parameter – in the target population!

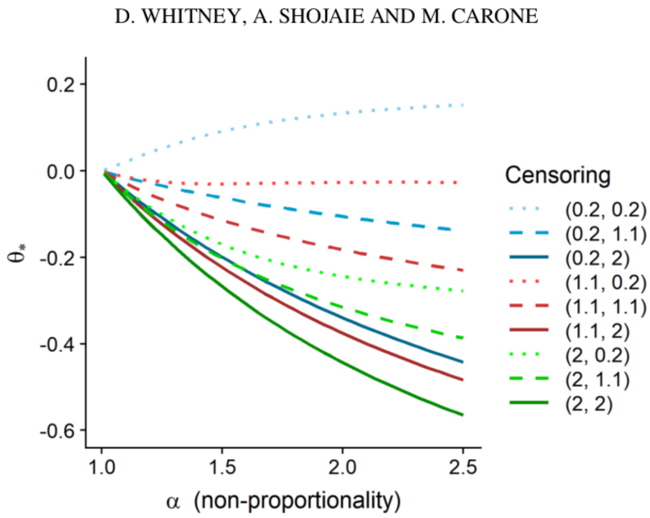


Figure from Whitney et al. [2019].

Identifiability – coarsening at random

To do estimation and inference we need to transform the problem $(\mathcal{Q}, \theta)$ into a problem concerning the observed data $(\mathcal{P}, \Psi)$, where $\{P_{Q,G} : Q \in \mathcal{Q}, G \in \mathcal{G}\}$.

First step is to *identify* our target parameter θ, i.e., write

$$\Psi(P_{Q,G}) = \theta(Q) \quad \text{for all } Q \text{ and } G.$$

Identifiability – coarsening at random

To do estimation and inference we need to transform the problem $(\mathcal{Q}, \theta)$ into a problem concerning the observed data $(\mathcal{P}, \Psi)$, where $\{P_{Q,G} : Q \in \mathcal{Q}, G \in \mathcal{G}\}$.

First step is to *identify* our target parameter θ, i.e., write

$$\Psi(P_{Q,G}) = \theta(Q) \quad \text{for all } Q \text{ and } G.$$

No assumptions about $\mathcal{G} \implies$ game over

For example, if $O = (R, RY)$ and $P(R = 0 \mid Y > 5) > P(R = 0 \mid Y \leq 5)$ we have a biased sample and we cannot learn the coarsening mechanism from the observed data.

Identifiability – coarsening at random

To do estimation and inference we need to transform the problem $(\mathcal{Q}, \theta)$ into a problem concerning the observed data $(\mathcal{P}, \Psi)$, where $\{P_{Q,G} : Q \in \mathcal{Q}, G \in \mathcal{G}\}$.

First step is to *identify* our target parameter θ, i.e., write

$$\Psi(P_{Q,G}) = \theta(Q) \quad \text{for all } Q \text{ and } G.$$

No assumptions about $\mathcal{G} \implies$ game over

For example, if $O = (R, RY)$ and $P(R = 0 \mid Y > 5) > P(R = 0 \mid Y \leq 5)$ we have a biased sample and we cannot learn the coarsening mechanism from the observed data.

Coarsening at random (CAR) $\implies$ game on

CAR states that the coarsening mechanism only depends on the observed data. [Heitjan and Rubin, 1991, Gill et al., 1997]

Identifiability – coarsening at random

To do estimation and inference we need to transform the problem $(\mathcal{Q}, \theta)$ into a problem concerning the observed data $(\mathcal{P}, \Psi)$, where $\{P_{Q,G} : Q \in \mathcal{Q}, G \in \mathcal{G}\}$.

First step is to *identify* our target parameter θ, i.e., write

$$\Psi(P_{Q,G}) = \theta(Q) \quad \text{for all } Q \text{ and } G.$$

No assumptions about $\mathcal{G} \implies$ game over

For example, if $O = (R, RY)$ and $P(R = 0 \mid Y > 5) > P(R = 0 \mid Y \leq 5)$ we have a biased sample and we cannot learn the coarsening mechanism from the observed data.

Coarsening at random (CAR) $\implies$ game on

CAR states that the coarsening mechanism only depends on the observed data. [Heitjan and Rubin, 1991, Gill et al., 1997]

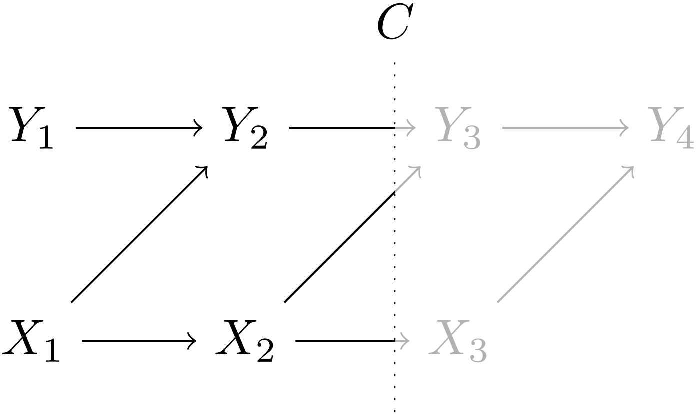
For example this holds if $R \perp\!\!\!\perp Y \mid X$.

CAR for longitudinal data

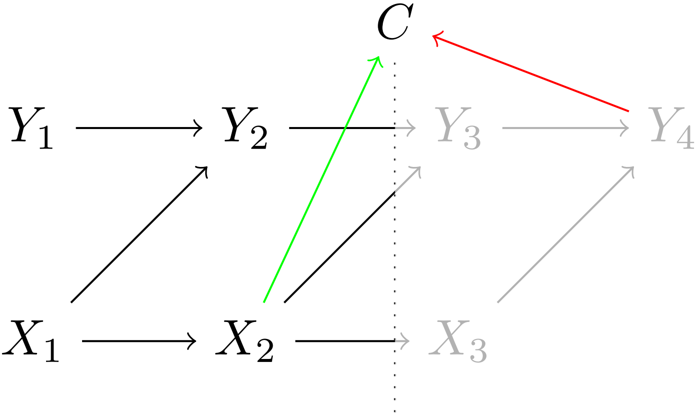
CAR for longitudinal data

$$
\begin{tikzcd}
Y_1 \arrow[r] & Y_2 \arrow[r] & Y_3 \arrow[r] & Y_4 \\
X_1 \arrow[r] \arrow[ur] & X_2 \arrow[r] \arrow[ur] & X_3 \arrow[ur] &
\end{tikzcd}
$$

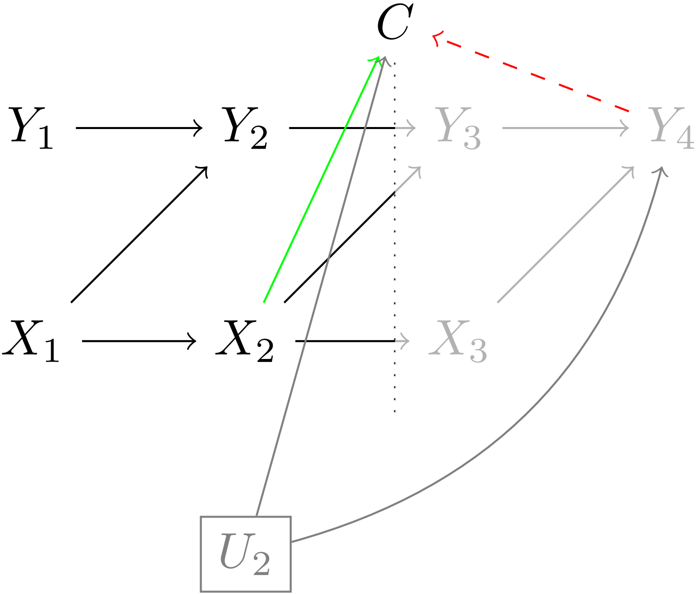
CAR for longitudinal data



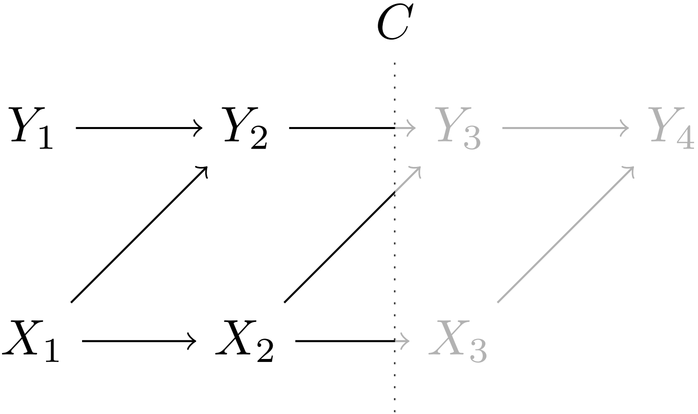
CAR for longitudinal data



CAR for longitudinal data



CAR for longitudinal data



CAR and counterfactual/potential outcomes

$$\text{Full data } (W, Y(0), Y(1)) \sim Q$$
$$\text{Observed data } (W, A, Y) \sim P_{Q,G} \text{ with } A \sim G$$

CAR and counterfactual/potential outcomes

Full data $(W, Y(0), Y(1)) \sim Q$
Observed data $(W, A, Y) \sim P_{Q,G}$ with $A \sim G$

The assumption of no unmeasured confounding states that

$$A \perp\!\!\!\perp \{Y(0), Y(1)\} \mid W. \qquad (*)$$

CAR and counterfactual/potential outcomes

$$\begin{aligned} \text{Full data} &\ \ (W, Y(0), Y(1)) \sim Q \\ \text{Observed data} &\ \ (W, A, Y) \sim P_{Q,G} \text{ with } A \sim G \end{aligned}$$

The assumption of no unmeasured confounding states that

$$A \perp\!\!\!\perp \{Y(0), Y(1)\} \mid W. \tag{$*$}$$

$$\begin{aligned} W &\ \ \text{is observed} \\ Y(0),\ Y(1) &\ \ \text{are partly unobserved} \\ \Longrightarrow &\ \ \text{CAR holds when we assume } (*) \end{aligned}$$

Efficiency theory under CAR

Nonparametric models stay nonparametric under CAR

CAR is the weakest assumption we can impose to ensure identifiability.

If $\mathcal{Q}$ is nonparametric and we assume nothing about $\mathcal{G}$ except car, then the induced model

$$\mathcal{P} = \{P_{Q,G} : Q \in \mathcal{Q}, G \in \mathcal{G}\}$$

will also be nonparametric.

Efficiency theory under CAR

Nonparametric models stay nonparametric under CAR

CAR is the weakest assumption we can impose to ensure identifiability.

If $\mathcal{Q}$ is nonparametric and we assume nothing about $\mathcal{G}$ except car, then the induced model

$$\mathcal{P} = \{P_{Q,G} : Q \in \mathcal{Q}, G \in \mathcal{G}\}$$

will also be nonparametric.

Information bounds under CAR

If we known the tangent space and the canonical gradient for the "ideal" statistical problem $(\mathcal{Q}, \theta)$, we can in many cases use projections and other Hilbert space techniques to find the tangent space and the canonical gradient for the observed statistical problem $(\mathcal{P}, \Psi)$.

A general methodology for doing this is presented in van der Laan et al. [2003] and Tsiatis [2007].

References

R. D. Gill, M. J. Laan, and J. M. Robins. Coarsening at random: Characterizations, conjectures, counter-examples. In *Proceedings of the First Seattle Symposium in Biostatistics*, pages 255–294. Springer, 1997.

D. F. Heitjan and D. B. Rubin. Ignorability and coarse data. *The annals of statistics*, pages 2244–2253, 1991.

A. Tsiatis. *Semiparametric theory and missing data*. Springer Science & Business Media, 2007.

M. J. van der Laan, M. Laan, and J. M. Robins. *Unified methods for censored longitudinal data and causality*. Springer Science & Business Media, 2003.

D. Whitney, A. Shojaie, and M. Carone. Comment: Models as (deliberate) approximations. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 34(4):591, 2019.