Advanced techniques in applied economics

Lecture 6: Unobserved confounding and adjustments

Piotr Zwiernik

upf. Universitat Pompeu Fabra Barcelona

Statistics, Probability and Machine Learning Research Group

BSE Barcelona School of Economics

Spring 2026

When the clean DAG picture breaks

What changes today?

In the previous lectures on DAGs we mostly acted as if the important variables were observed. Today we ask what happens when a relevant common cause is hidden.

Main difficulty

With unobserved confounders

- back-door adjustment may fail,
- graphs on observed variables may be misleading,
- standard regression can be badly biased.

Main response. We often can partially adjust for unobserved confounders.

- instrumental variables,
- proxy controls,
- latent factor adjustments,
- structured economic assumptions.

Main question: How can we reason causally when the graph contains hidden variables?

Three recurring economic examples

Education and earnings

D = schooling, Y = wages,
hidden confounder = ability or family background.

Training programs

D = program participation, Y = later earnings,
hidden confounder = motivation or employability.

Firm outcomes

D = policy or treatment exposure, Y = productivity or profits,
hidden confounder = latent firm quality or local demand.

The same statistical problem appears in many applied settings: treatment choice and outcomes are both affected by something we do not observe.

Roadmap for today

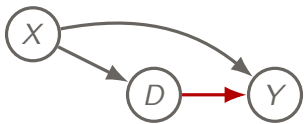
1. hidden confounding and why the clean DAG story breaks
2. a brief glimpse of MAGs
3. IV as the benchmark workaround
4. proxy controls
5. latent factor adjustments

The goal is no longer perfect recovery of the causal graph. The goal is careful reasoning under partial observability.

Part 1: Hidden confounding

Recall: the clean adjustment story

Suppose we want the effect of treatment D on outcome Y.



Adjustment logic If X blocks all back-door paths from D to Y, then adjusting for X may identify the causal effect.

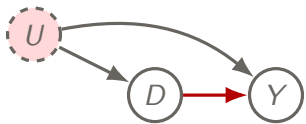
Example

D could be years of education, Y later earnings, and X observed prior test scores.

For the back-door criterion, see Pearl, *Causality*, Chapter 3.

What if the confounder is hidden?

Now suppose the true common cause U is unobserved.



Main problem

The back-door path

$$D \leftarrow U \rightarrow Y$$

is open, but we cannot condition on U because it is hidden.

Example

Suppose more motivated individuals are more likely to enroll in a training program. They also tend to earn more later even without the program. Then the observed association between training and wages mixes the causal effect of training with selection on hidden motivation.

A regression view of omitted-variable bias

Suppose

$$Y = \alpha D + h(W) + \lambda U + \varepsilon, \qquad D = s(W) + \rho U + \nu,$$

where U is unobserved and $\mathbb{E}(\varepsilon | D, W, U) = 0$, $\mathbb{E}(\nu | W, U) = 0$.
After partialling out W, the residualized equations are

$$\widetilde{Y} = \alpha \widetilde{D} + \lambda \widetilde{U} + \varepsilon, \qquad \widetilde{D} = \rho \widetilde{U} + \nu.$$

What goes wrong?

The remaining variation in D is correlated with the omitted term:

$$\mathrm{Cov}(\widetilde{D}, \lambda \widetilde{U} + \varepsilon) = \lambda \rho \, \mathrm{Var}(\widetilde{U}) \neq 0.$$

So we cannot recover α by regressing $\tilde{Y}$ on $\tilde{D}$.

For omitted-variable bias in regression language, see Wooldridge, *Econometric Analysis of Cross Section and Panel Data*.

A tiny omitted-variable simulation

Suppose

$$D = U + \varepsilon_D, \qquad Y = \mathbf{1.0}\, D + U + \varepsilon_Y.$$

```
> set.seed(1)
> n <- 3000
> U <- rnorm(n)
> D <- U + rnorm(n)
> Y <- 1.0 * D + U + rnorm(n)

> coef(lm(Y ~ D))
(Intercept) D
-0.01500857 1.50003799
```

```
# impossible benchmark: if U were observed

> coef(lm(Y ~ D + U))
(Intercept) D U
-0.01555177 0.99209827 0.98495893
```

Interpretation

The naive regression on D alone mixes the true causal effect with hidden confounding from U.

Part 2: A brief glimpse of MAGs and PAGs

Why do we need something beyond DAGs?

With hidden variables, a DAG on the observed variables alone is no longer the right target.

What changes when some variables are hidden?

- some directed effects remain visible,
- some dependence comes from hidden common causes,
- and the observed CI structure is no longer well represented by an ordinary DAG.

So after marginalizing over hidden variables, the right graphical language is often a **mixed graph** rather than a DAG.

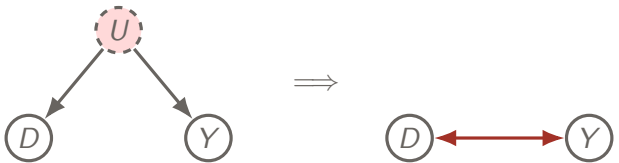
For MAGs and latent-variable graphical structure, see Richardson and Spirtes (2002), *Annals of Statistics*. For discovery with latent variables, see Zhang (2008) and Colombo et al. (2012), both in *Annals of Statistics*.

First intuition: bidirected edges for hidden confounding

A **mixed graph** is simply a graph that may contain different edge types, for example

$$A \to B, \qquad A \leftrightarrow B.$$

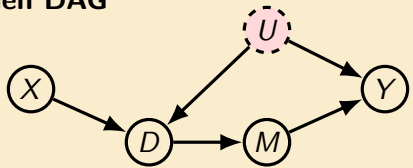
Useful first intuition: If two observed variables share a hidden common cause, we often represent this by a bidirected edge.



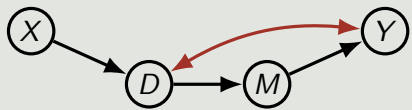
So a bidirected edge is a convenient way to say: "these two observed variables may share hidden confounding."

A tiny hidden-DAG example

Hidden DAG



Corresponding mixed graph



Observed CI

In this example, the only observed CI statement is $X \perp\!\!\!\perp M \mid D$.

The main point is not to master the formal definition of MAGs today. The main point is that mixed graphs give a convenient language for CI structure when hidden variables are present.

(Really) Fast Causal Inference (R)FCI

In practice, latent-variable discovery usually returns not a single MAG, but an equivalence class represented by a **PAG**. .

Typical workflow

1. start from observed data only,
2. test conditional independences,
3. run FCI or RFCI,
4. output a PAG.

Why PAGs look more complicated

A PAG must summarize what is identifiable when

- some variables may be hidden,
- some directions are not identifiable,
- several latent-variable structures may fit the same observed CI pattern.

A PAG is the latent-variable analogue of a CPDAG: it tells us what the CI information can and cannot identify.

see Colombo et al., *Learning high-dimensional directed acyclic graphs with latent and selection variables*, Ann Stat 2012.

How to read a PAG

Basic reading guide

- **missing edge**: some conditional independence is suggested by the data,
- **arrowhead**: this endpoint is more constrained,
- **circle mark**: this endpoint is not fully identified from CI information alone,
- **bidirected-looking relation**: possible hidden confounding.

The right mindset

Do not read a PAG as "the true causal graph." Read it as:

the part of the causal structure that remains visible after allowing hidden variables.

Why a DAG-based learner can be misleading

Consider the three-variable mixed graph

What it says

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

But the association between D and Y is not an ordinary directed effect: it may come from hidden confounding.

Why PC can mislead

A DAG learner may see the same skeleton $X - D - Y$ and orient it as $X \to D \leftarrow Y$, because it has no way to represent hidden confounding.

So with latent variables, DAG learners may replace hidden confounding by ordinary arrows.

An economics-style use case for RFCI

Suppose we observe variables such as

education (E), wages (W), test score (T), family background (F), job training (J),

but an important variable such as **ability** is hidden.

Typical concern A DAG learner on the observed variables may confuse

- direct effects,
- indirect effects,
- and hidden confounding.

What RFCI tries to recover

From observed conditional independences, RFCI estimates a PAG: a mixed graph showing

- which adjacencies are supported by the data,
- where hidden confounding may be present,
- and which directions remain uncertain.

Running RFCI in R: a toy example

```
library(pcalg)
set.seed(1)
n <- 3000
U <- rnorm(n) # hidden
X <- rnorm(n)
D <- 0.8*X + 0.9*U + rnorm(n)
M <- 0.9*D + rnorm(n)
Y <- 0.8*M + 0.9*U + rnorm(n)

dat <- cbind(X, D, M, Y)
colnames(dat) <- c("X","D","M","Y")

suffStat <- list(C = cor(dat), n = nrow(dat))

fit <- rfci(suffStat, indepTest =
     gaussCItest, p = ncol(dat), alpha = 0.01)
plot(fit)
```

Hidden DAG behind the simulation

$X \to D \to M \to Y, \qquad U \to D,\ U \to Y.$

Here U is not included in dat.

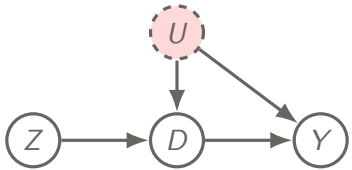
What to look for

- missing edges: possible CIs statements,
- endpoint marks: possible hidden confounding,
- circle marks: directions not identified from CI information alone.

Part 3: IV as the benchmark workaround

IV intuition

If the confounder is hidden, the classical econometric workaround is to find an observed variable Z that moves the treatment but is otherwise exogenous.



Z is a valid instrument if:

- **Relevance:** Z affects D.
- **Exogeneity:** Z is independent of the hidden confounder.
- **Exclusion:** Z affects Y only through D.

Example

D = schooling, Y = earnings, Z = school reform or college proximity.

Applied economics examples of IVs

Military service
Use draft lottery numbers to generate exogenous variation in service.

Education and earnings
Use distance to college, local college openings, or schooling reforms to shift education without directly shifting earnings.

Health and detention
Use physician or judge assignment when assignment is as-good-as random and decision-makers differ in strictness.

Main lesson

IV is often the cleanest benchmark when hidden confounding is serious and a credible instrument is available.

For an econometric treatment of IV, see Angrist and Pischke, *Mostly Harmless Econometrics*, or Wooldridge, *Introductory Econometrics*.

Why two-stage least squares works

Suppose the structural equation is $Y = \alpha D + \varepsilon$, where D is endogenous, so $\mathrm{Cov}(D, \varepsilon) \neq 0$.

Instrument assumptions A valid instrument Z satisfies

$$\mathrm{Cov}(Z, D) \neq 0, \qquad \mathrm{Cov}(Z, \varepsilon) = 0.$$

So Z is related to the treatment, but unrelated to the structural error.

Stage 1: isolate exogenous variation in D
Project D onto Z:

$$D = \pi Z + \text{residual}.$$

This gives fitted values $\widehat{D}$, the part of D explained by the instrument.

Stage 2: estimate the effect
Regress Y on $\widehat{D}$:

$$Y = \alpha \widehat{D} + \text{error}.$$

This setting allows us to incorporate additional controls as well.

Why this targets α

The variation in $\widehat{D}$ is uncorrelated with ε, unlike the raw treatment D.

A tiny IV simulation

Suppose

$$D = 0.8Z + U + \varepsilon_D, \qquad Y = \mathbf{1.0}\, D + U + \varepsilon_Y,$$

where Z is independent of U.

```
set.seed(1)
n <- 4000
Z <- rnorm(n)
U <- rnorm(n)
D <- 0.8*Z + U + rnorm(n)
Y <- 1.0*D + U + rnorm(n)

coef(lm(Y ~ D))

library(AER)
coef(ivreg(Y ~ D | Z))
```

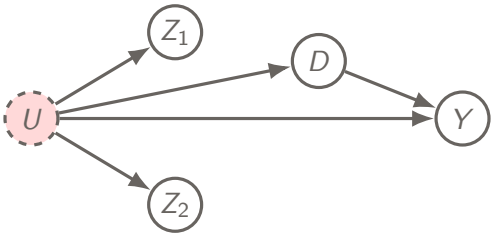
OLS is biased by the hidden confounder U. The command `ivreg(Y ~ D | Z)` implements 2SLS: it first extracts the part of D predicted by Z, then uses that variation to estimate the causal effect.

With heterogeneous treatment effects, IV generally identifies a local effect tied to the units whose treatment is moved by the instrument.

Part 4: Proxy controls and latent factor adjustment

The basic proxy idea

Even if we do not observe the confounder U, we may observe variables that are informative about it.



Idea

The proxies Z_1, Z_2 are not themselves the confounder, but they carry information about it.

Applied examples

Test scores, prior grades, neighborhood variables, baseline outcomes, or balance-sheet variables may all serve as noisy measurements of an unobserved confounder.

One proxy vs several proxies

A single noisy proxy is often not enough to remove confounding.

One proxy

If

$$Z = U + \text{noise},$$

then conditioning on Z may reduce confounding, but usually does not eliminate it.

Several proxies

With multiple noisy measurements of the same latent source, we can often extract a better approximation to the hidden confounder.

Proxy adjustment is usually an approximation strategy: it often reduces bias, but it does not automatically identify the causal effect.

For modern proxy-based identification ideas, see Miao, Geng, and Tchetgen Tchetgen (2018), *Biometrika*.

A tiny proxy-control simulation

Suppose

$$Z_1 = U + \eta_1, \qquad Z_2 = U + \eta_2, \qquad D = U + \varepsilon_D, \qquad Y = 1.0\,D + U + \varepsilon_Y.$$

```
set.seed(1)
n <- 3000
U <- rnorm(n)
Z1 <- U + rnorm(n)
Z2 <- U + rnorm(n)
D <- U + rnorm(n)
Y <- 1.0*D + U + rnorm(n)

coef(lm(Y ~ D))
```

```
coef(lm(Y ~ D + Z1))
coef(lm(Y ~ D + Z1 + Z2))
```

Interpretation

One proxy helps a bit; multiple proxies help more. The point is not exact identification, but partial recovery of the hidden confounder and partial bias reduction.

Latent factor adjustments

Structured proxy adjustment
Suppose many observed variables are all affected by the same hidden source U. Instead of using them one by one as proxy controls, we summarize them through a low-dimensional factor model.

Adjustment strategy

Estimate one or a few latent factors, then include those estimated factors as additional controls in the outcome model.

Economic reading
This is natural when hidden confounding comes from business-cycle conditions, latent productivity, local demand, or market-wide financial stress.

Panel-data intuition

If many firms, regions, or industries are exposed to the same hidden macro shock, then a few latent factors extracted from many observables may absorb part of that confounding.

For interactive effects and latent-factor adjustments in panel settings, see Bai (2009), *Econometrica*.

A tiny latent-factor adjustment in R

```r
set.seed(1)
n <- 2500
U <- rnorm(n)

X1 <- 0.8*U + rnorm(n)
X2 <- 0.7*U + rnorm(n)
X3 <- 0.9*U + rnorm(n)
D <- U + rnorm(n)
Y <- 1.0*D + U + rnorm(n)

X <- cbind(X1, X2, X3)
```

```r
pc1 <- prcomp(X, scale. = TRUE)$x[,1]

coef(lm(Y ~ D))
coef(lm(Y ~ D + pc1))
```

Interpretation

The first principal component acts as a rough estimate of the hidden confounder and can reduce omitted-variable bias.

Take-away

- Hidden confounders break the clean adjustment logic of observed-variable DAGs.
- MAGs give a compact language for the observed implications of latent confounding.
- IV is the benchmark workaround when a credible instrument is available.
- Proxy controls use noisy measurements of the hidden confounder.
- Latent factor adjustments use many observed variables to estimate a hidden source.

When the graph is only partially observed, the goal shifts from perfect recovery to careful partial adjustment.

Looking ahead

Next lecture
We move from hidden confounding and partial adjustment to a different kind of structure: positive dependence and total positivity.

Bridge forward

Today the main difficulty was that hidden variables distort causal interpretation. Next time we will see that even without a causal story, the *sign* of dependence can itself be useful structure for modeling and estimation.