Causality

without headaches

Benoît Rostykus
Senior Machine Learning Researcher

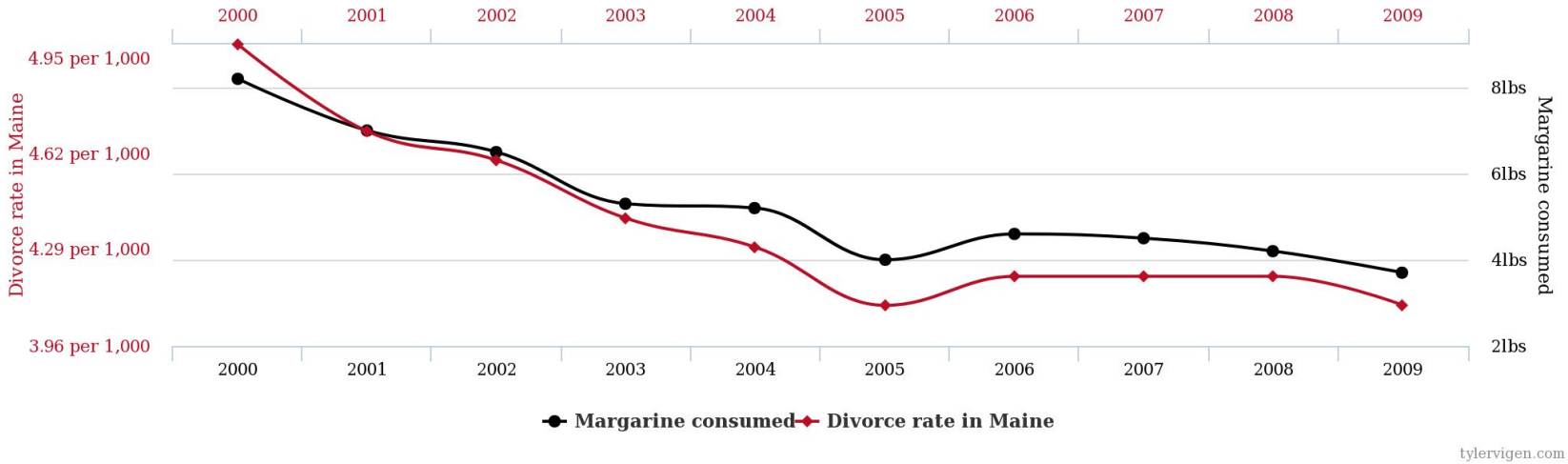
Feb. 23rd, 2018
Talk @ Amazon, SF

NETFLIX

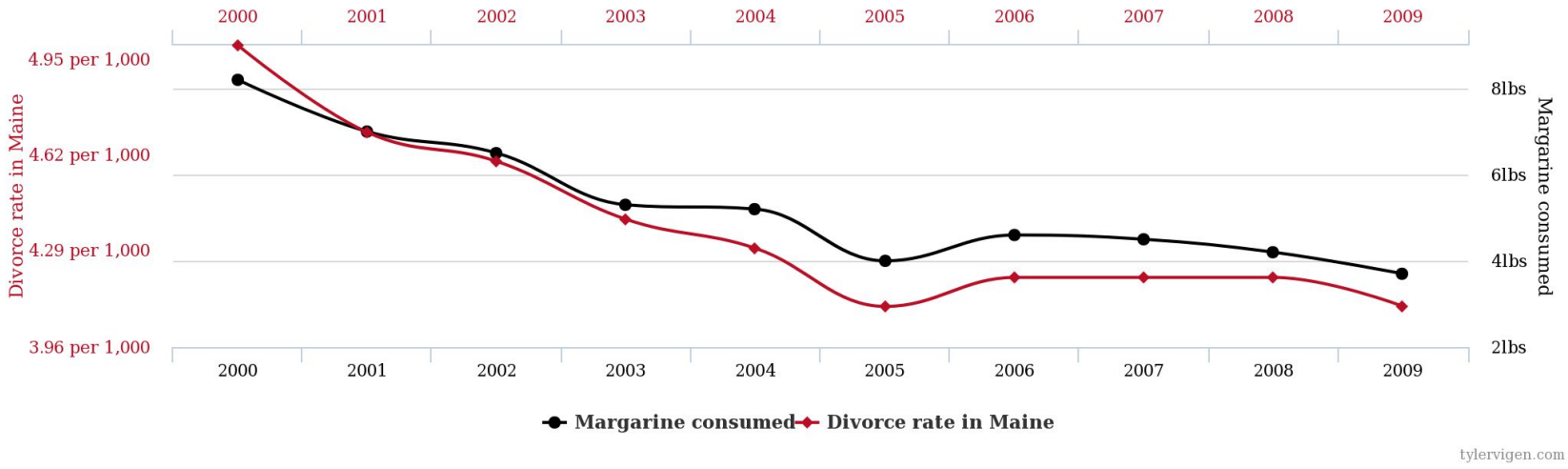
Content

- Correlation ≠ causation
- Randomization & counterfactuals
- Experimental vs observational data
- Inverse Propensity Scoring
- Instrumental variables
  - Generalized Method of Moments
  - Scalable IV regression
  - Weak instruments
- Conclusion

Correlation ≠ causation



Should you stop buying margarine to save your marriage?



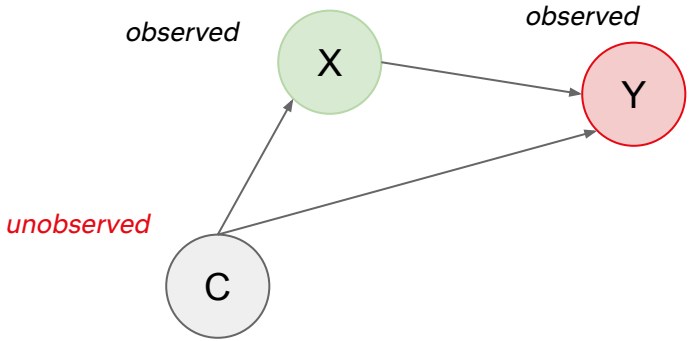
Or should you stay married to eat less margarine?

Correlation(X, Y) is high, does it mean...

… X causes Y ?
… Y causes X ?

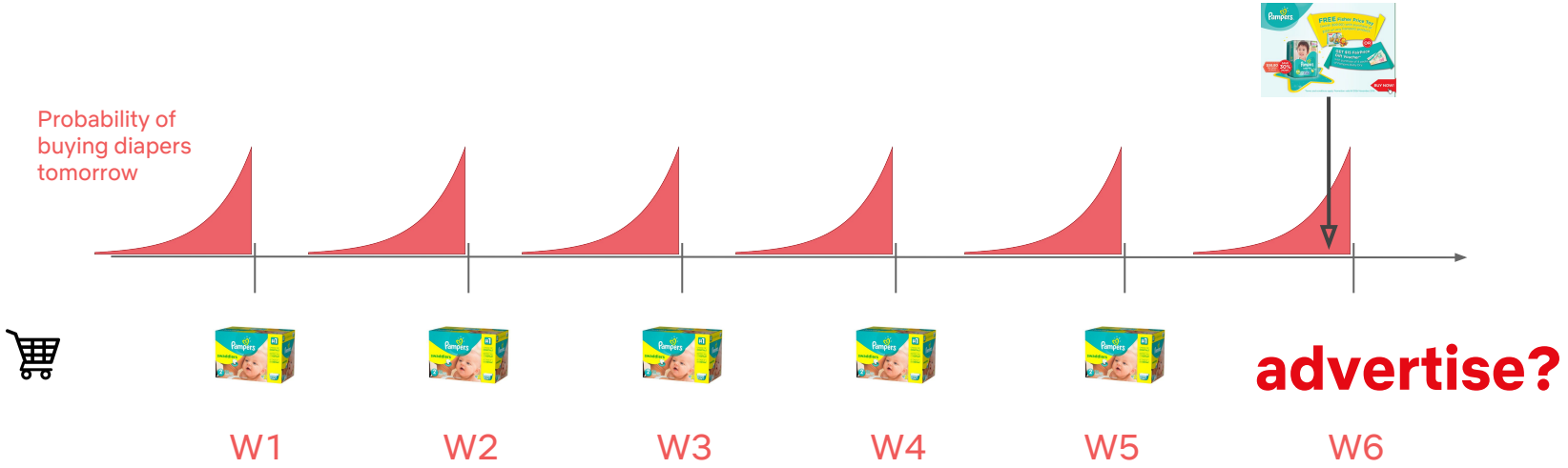
in general, **neither**

most common reason: **unobserved confounder**



“Omitted Variable Bias”

Advertising

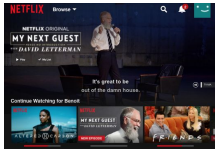


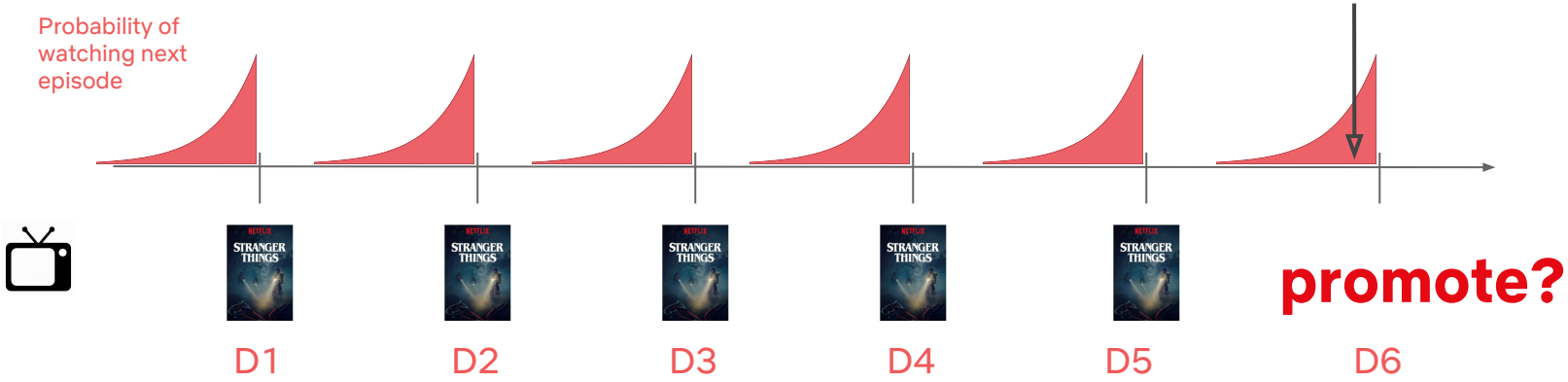
- High probability of conversion the day before weekly groceries irrespective of adverts shown
- Effect of Pampers ads is null in this case.

Traditional (correlational) machine learning will fail and waste $ on useless ads

in practice, **Cost-Per-Incremental-Acquisition** *can be* **> 100x Cost-Per-Acquisition** *(!!!!!)*

Recommendations





Netflix homepage is an **expensive real-estate** (high opportunity cost):

- so many titles to promote
- so few opportunities to recommend

Traditional (correlational) ML systems:

- **take action if probability of positive reward is high, irrespective of reward base rate**
- **don't model incremental effect of taking action (showing recommendation, ad etc.)**

Surely we can do better.

Randomization & counterfactuals

Typical ML pipeline

1) Build model predicting reward probability
2) In AB test, pick UI element that maximizes predicted reward
3) If long-term business metric lift is green, roll out

Typical ML pipeline

1) Build model predicting reward probability
2) In AB test, pick UI element that maximizes predicted reward
3) If long-term business metric lift is green, roll out

i.e

1) Learn P(reward)
2) Max P(reward) to pick arm
3) Evaluate on #(rewards | new policy) - #(rewards | old policy)

Typical ML pipeline

1) Build model predicting reward probability
2) In AB test, pick UI element that maximizes predicted reward
3) If long-term business metric lift is green, roll out

i.e

1) Learn P(reward)
2) Max P(reward) to pick arm
3) Evaluate on #(rewards | new policy) - #(rewards | old policy)

mismatch!

AB tests

- Offer lift measurement by randomizing treatment (= algo)
- Typically user-level counterfactuals
- Counterfactual: “What would happen if we were to use this new algo?”

AB tests

- Offer lift measurement by randomizing treatment (= algo)
- Typically user-level counterfactuals
- Counterfactual: “What would happen if we were to use this new algo?”

We can generalize this!

Counterfactuals: “what would happen if?”

1) Randomize treatment application (binary 1/0 or treatment intensity)
2) Log what happens

- Granularity
  - User-level
    - What happens to user metric X if I use algo A vs B?
    - What happens to user metric X if I always take action A vs action B?
    - What happens to user metric X if I always take action A vs ∅ ?
    - What happens to user metric X if I got treated 20% of the time by action A vs ∅ ?
  - Session-level
  - Impression-level
- Different flavors offer different answers to different causal questions
  - User-level: what happens to retention if we use new algo?
  - Session-level: what happens to session play rate if we use new algo?
  - Impression-level: what happens to CTR if we use new algo?
- User-level not always possible
  - Cannot holdback Strangers Things to some members to see impact on retention

Experimental vs observational data

- When we’re in control of the production system to produce counterfactuals, we call that **experimental data**
- When we don’t control part of the randomization process, we call that **observational data**

Incrementality modeling

Simple experimental example

- On X% of traffic, take no action (or random action)
- On (100-X)% of traffic, take action
- X is typically small because it has a product cost (quality / efficiency / …)

- From collected data, learn:
  - P(reward | features, action) = **f(features)**
  - P(reward | features, no action) = **g(features)**
  - Predicted lift: **lift(features) = f(features) - g(features)**
- Use incremental model in production:
  - Max over arms of lift(features(arm))

Incrementality modeling

Pros

- simple
- no model assumptions, plug your favorite ML technique

Cons

- 2 models
- X is small
  - limit on g model accuracy
  - asymmetrical
- doesn’t explicitly model lift
  - can be hard to calibrate
  - offline metrics?

Inverse Propensity Scoring

One IPS solution

Generalizing the previous example...

In production:

- take action with P(treatment | features)
- take counterfactual action with 1 - P(treatment | features)

We can be in control of P (experimental) or not (observational)

Even when we control P, we might want it non-binary (smooth)

- to control for product quality cost
- to provide enough variance if there's sequential treatment and long-term reward

One IPS solution

1) **Learn model of P(treatment | features)** g_α

2) **Learn incremental model** f_θ **through weighted MLE:**

$$\min_\theta \sum_{i \in \text{treated}} \frac{1}{g_\alpha(x_i)} \ell\left(y_i, f_\theta(x_i, t_i = 1)\right) + \sum_{i \in \text{not treated}} \frac{1}{1 - g_\alpha(x_i)} \ell\left(y_i, f_\theta(x_i, t_i = 0)\right)$$

t_i : treatment variable (usually binary)

$f_\theta(x, t = 1) - f_\theta(x, t = 0)$: predicted lift

f_θ and g_α can have different features set

if sequential treatment, need to condition on full treatment path

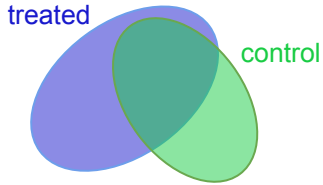
One IPS solution

Pros

- unbiased in theory if no unobserved confounders
- explicitly model (treatment/control) covariates shift
- generic weighted MLE
  - plug in your favorite model
  - your usual ML library already supports it

Cons

- Not robust to unobserved confounders
- g_α needs to have *enough variance* over the features space
- IPSing can blow up variance of f_θ estimate
  - usually resort to clipping
- if g_α is biased, what happens to f_θ ?



Application: Ad incrementality

Does online advertising work?

Application: Ad incrementality

clickbait headlines:

CNBC article, 02/08/2018

P&G Cuts More Than $100 Million in 'Largely Ineffective' Digital Ads

THE WALL STREET JOURNAL. article, 07/27/2017

Application: Ad incrementality

Main problem: measurement

All advertising platforms like **Facebook** and **Google** report on metrics such as Cost-Per-Click (CPC) or Cost-Per-Action (CPA)

Based attribution methodologies such as:

- Last click (only give credit to the last ad clicked before conversion)
- Last view (only give credit to the last ad viewed)
- Any view (any ad “viewed” gets the credit)
- Arbitrary combinations of the previous with fudge factors

These are **non-rigorous ways of estimating causal effect of an ad**

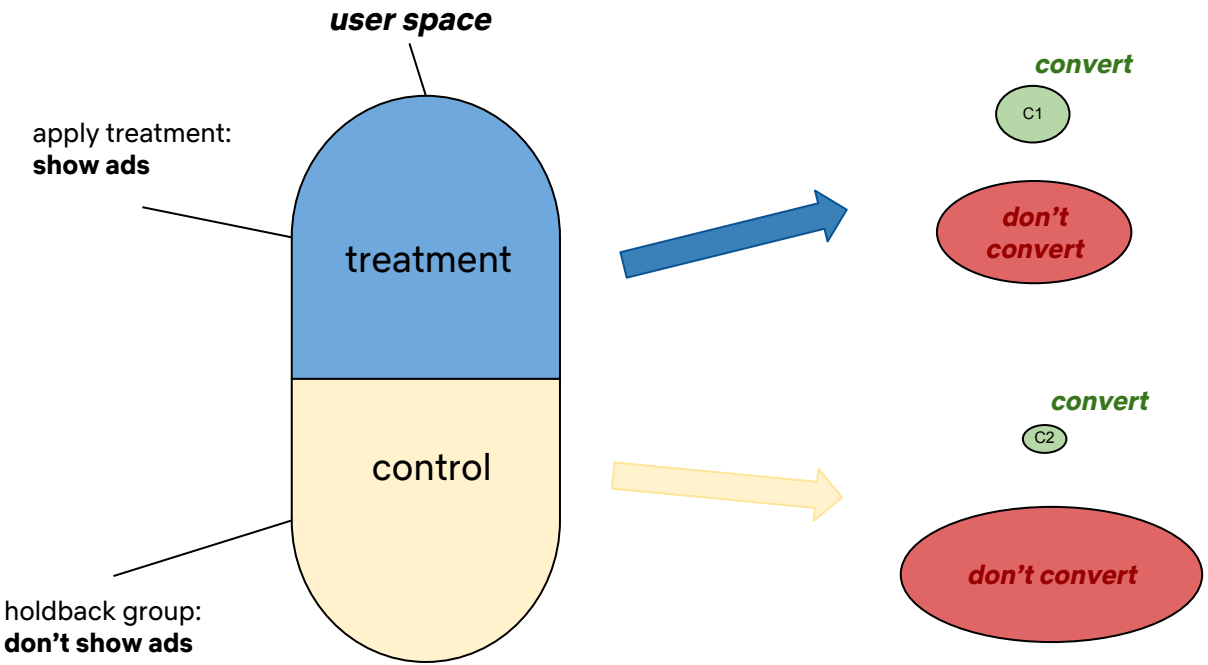
In practice **metrics reported** **over-inflate ad effect by 1 or 2 orders of magnitude**
… because most people would convert anyway
Cost-Per-Incremental-Action > 100x CPA

Rigorous alternative:

1) proper lift measurement using counterfactuals (“ghost bids” / “ghost ads”)
2) Incrementality-based bidding to optimize for ad effect

Application: Ad incrementality

in an ideal world...

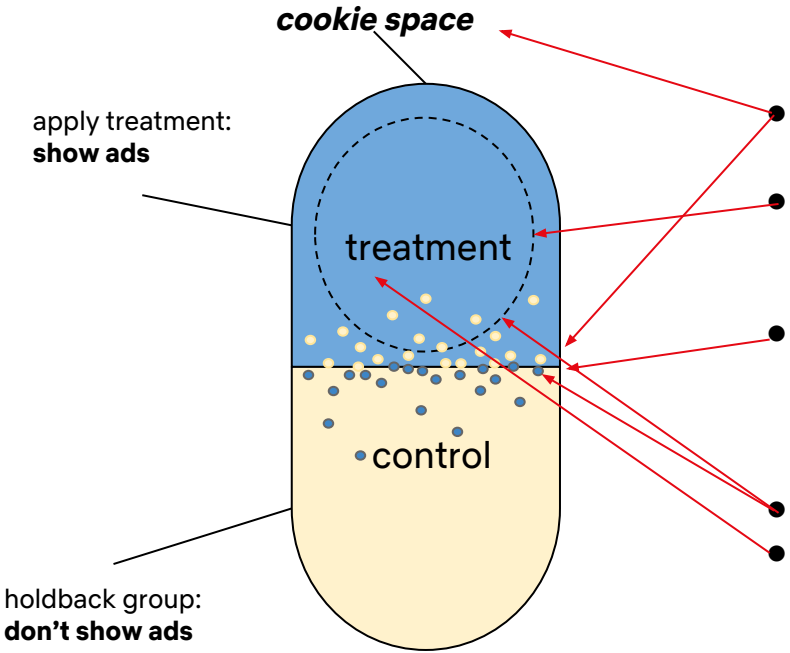


incremental conversions:

C1 - C2

Application: Ad incrementality

in a real world...



- Non-perfect identity
  - cookie ≠ user, cross-device tracking...
- Non-perfect non-random compliance
  - We don't show ads to all treatment cookies
  - Non-deterministic: due to auction mechanisms
- Cross-channel issues
  - Can't easily compare numbers across platforms
  - Holdback group pollution: FB control group is exposed to YT ads
- Heterogeneous logging/reporting tech
- Incrementality varies by ad characteristics
  - Need scores for each impression

Application: Ad incrementality

incremental expected revenue of an impression:

$$LTV(U) \times [\mathbb{P}(conv < 10 \text{ days} | X, U, w = 1) - \mathbb{P}(conv < 10 \text{ days} | X, U, w = 0)]$$

U: user features (browsing history, impressions history...)

X: bid request features (domain, exchange, site categories, ad size, ad position, time of day...)

w: was the auction won (boolean flag)

$LTV(U)$: expected revenue if U subscribes to Netflix

Application: Ad incrementality

Trained through **MLE with IPS**:

$$L(\theta) = \sum_{i \in \mathcal{D}_u} r(X_i) \log(\mathbb{P}_\theta(T = T_i | X_i, U_i, w_i)) + \sum_{i \in \mathcal{D}_c} r(X_i) \log(\mathbb{P}_\theta(T > C_i | X_i, U_i, w_i))$$

$\mathcal{D}_u$: bids associated with converting cookies
$\mathcal{D}_c$: bids associated with non-converting cookies
T_i: observed conversion delay from bid event
C_i: censoring time (joining window length)
$r(X_i)$: IPS weight

} bids = impressions + lost auctions

Auction is a **non-random process** which decides if the treatment (impression) is applied
We need to learn it to get an unbiased estimate of the treatment effect:

$$r(X_i) = \begin{cases} \dfrac{1}{\mathbb{P}(win | X_i)} & \text{if the auction was won} \\ \dfrac{1}{\mathbb{P}(lost | X_i)} & \text{if the auction was lost} \end{cases}$$

Instrumental variables

Instrumental variable

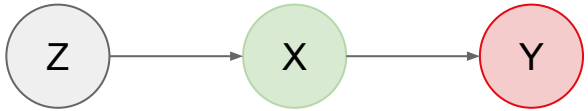
Under the following model:

$$y = f_\theta(x) + \epsilon$$

An **instrument** z is an observed variable following the 2 properties:

1. $\epsilon \perp\!\!\!\perp z$
2. z is correlated with x

z can only influence y through x



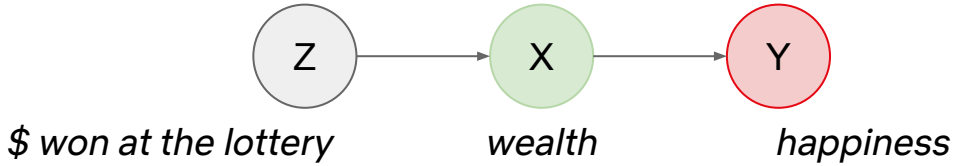
In practice, we replace 1 with a weaker hypothesis:

$$\mathbb{E}\left[\epsilon | z\right] = 0$$

Instrumental variable

Example: Does money make people happy?

Reasonable instrument:



But:

- conditioned on playing (specific demographics)
- Z can affect Y irrespective of $ (fun of playing)

In practice, **finding good instruments** for observational data **is hard**:

"*Many empirical controversies in economics are essentially disputes about whether or not certain variables constitute valid instruments.*" - Davidson & McKinnon book

Instrumental variable (IV) Regression

- The idea is that we can debias our model by using the fact that our instruments explain the endogenous features independently from the regression residual

- **Bread and butter of econometrics**
  - Because we don't have parallel universes to run AB tests on economic policies
  - *Observational data* is sometimes all we have

Let's dive into the details since it is less familiar to people with an ML background

IV regression

From hypothesis:

$$\mathbb{E}\left[\epsilon|z\right] = 0$$

We derive:

$$\mathbb{E}\left[Z^\top \mathcal{E}(\theta)\right] = \overrightarrow{0}$$

with: $X \in \mathbb{R}^{n \times d_X}$ $Z \in \mathbb{R}^{n \times d_Z}$ $Y \in \mathbb{R}^n$ $\mathcal{E}_i(\theta) = y_i - f_\theta(x_i)$ $\forall i \in [\![1, n]\!]$

Because conditional expectation is an orthogonal projection*

*: see here, chapter 4.3

GMM for IV regression

From there, we can see that the inference becomes:

$$\hat{\theta} = \arg\min_{\theta} \mathcal{E}(\theta)^\top Z Z^\top \mathcal{E}(\theta)$$

This is called (the functional form of):
the **G**eneralized **M**ethod of **M**oments for IV regression

The usual econometrics solution to this problem in the linear case is 2-stage least square (2SLS), which expresses the solution through matrix inversion:

$$\hat{\theta} = \left(X^\top Z (Z^\top Z)^{-1} Z^\top X\right)^{-1} X^\top Z (Z^\top Z)^{-1} Z Y$$

… this works for small datasets, but breaks ($O(n^3)$ complexity, $O(n^2)$ storage) for internet-scale data or non-linear models

See this book for thorough treatment of GMM

Scalable GMM for IV regression

Joint work with T. Jebara,
to be published

We want a solution that scales linearly with:

- # of training points
- # of non-zero features per row (sparse high dimensional X)
- # of non-zero instruments per row (sparse high dimensional Z)

We are trying to minimize:

$$L(\theta) = \sum_{m=1}^{d_Z} \left(\sum_{i=1}^{n} z_{i,m} \epsilon_i(\theta) \right)^2$$

or:

$$L(\theta) = \sum_{i,j} \left[\epsilon_i(\theta) \epsilon_j(\theta) \sum_{m=1}^{d_Z} z_{i,m} z_{j,m} \right] = \sum_{i,j} \ell_{i,j}(\theta)$$

Scalable GMM for IV regression

Joint work with T. Jebara, to be published

Idea: pairwise importance-sampling SGD

$$L(\theta) \propto \sum_{i=1}^{n}\sum_{j=1}^{n} \epsilon_i(\theta)\epsilon_j(\theta)p(i,j)$$

$$p(i,j) = \sum_{m=1}^{d_Z} p(i|m)p(j|m)p(m)$$

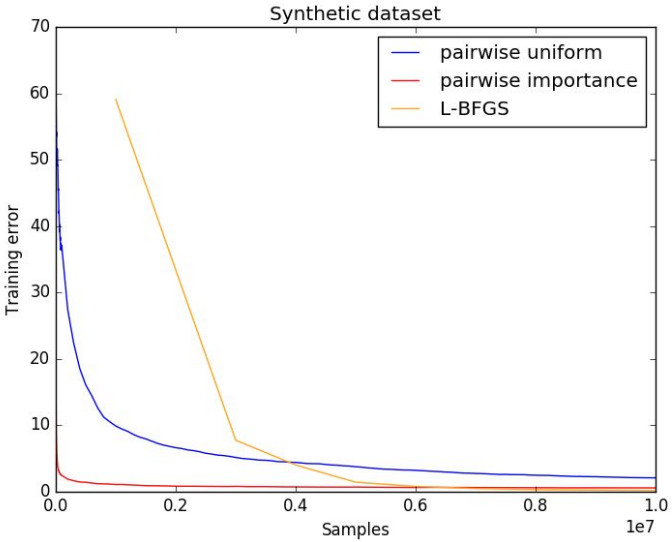
$$\begin{aligned} p(i|m) &= \frac{z_{i,m}}{\sum_{k=1}^{n} z_{k,m}} \\ p(j|m) &= \frac{z_{j,m}}{\sum_{k=1}^{n} z_{k,m}} \\ p(m) &= \frac{(\sum_{i=1}^{n} z_{i,m})^2}{\sum_{n=1}^{d_Z}(\sum_{i=1}^{n} z_{i,n})^2} \end{aligned}$$

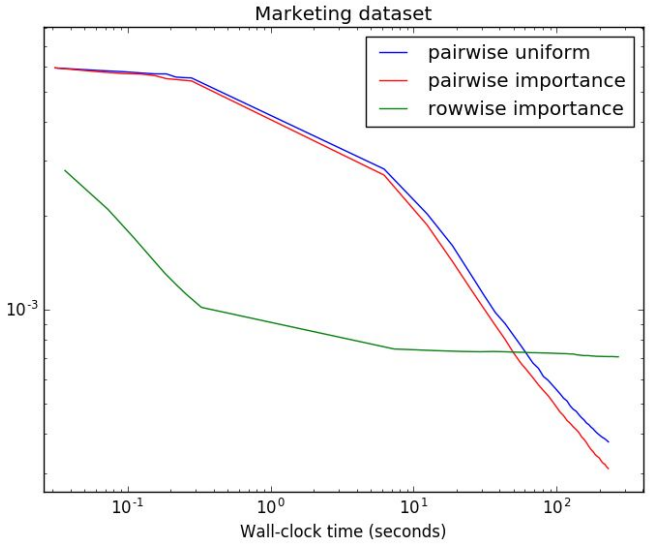
More likely to sample rows which matter (large instruments): will converge fast

Scalable GMM for IV regression

Joint work with T. Jebara,
to be published

Converges faster than known alternatives for non-linear GMM:





Bonus: Extra variance-reduction around the point-estimate thanks to importance sampling!
(hypothesis: same effect as efficient GMM)

Weak instruments

Joint work with D. Hubbard

Now that ***we can*** run IV regression on problems:
- with millions of features
- with millions of instruments

Should we do it?

Problem: when instruments are weak (low correlation with X), the IV estimator is biased

… first towards the correlational answer
… but then unbounded

Causal answer can become worse than correlational one with IV too!

Weak instruments

Joint work with D. Hubbard

Simple 1D experiment

$$hours = \pi_0 + \pi_1 z + \pi_2 wealth + v$$

$$y = \beta_0 + \beta_1 hours + \beta_2 wealth + u$$

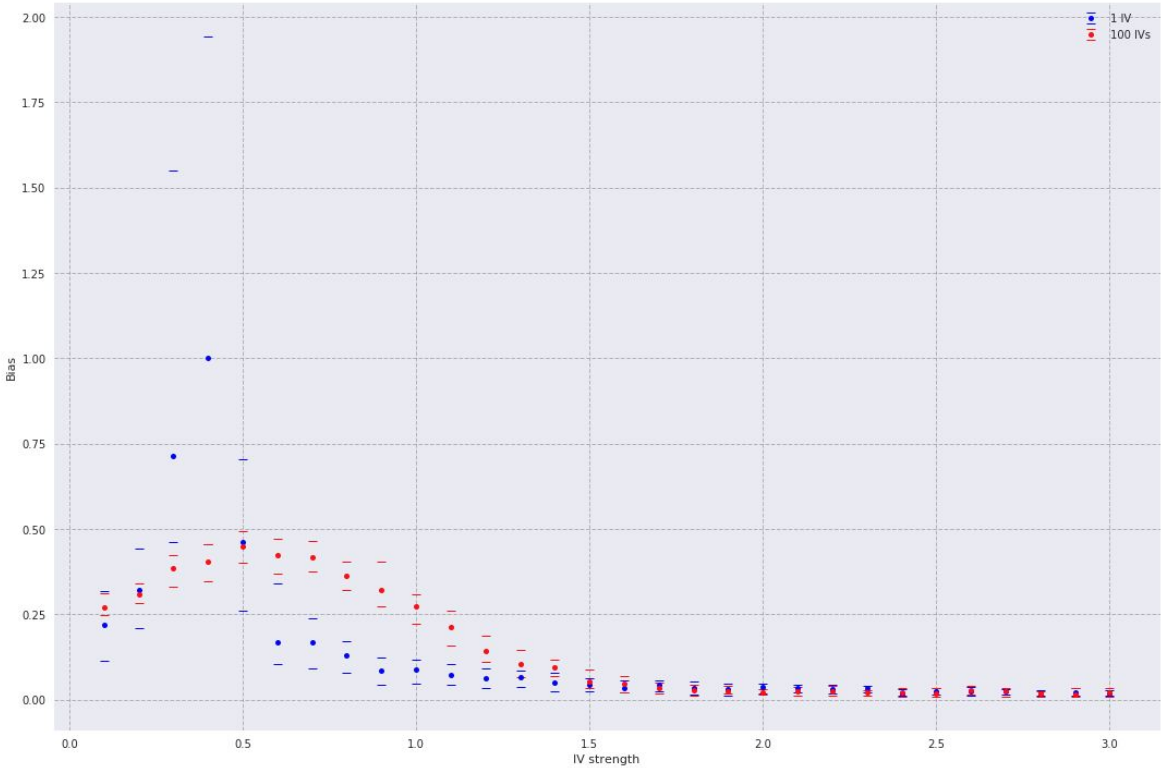
$$\begin{pmatrix} u \\ v \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_u & \rho \\ \rho & \sigma_v \end{pmatrix}\right]$$

observed: hours, z, y
unobserved: wealth

Try to recover: $\beta_1 = 0.75$

Weak instruments

Joint work with D. Hubbard



- Can 100 weaker IVs replace 1 good IV? It depends
- Non-monotonic behavior in the very weak regime :(

Weak instruments

Joint work with D. Hubbard

Controlling for instruments quality is crucial

How to do it in a meaningful and scalable way?

- Partial answer from the statistics literature:
  - Partial F-tests
  - Cragg–Donald
  - Anderson-Rubin
  - …
- “Regularize” the instruments
  - Which cross-validation metric to use? Circular problem. No ground truth!

IPS-MLE vs IV-GMM

(other methods exist! Ex: propensity matching)

Both are unbiased and consistent when there are no unobserved confounders.
Typical estimates have higher variance than their correlational counterparts.

IPS-MLE

- Familiar to ML people
- More flexible on model class
- Easy to scale
- Less theoretical guarantees
- Not robust to unobserved confounders
- Bias and variance come from IPS weight

IV-GMM

- Familiar to econ people
- Mostly gaussian residuals
- Harder to scale
- More theoretical guarantees
- Robust to unobserved confounders
- Bias and variance come from IV strength

In both cases, there is **no built-in fallback** to correlational answer if randomization is poor

Conclusion

Applications

Plenty of use cases for causal inference at Netflix

- Advertising
- Causal recommendations
- Content valuation
- Increased experimentation power
- ...

Causal inference in practice

Hard! because:

Causal effects are small
Asymptotic unbiasedness is useless if the variance dominates, even on large datasets
Variance grows even more when there is sequential treatment

Unobserved confounders can have bigger magnitude than what we try to measure

Plenty of unsatisfactory / unanswered questions in the literature

No clean ground-truth
All estimators have their flaws.
Hard (impossible) to measure and compare biases offline on large-scale problems

When it matters

Correlational models are fine...

- When we only care about fitting the data / predicting
- When your model predictions won't interact with the product

Causal models can help...

- **When there's a "why?"**
  - "why did NFLX stock price move today?"
- **When there's a "what would happen if?"**
  - "what would happen to streaming if iOS app was 10% slower?"
- **To build cost-efficient ML algorithms**
  - incremental models factor in *the effect of taking the action suggested*
  - aligned with business metrics *lift* : maximize likelihood of green AB test
  - … it's just a greedy one-step-ahead Reinforcement Learning strategy

Thank you.

NETFLIX

Appendix

Clearing-price modeling for ad lift

We build an estimator of the clearing price distribution to be able to derive $\mathbb{P}(win|X_i)$

Challenges:

- Highly contextual (domains, liquidity, market dynamics…)
- Partial information
  - clearing price observed only when auction is won
  - otherwise bid price is a right-censoring point

Solution:

Joint censored conditional quantiles regression of the clearing price for a grid of 60 quantiles

=> **Non-parametric** estimate of the clearing price distribution

- includes features such as domain, ad position...