Estimating regions of positive treatment effect

Aahlad Manas Puli
NYU
Joint work with Uri Shalit and David Sontag

CILVR Seminar March 2017

Primer on causal effects

Q : How does treatment A affect this population on average compared with treatment B? (Average treatment effect **ATE**)

A : Randomly assign people to treat (treated group), to not treat (control group). Compute average of each group separately. Compute difference.

This is a **randomized study.** Expensive or hard to conduct on large populations.

But general patient records easier to obtain and relatively inexpensive. This is an **Observational study.** Suffers from selection bias.

What else can we ask?

Causal effects …

A more general problem is to estimate the **Individual treatment effect (ITE)**.

Treatments A and B are used to treat Diabetes and patient **x** has diabetes.

Let the outcome, if treated with B, be ***$y_1(x)$,*** and if treated by A, be ***$y_0(x)$.***

$$ITE(x) := y_1(x) - y_0(x)$$

If ***ITE(x)*** is positive treat with B, with A otherwise.

Responders - Example

Drug Metformin (**treatment A**) is the standard for diabetes. We want test a much higher dose of Metformin (**treatment B**). (Why is higher dose bad?)

For most of the the general population the increase in dose doesn't affect blood-sugar level (**the outcome**) much more than the standard dose.

But … it turns out people with two copies of a variant of a gene SLC2A2 respond much better to 1.5 times the standard dose (Zhou et al., *Nature Genetics,* 2016)

So generally, to compare the effect on two sets of people with treatments **A**, **B**

> Find patients whose response to **B** is different from their response to **A**.

Individual Treatment Effect (ITE): Responders

Responders are the subpopulation who have **positive ITE.**

Often either very few people are responders.

… might be ignored as outliers.

or the effect itself might be **very small.**

… might be overlooked as noise.

Back to the problem

Given a population, find the subset that responds to treatment.

Formally, we have $\boldsymbol{y_1(x)}$ which is the treated outcome

$\boldsymbol{y_0(x)}$ which is the control outcome.

Find all $\boldsymbol{x}$ which have $\boldsymbol{y_1(x) - y_0(x) > 0}$.

$\boldsymbol{y_0(x)}$ can be thought of as the **background function**. The interesting effects only occur in $\boldsymbol{y_1(x)}$, thought of as **foreground function.**

Connection to Contrastive learning

Consider this setting : Assume we want to cluster diabetes patients. But any clustering would cluster based on age because it accounts for most of the effects on health.

How does one discount the effect of age?

Contrastive learning as defined by Zou et al. (NIPS 2013) aims to learn topics from a **foreground** distribution which don’t exist in a **background** distribution.

Zou et al. use **tensor decompositions** on moment tensors.

Back to the problem

Given a population, find the subset that responds to treatment.

Formally, we have $\boldsymbol{y_1(x)}$ which is the treated outcome

$\boldsymbol{y_0(x)}$ which is the control outcome.

Find all $\boldsymbol{x}$ which have $\boldsymbol{y_1(x) - y_0(x) > 0.}$

$\boldsymbol{y_0(x)}$ can be thought of as the **background function**. The interesting effects only occur in $\boldsymbol{y_1(x),}$ thought of as **foreground function.**

Predicting responders : The obvious

Problem : Find all ***x*** which have $\boldsymbol{y_1(x) - y_0(x) > 0}$.

Solution : Fit $\boldsymbol{y_1(x)}$ *and* $\boldsymbol{y_0(x)}$ separately and take **difference.**

Cluster the differences $\boldsymbol{y_1(x) - y_0(x)}$.

Required assumption: Good estimation of $\boldsymbol{y_1(x)}$ *and* $\boldsymbol{y_0(x)}$.

When might this fail?

The obvious might fail?

1. **Weak signal**: The variations in $\boldsymbol{y_1(x)}$ - $\boldsymbol{y_0(x)}$ are small.
   Features like **age** and **location** cause most of the variation in blood-glucose.

2. **Small region*:*** Very few people react positively.

What could be improved?

1. Using simplicity of the response
2. Accounting for the responders explicitly

Model

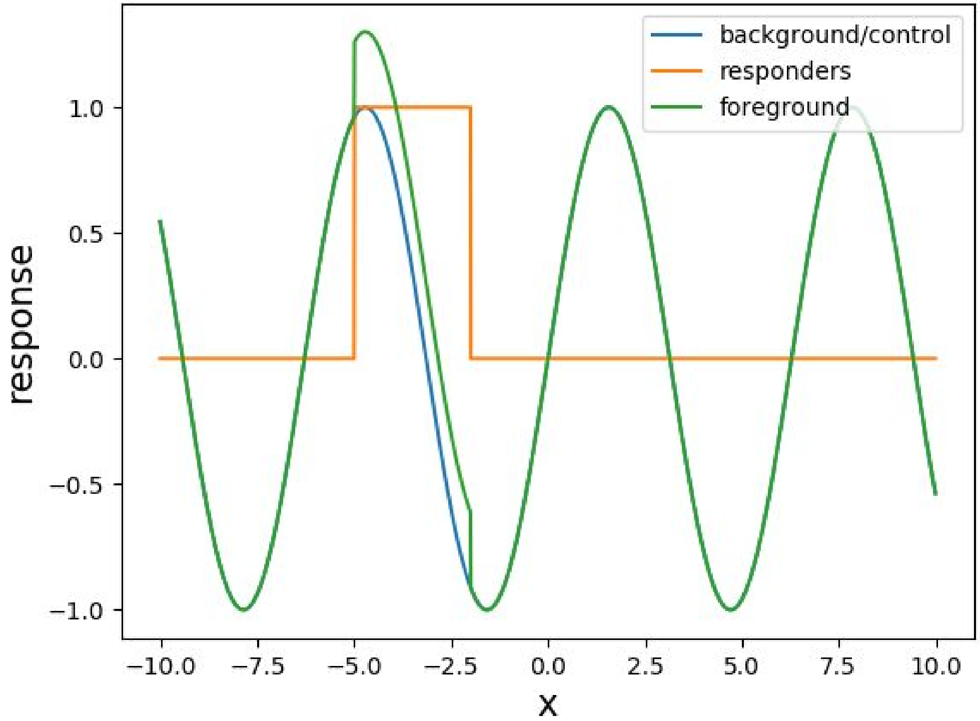
We assume an additive model for $y_1(x)$ *and* $y_0(x)$. Assume functions $f(x),g(x)$ with noise e

$$\begin{aligned} y_0(x) &= f(x) + e \text{ (many samples)} \\ y_1(x) &= f(x) + r(x) \cdot g(x) + e \text{ (few samples)} \end{aligned}$$

f(x) is background function

r(x) is a binary function that models the **responders** directly.

g(x) is the response (or) magnitude of effect.


background/control
responders
foreground
response
x
1.0
0.5
0.0
−0.5
−1.0
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0

$Y_1(x)$ and $Y_0(x)$

Model - why?

$$y_0(x) = f(x) + e$$
$$y_1(x) = f(x) + r(x) \cdot g(x) + e$$

We decouple the strength of the effect from those that show it.

1. We don’t ignore responders; **by explicitly focusing on Responders.**

2. Simpler estimation when **response** is simple.
   If drug **B**’s response is constant, we need few responders.

3. Could incorporate **prior knowledge** into formulation.

More model ...

Considering different settings where the decomposition helps

$$y_0(x) \;= f(x) \;+ e$$
$$y_1(x) \;= f(x) \;+ r(x) \bullet g(x) \;+ e$$

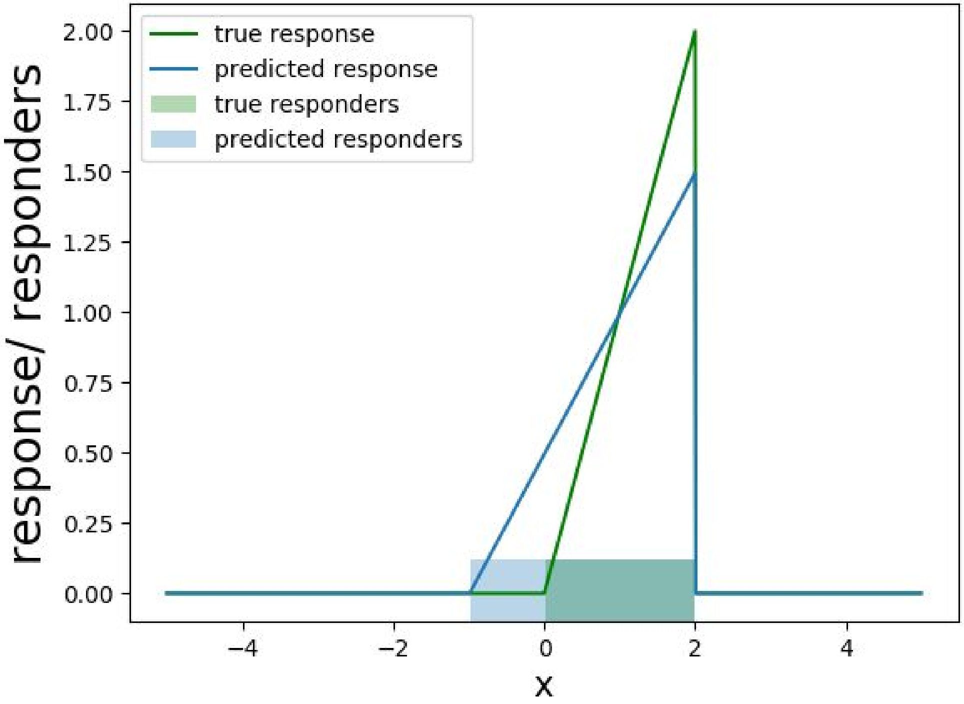
	g(x) simple	g(x) complex
r(x) simple	Decomposition helps	Not considered yet
r(x) complex	**Decomposition helps**	Decomposition doesn’t help

What we propose

Background f(x) can be estimated very precisely from controls

Any treated units that deviate from prediction of f(x) are “suspected responders”

Regression error vs. classification error



Why validation error?

For **linear response**, if **S** has **at least k non-responders**, with probability at least **1-δ** the validation risk has a lower bound.

$$\mathbf{R_{val}(\hat{f}) \geq k\gamma^2 + noise + O(1/\sqrt{n}))}$$

$\mathbf{R_{val}(f)}$ - Validation error

γ - Error outside R at least with probability at least **1-δ**

What we propose

Background f(x) can be estimated very precisely from controls

Any treated units that deviate from prediction of f(x) are “suspected responders”

1. Fit background function $\boldsymbol{f(x) = y_0(x)}$ on the control samples.
2. For each candidate set **S** of **suspected responders:**
   - Split **S** into train and validation
   - Fit foreground function $\boldsymbol{\hat{g}(x) = y_1(x) - f(x)}$ on $\boldsymbol{x \in S_{train}}$
3. Choose set **S*** that corresponds to the minimal validation error of $\boldsymbol{\hat{g}(x)}$ on $\boldsymbol{S_{val}}$
4. Fit function $\boldsymbol{\hat{r}(x)}$ to members of **S***

Problem: Exponentially many subsets **S**

How to control the number of candidate sets: Two way thresholding

Intuitively the **responders** give large absolute residuals.

Split the **residuals (y_1(x) - $\bar{f}$(x))** into 3 parts.

1. **Above T_1 :**
   Suspected Responder
2. **Below T_0 :**
   Suspected Non-responder
3. **Between T_1 and T_0** :
   Ignored in learning the split.

Why 2 thresholds?
To induce a trade-off between noise and number of samples.

In figure, ***g(x) is a constant 3.***
Thresholds at **4** and **-1.**

Main algorithm

1. Create held-out validation set.
2. Fit ***f(x)*** only on ***y_0(x).***
3. For all **threshold pairs**(to predict responders)
    a. Fit classifier **c(x)** on label for **suspected responders** and predicted non-responders.
    b. Fit the ***g(x)*** on ***y_1(x) - f(x)*** on all those treated in training with **c(x) = 1.**
    c. Compute validation error like before.
4. Return the threshold pair (and resulting **c(x)**), with the lowest validation error.

Improvements: **O(n^2)** number of subsets.
Much lesser chance of overfitting.

Experiments - Synthetic data

We use a linear region for now. Cases considered :

1. g(x) = 3, constant. 2 dimensional data
2. g(x) = 3, constant. 100 dimensional data
3. g(x) is linear (sampled). 2 dimensional data.

Baseline

1. Fit rbf-kernel Support vector regression on control and treated functions.
2. Cluster residuals on the treated samples using k-means. 2-clusters.
3. Fit a linear SVM with cluster ids as labels.

Experiments - Synthetic data (Linear region, 100 experiments)

Method	*g(x)*	*d*	*Sample size*	Binary Classification error (mean,median)
Baseline	Constant 3	2	5000	3.4%, 0.6%
Our method	**Constant 3**	**2**	**5000**	**1.4%, 0.4%**
Baseline	Constant 3	100	5000	4.0%, 0.4%
Our method	**Constant 3**	**100**	**5000**	**2.8%, 0.27%**
Baseline	Linear sampled	2	5000	30%, 24%
Our method	**Linear sampled**	**2**	**5000**	**5.4%, 1.2%**

Applications

- Causal Inference and effect prediction -
    - Predicting responders.
    - Predicting ITE better.
- Find other supervised foreground / background phenomena
    - Find interesting cases with weak foreground over strong background
    - Foreground active in small subset of space

Future work

1. More baselines!
2. Real-world experiments
3. A better way to choose thresholds.
4. A more general proof for using regression error as surrogate.
5. Finer analysis trading off complexity between **responders** and **strength of effect.**

Thanks! Questions?