DeepIV: A Flexible Approach for Counterfactual Prediction

Jason Hartford and Kevin Leyton-Brown
University of British Columbia

Greg Lewis* and Matt Taddy†
Microsoft Research & NBER / † University of Chicago
I need a model that predicts the effect of price on ticket sales
We can raise prices and get more sales!
Prediction with confounding effects

\[ y = g(p) \]

Annual Sales, \( p = g(p, x) + 0.00 \times e \)

Ticket Price ($), \( p = f(x, z, 0.00 \times e) \)
Prediction with confounding effects

\[ y = g(p, e) \]

Annual Sales, \( p = g(p, x) + 0.00 \times e \)

Sales

Price

\( p \)
Prediction with confounding effects

\[ y = g(p, e) \]

Automated pricing engine increases prices as the plane fills

\[ p = f(e) \]
The observational distribution

\[ y = g(p, x) \]

\[ p = f(x) \]

"features / observed confounders"

\[ x \]

"policy / treatment"

"response"
The **interventional distribution**

```
features / observed confounders
```

\[ x \]

\[ \text{Holidays} \rightarrow \text{Sales} \]

\[ \text{Price} \leftarrow \text{Sales} \]

\[ \text{Set } p = \hat{p} \]

\[ P(y|\text{do}(\hat{p}), x) \]

**“response”**

**“policy / treatment”**
Identification of causal effects

If $x, p \& y$ observed, $P(y|\text{do}(p), x)$ is identified. See e.g. [Athey et al. 2016], [Shalit et al. 2017]

\[ y = g(p, x) \]

\[ p = f(x) \]
Identification of causal effects

Not identified without further assumptions

"features / observed confounders"

x

Holidays

Sales

"response"

y = g(p, x, e)

Price

Conference

"latent / unobserved confounders"

p = f(x, e)

"policy / treatment"
Identification of causal effects

Variable that only affects the response \textbf{indirectly} via its effect on price

\[ y = g(p, x) + e \]

Additive latent effects

\[ p = f(x, z, e) \]

"policy / treatment"

"features / observed confounders"

"instrument"

"response"

Conference

"latent / unobserved confounders"
Simulate a world without latent effects on price

Annual Sales, $p = g(p, x) + 1.00 \times e$

Ticket Price ($), $p = f(x, z, 1.00 \times e)$
Simulate a world without latent effects on price
The learning problem

These assumptions imply the following identity\(^1\),

\[
E[y|x, z] = E[g(p, x)|x, z] = \int g(p, x) dF(p|x, z)
\]

So we can recover \(g(p, x)\) solve the implied inversion problem...

\[
\min_{g \in \mathcal{G}} \sum_{t=1}^{n} \left( y_t - \int g(p, x_t) dF(p|x, z) \right)^2
\]

1. This holds if \(E[e|x] = 0\). In general we recover \(g(p, x)\) up to a constant wrt \(p\) – see paper for details.
A two-stage solution

\[
\min_{g \in G} \sum_{t=1}^{n} \left( y_t - \int g(p, x_t) \, dF(p \mid x, z) \right)^2
\]

**Stage 1:** fit \( F_{\phi}(p \mid x, z) \) using the model of your choice.

We use **mixture density networks** [Bishop 94]

\( \hat{F}_{\phi}(p \mid x, z) \)

**Stage 2:** train network \( \hat{g}_{\theta} \) using stochastic gradient descent with monte-carlo integration.

\[
\nabla L(\theta) = -2 \left( \frac{1}{\mid \hat{p}_1 \mid} \sum_{\hat{p}_1 \sim \hat{F}(p \mid x, z)} \hat{g}(\hat{p}_1, x_t) \right) \times \left( \frac{1}{\mid \hat{p}_2 \mid} \sum_{\hat{p}_2 \sim \hat{F}(p \mid x, z)} \nabla_{\theta} \hat{g}(\hat{p}_2, x_t) \right)
\]

At each SGD iteration

Sample

\( \hat{g}_{\theta}(p, x) \)
Causal Validation

• In general, out-of-sample validation causal models is challenging / impossible...

• But… both our losses depend only on observable quantities and reflect causal loss, so we can simply use standard validation sets.
Evaluation

Simulation & Bing Ads Experiments
Simulation Experiments

- Price Sensitivity
- Jan, Jun, Dec
- Holidays
- Customer type
- Conference
- Customer features
- Fuel Cost
- Ticket Price
- Ticket Sales

$\rho$ lets us smoothly vary the correlation between sales and price.

$t \sim U\{0, 1, \ldots, 6\}$
Simulation – low dimensional feature space
Simulation – low dimensional feature space

Out-of-Sample MSE (log scale)

- $\rho = 0.1$
- $\rho = 0.25$

Training Sample in 1000s

FFNet
Simulation – low dimensional feature space
Simulation – low dimensional feature space

Out-of-Sample MSE (log scale)

$\rho = 0.1$

$\rho = 0.25$

$\rho = 0.5$

$\rho = 0.75$

Training Sample in 1000s

FFNet
Simulation – low dimensional feature space
Simulation – low dimensional feature space
Simulation – low dimensional feature space

Out-of-Sample MSE (log scale)

$\rho = 0.1$

$\rho = 0.25$

$\rho = 0.5$

$\rho = 0.75$

$\rho = 0.9$

[Darolles et al. 2011]
Simulation – low dimensional feature space
Implications and future directions

• We recover heterogeneous treatment effects in settings with unobserved confounding effects for both discrete and continuous variables… and SGD scales naturally to very large datasets.

• Can leverage the flexibility of deep nets for rich data types. E.g. raw text in our Bing ads application experiments / images in simulation.

Future work:

• Methods for uncertainty estimates over predictions.