

Economic AI

Matt Taddy (Chicago + MSR) Greg Lewis (MSR)

Jason Hartford (UBC) Kevin Leyton-Brown (UBC)

Kui Tang (Columbia) Mengting Wan (UCSD) Richard Li (UW)

Matt Goldman (MSFT) Justin Rao (MSR) Di Wang (MSFT)

What do economists do?

320

	JUNE 1980			
	TABLE 2—TOTAL EXPENDITURE AND OWN-PRICE ELASTICITIES			
	Unconstrained e_{ii}	Homogeneous e_i	First-Differences Model e_{ii}	Homogeneous e_i
Food	0.21	-0.07	-0.01	0.17
Clothing	2.00	-0.92	0.04	2.92
Housing	0.30	-0.31	1.51	-0.02
Fuel	1.67	-0.28	0.79	0.84
Drink and Tobacco	1.22	-0.60	1.37	1.17
Transport and Communication	1.23	-1.21	-0.48	-0.67
Other goods	1.21	-0.72	-0.16	-0.23
Other services	1.40	-0.93	-0.62	-0.52
		1.28	-0.78	-0.78
		1.73	1.00	0.00
		1.15	1.37	-0.31
			1.14	-0.00
			2.03	-0.67
			1.03	-0.22
				-0.94
				0.04
				-0.31
				0.00
				-0.67
				0.17
				2.92
				-0.02
				0.84
				1.17

the D.W. statistic shows a sharp

heterogeneity is not a new

problem; Ray Byron;

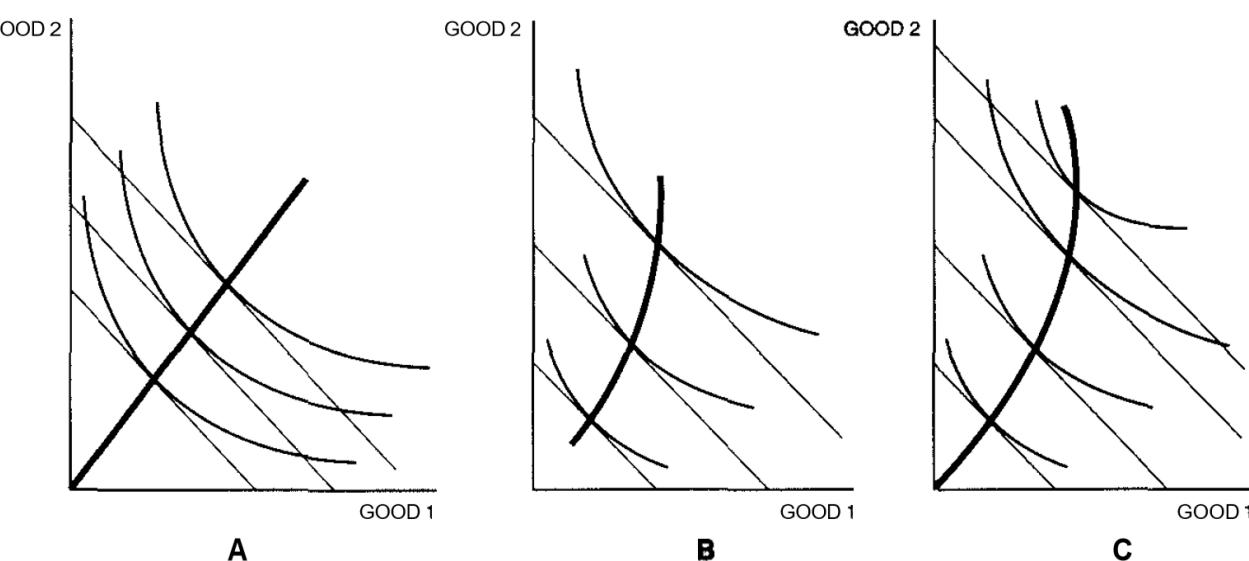
described to a

problem as far

as temporal

A discussion of aggregation above, is that it assume that k , the distribution of household income structure

Finally, the assumption of separability



Income expansion paths. Panel A depicts unit elastic demands, in panel B good 2 is a luxury good, and in panel C, good 1 is an inferior good.

What do they need to do today?

Microsoft Azure

New HDInsight Cluster

Cluster Type configuration

Cluster Type configuration

Report a bug

Cluster Type configuration

Learn about HDInsight and cluster versions. [Learn more](#)

Cluster Type: Spark

Cluster Tier ([more info](#)): STANDARD

Operating System: Linux

Version: Spark 2.0.0 (HDI 3.5)

Administration

Manage, monitor, connect

Scalability

On-demand node scaling

99.9% Uptime SLA

toddler shoes

Web Images Videos Maps News Explore

91,800,000 RESULTS Any time

[Shop DSW Kids Shoes | dsw.com](#)
www.dsw.com/kids · DSW, Inc.

The Latest **Kids** Styles @ Participating DSW Stores Today!
Find the Best Selection of Infant, Toddler, and Best Brands for Infants, Toddlers, and ...
taddy@microsoft.com

Sign Up for DSW® Rewards
Earn a \$10 Certificate with Your First Purchase. Free to Enroll!

Find a Store Near You
More Than 480 Locations Available. Shop at a DSW® Near You Today!

Ads

New Balance 150 Slip \$39.99 New Balance

Off Sale - Last Day!

What can we (AI/ML) do to help?

The **dimension and complexity** of the problem space has exploded
We can develop ML to navigate this space: **stay safe and automate**

Econ AI is about systems of ML tasks for econometric questions

Example: Demand System

Suppose that you have transactions 't' on products 'j'.

Write the quantity bought 'q' as

$$q_{tj} = \alpha_{tj} + \gamma_j p_{tj} + e_{tj}$$

a function of utility we can (α_{tj}) and can't (e_{tj}) see, plus price p_{tj} .

You need to have a model like this to target customers or set prices.

But it's a system!

For example: There many different products

Demand for j depends on **substitutes** and **complements**

Or: where does price come from?

$$\log p_{tj} = \varphi_{tj} + \psi_j q_{tj}^* + \nu_{tj}$$

and the *demand system* is in equilibrium when $q_{tj}^* = q_{tj}$

Product Co-occurrence

Ignoring price and observable demand shifters...

$$q_{tj} = \alpha_{tj} + \gamma_j \log p_{tj} + \varepsilon_{tj}$$

Classic “data mining” seeks *association rules* in *market baskets*:

Find $j \neq k$ pairs so that $\mathbb{E}[q_{tj}q_{tk}] \gg \mathbb{E}q_{tj}\mathbb{E}q_{tk}$

We can do better

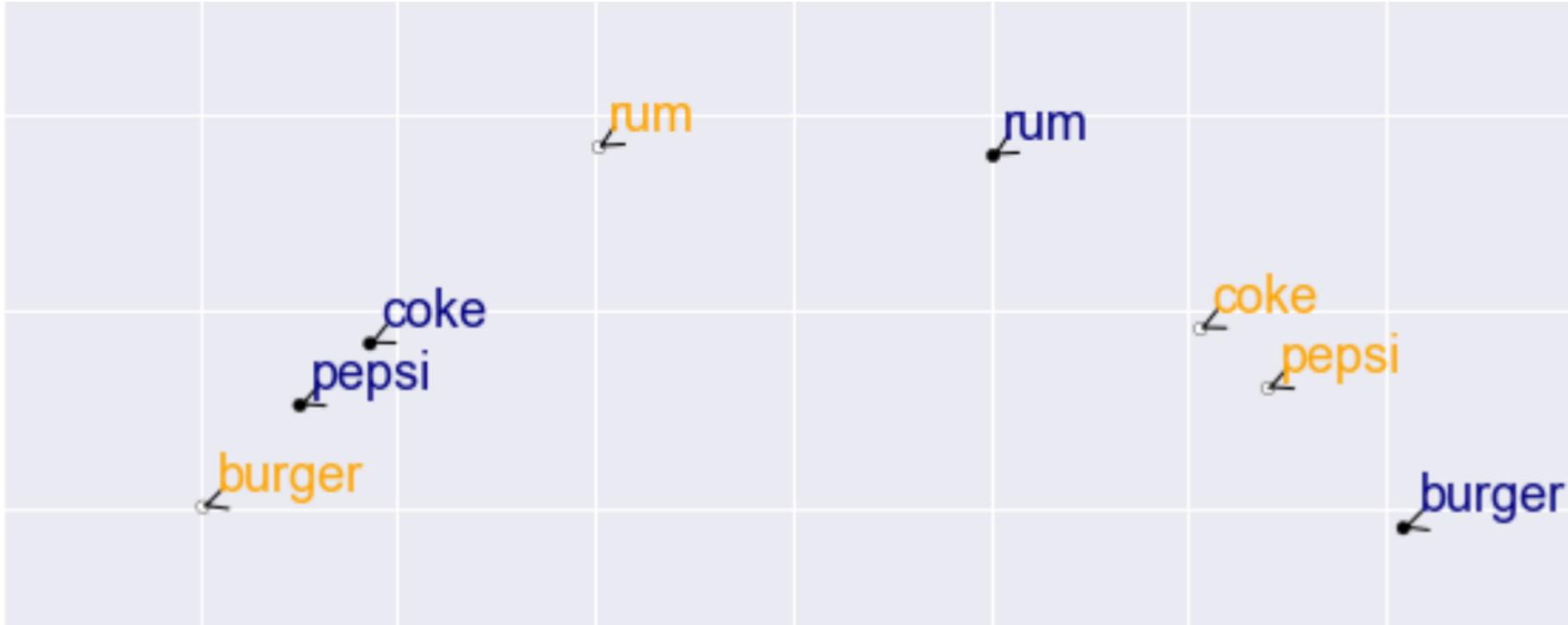
From word to product embedding

Words are like products and sentences are like baskets

Tools like `word2vec` and `glove` map from the J -dimensional discrete space to [two] vector *embeddings* in \mathbb{R}^S , and $S \ll J$

$$\max_{\mathbf{U}, \mathbf{V}, \mathbf{C}} \left\{ \sum_t \sum_{j,k} q_{tj} q_{tk} \mathbf{u}'_j \mathbf{v}_k - A(\mathbf{U}, \mathbf{V}, \mathbf{C}) \right\}$$

Product Embeddings



substitutes (synonyms) are close in the same vector space
complements (topical words) are close across vector spaces

Moving to a demand system (AID)

It's *almost* ideal:

$$s_t = \alpha + \Gamma \log(p_t) + \beta \log \frac{e_t}{\phi_t} + \varepsilon_t$$

s_{tj} is the **budget share** for product j in basket t and e_t is the budget
($e_t = \sum_j \$_{tj}$ and $s_{tj} = \$_{tj}/e_t$)

ϕ_t is the **translog price index** $\sum_j \log p_{tj} [\alpha_j + \sum_k \gamma_{jk} \log p_{tk}]$

But not so easy...

In grocery stores, we have 100,000s of products

Each week, at each store, for each product,
we will have different demand and elasticity and income effects.

$$\mathbb{E}s_{tj} = \alpha_{tj} + \sum_k \gamma_{tjk} \log p_{tj} + \beta_{tj} \frac{e_t}{\phi_t}$$

Featurize and Embed

Grab massive \mathbf{x}_{tj} that *hierarchically* encode products and transactions

$$\mathbb{E}s_{tj} = \mathbf{x}'_{tj}\boldsymbol{\alpha} + \mathbf{x}'_{tj}\boldsymbol{\delta} \log p_{tj} + \mathbf{x}'_{tj}\boldsymbol{\beta} \log \left(\frac{e_t}{\phi_t(\mathbf{x}_{tj})} \right) + \boldsymbol{\Gamma} \log \mathbf{p}_t$$

and use a symmetric square matrix factorization $\boldsymbol{\Gamma} = \mathbf{U}\mathbf{V}' + \mathbf{V}\mathbf{U}'$

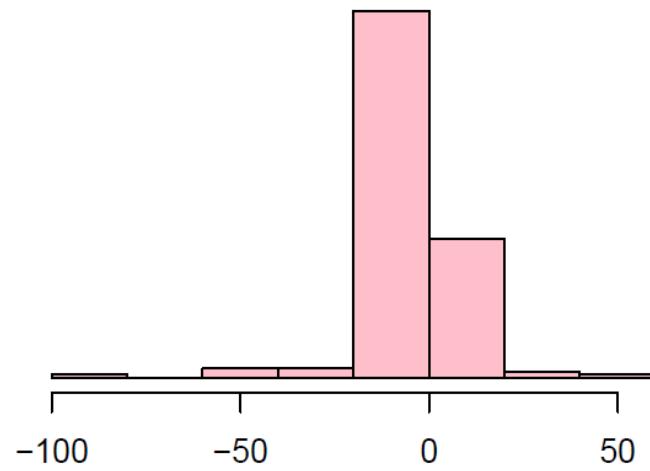
Note we are using the actual trans-log and solving a bilinear system

Beer Store own-price (compensated) elasticities

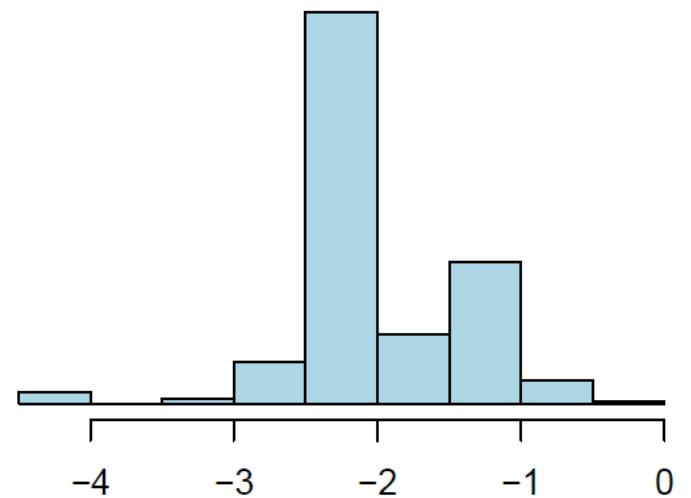
one shared: $x_{tj} = 1$

$$\frac{dq}{dp} \frac{p}{q} = -0.23$$

brand-specific: $x_{tjk} = \mathbb{1}_{[k=j]}$

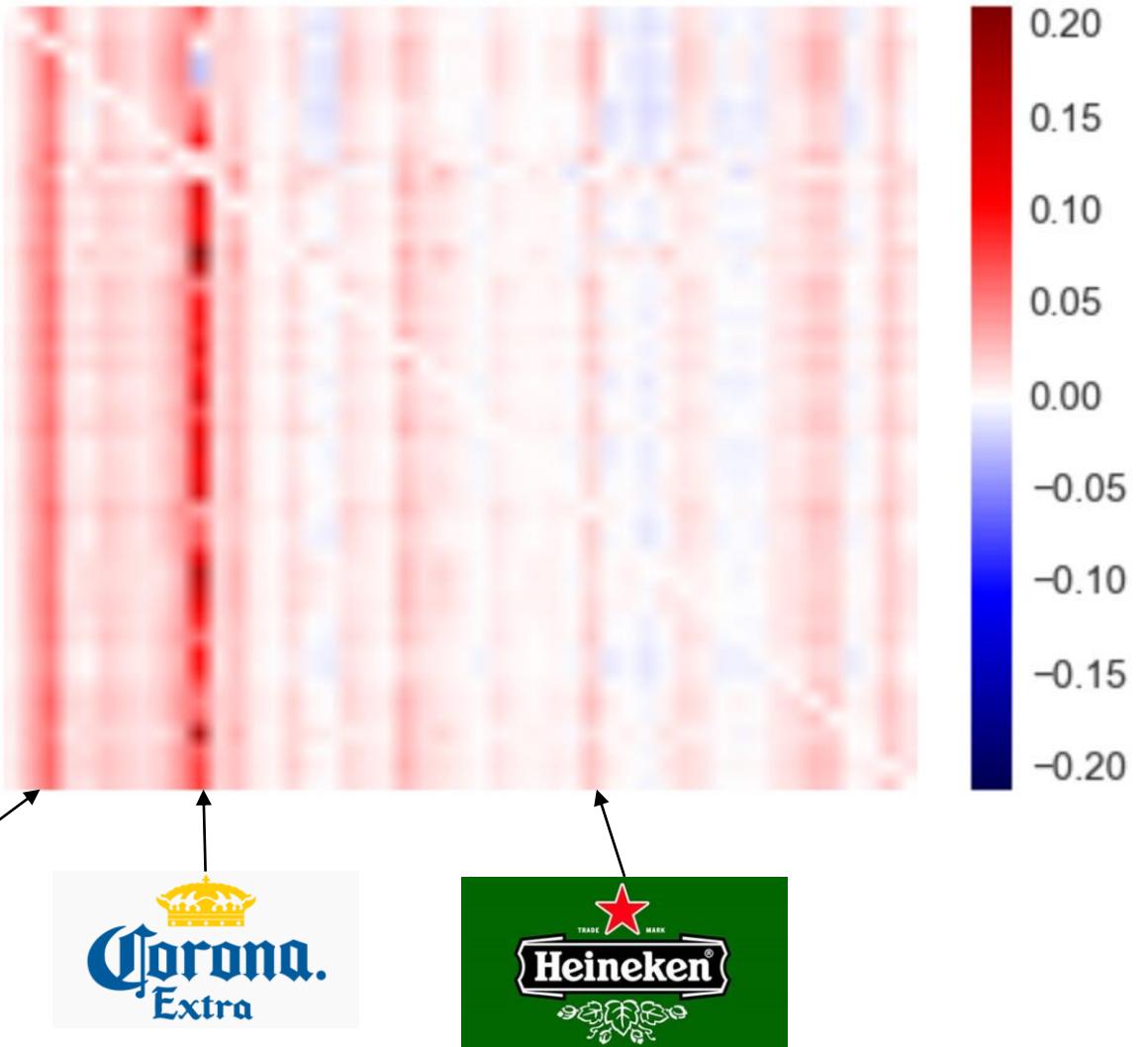


x_{tj} =featurized description

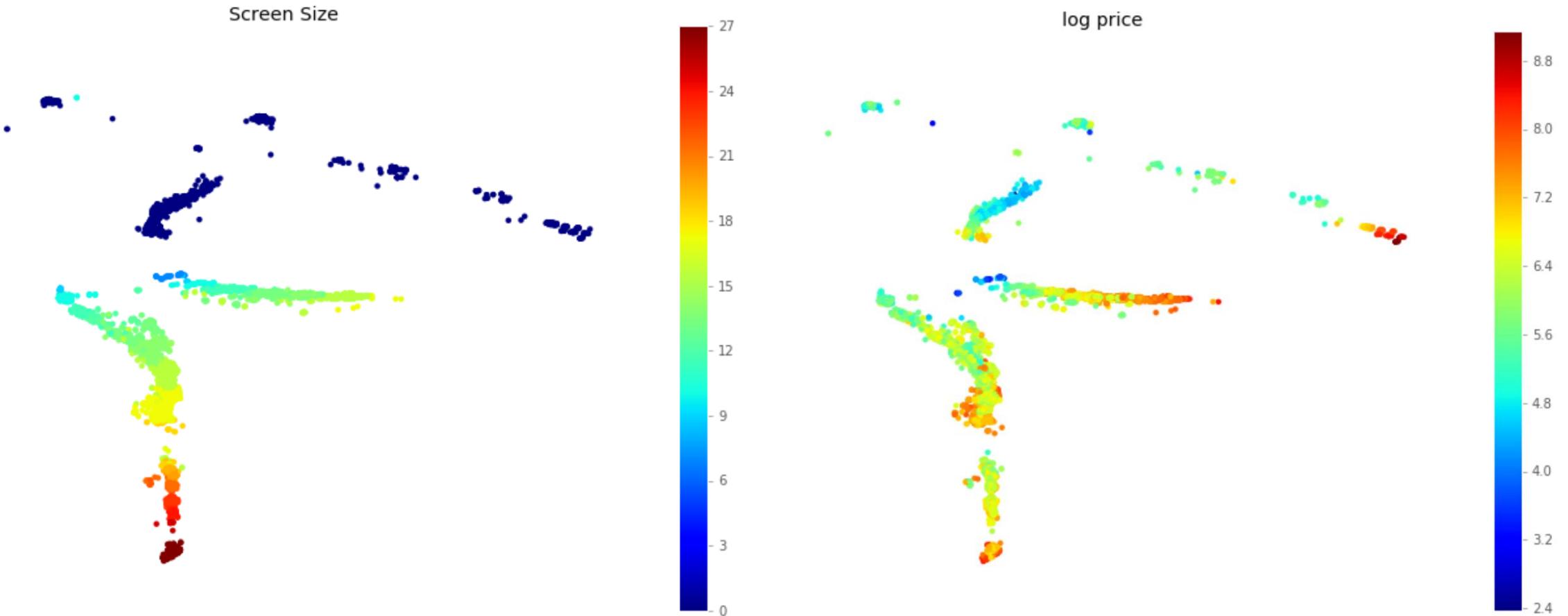


Beer Store cross-product compensated elasticities

Elasticity matrix (omitting diagonal)



Deep Hedonics



But wait... it's still a system

Recall: where does price come from?

$$\log p_{tj} = \varphi_{tj} + \psi_j q_{tj}^* + \nu_{tj}$$

and the *demand system* is in equilibrium when $q_{tj}^* = q_{tj}$

This equilibrium introduces ‘price endogeneity’: $\mathbb{E}[p_{tj} e_{tj}] \neq 0$

Endogenous Errors

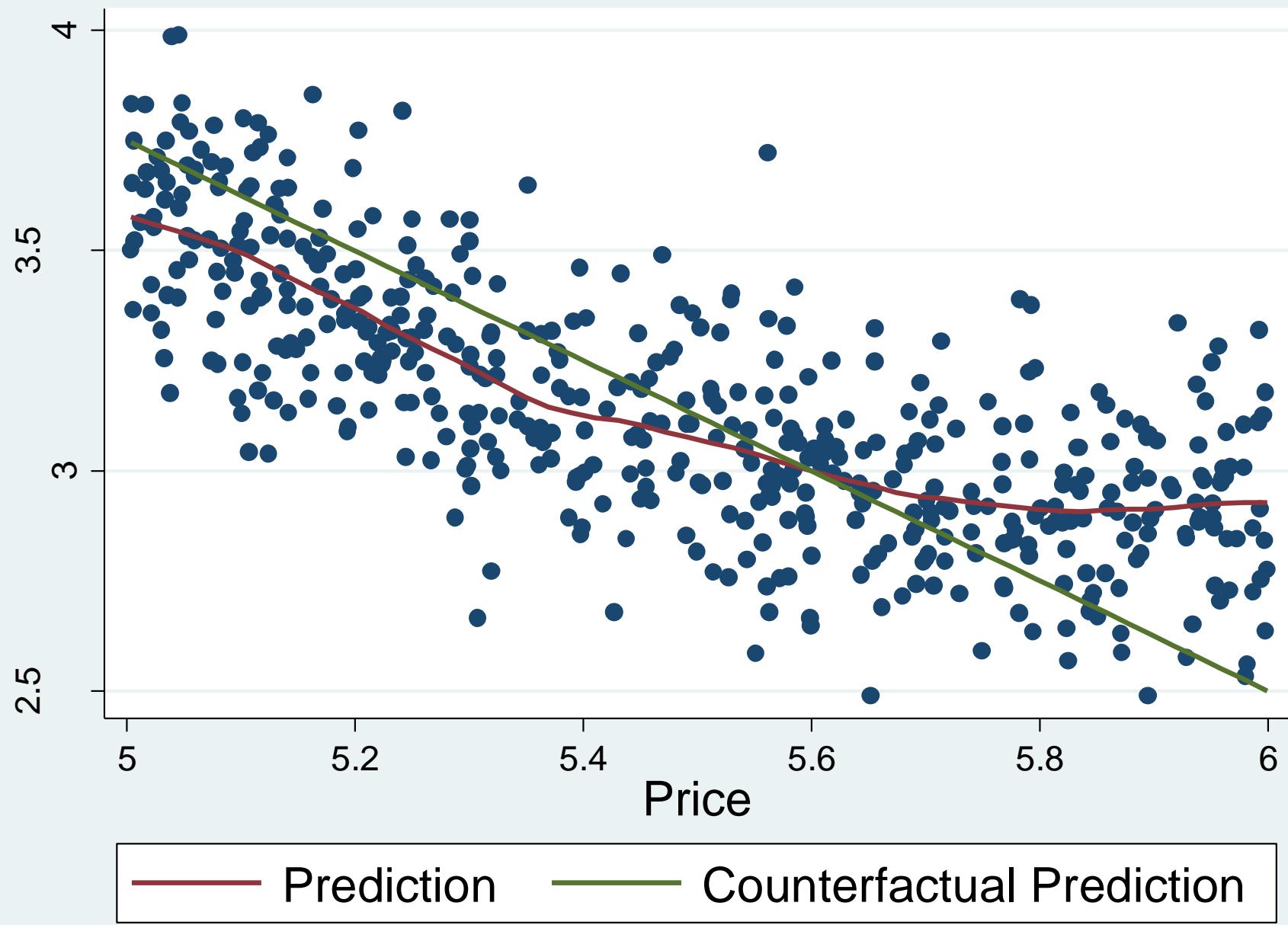
$$y = g(p, \mathbf{x}) + e \text{ and } \mathbb{E}[p e] \neq 0$$

If you estimate this using naïve ML, you'll get

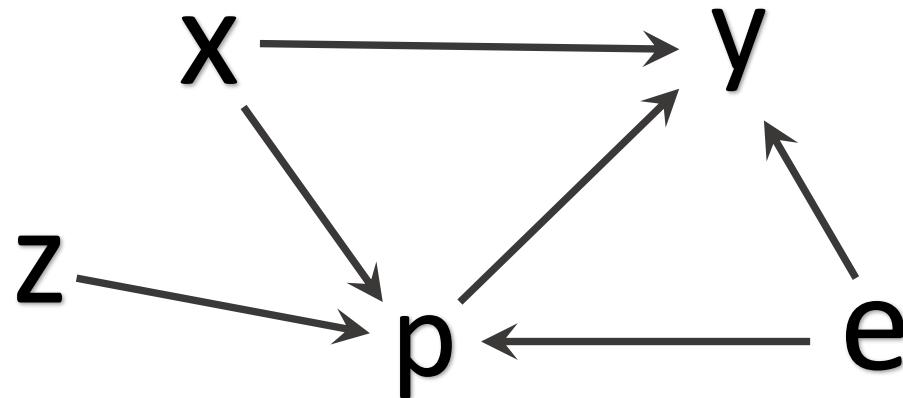
$$E[y|p, \mathbf{x}] = E_{e|p}[g(p, \mathbf{x}) + e] = g(p, \mathbf{x}) + E[e|p, \mathbf{x}]$$

This works for prediction. It doesn't work for counterfactual inference:

What happens if I change p independent of e ?



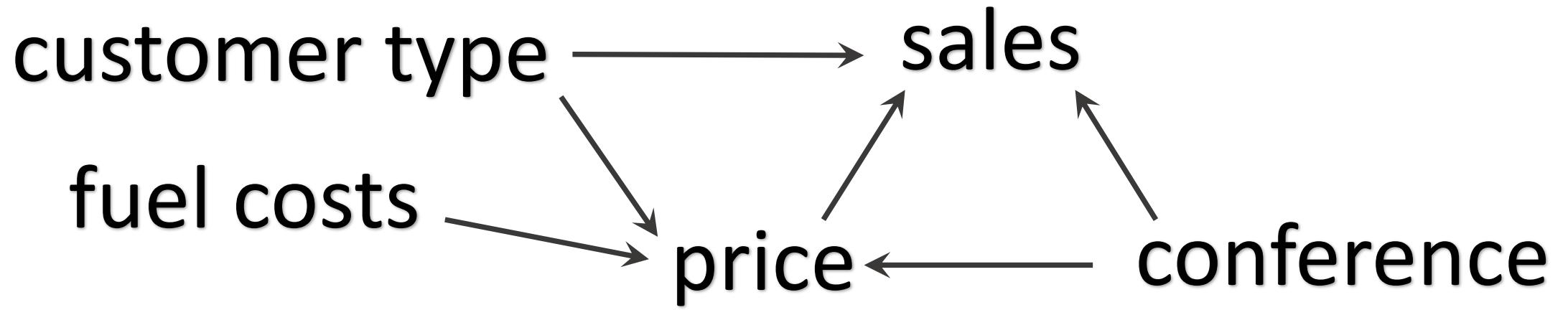
Instrumental Variables (IV)



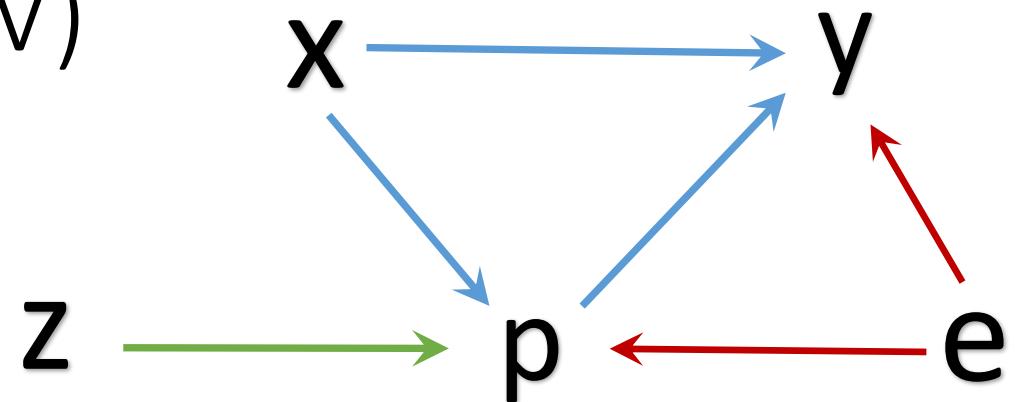
In IV we have a special $z \perp e$ that influences policy p but not response y .

- Supplier costs that move price independent of demand (e.g., fish, oil)
- Any source of treatment randomization (intent to treat, AB tests, lottery)

Instrumental Variables (IV)



Instrumental Variables (IV)



The *exclusion structure* implies

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

So to solve for *structural* $h(p, x)$ we have a new learning problem

$$\min_{g \in G} \sum \left(y_i - \int g(p, x_i)dF(p|x_i, z_i) \right)^2$$

$$\min_{g \in G} \sum \left(y_i - \int g(p, x_i) dP(p|x_i, z_i) \right)^2$$

2SLS:

$$p = \beta z + \nu \text{ and } g(p) = \tau p \text{ so that } \int g(p) dP(p|z) = \tau \hat{p} = \tau \hat{\beta} z$$

So you first regress p on z then regress y on \hat{p} to recover $\hat{\tau}$.

This requires strict assumptions and homogeneous treatment effects.

$$\min_{g \in G} \sum \left(y_i - \int g(p, x_i) dP(p|x_i, z_i) \right)^2$$

Or look to nonparametric 2SLS like in Newey and Powell:

$$g(p, x_i) \approx \sum_k \varphi_k(p, x_i) \text{ and } \varphi_k(p, x_i) \approx \sum_j \phi_{kj}(x_i, z_i)$$

But this requires careful crafting and will not scale with $\dim(x)$

$$\min_{g \in G} \sum \left(y_i - \int g(p, x_i) dF(p|x_i, z_i) \right)^2$$

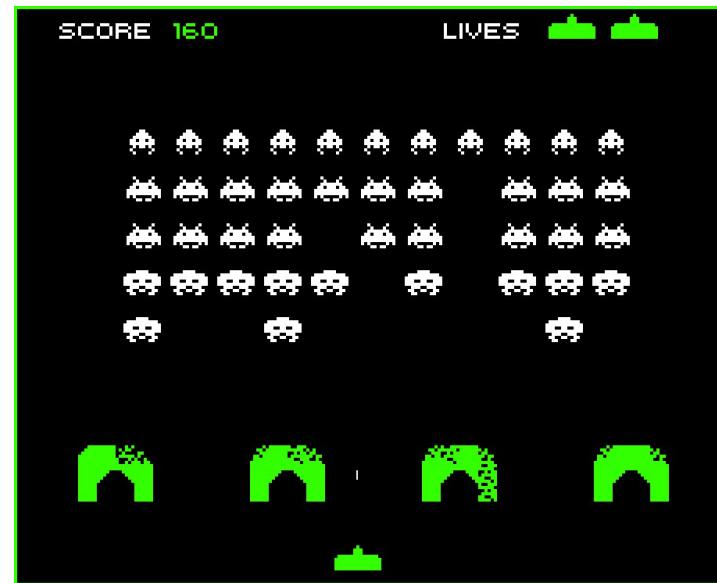
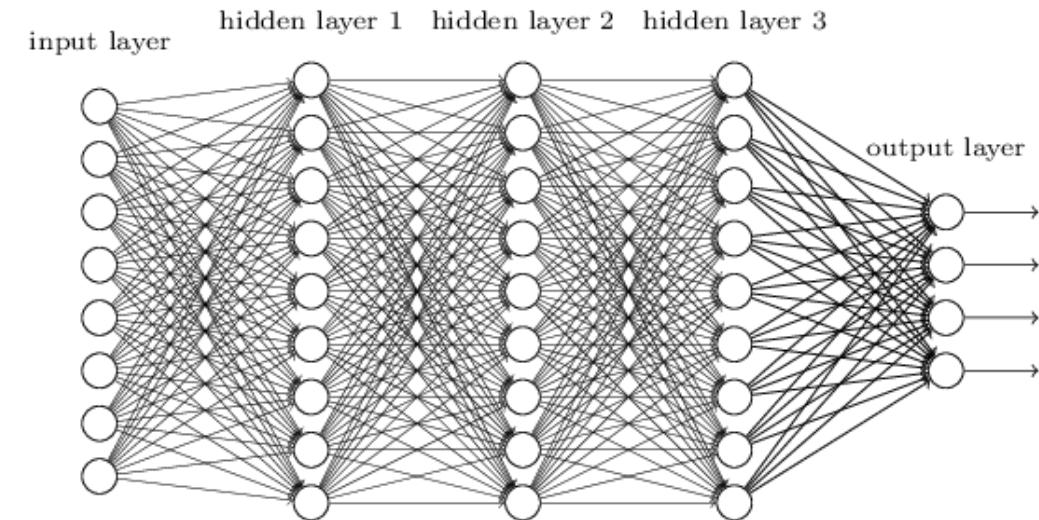
Instead, we propose to target the integral loss function directly

For discrete (or discretized) treatment

- Fit distributions $\hat{F}(p|x_i, z_i)$ with probability masses $\hat{f}(p_b|x_i, z_i)$
- Train \hat{g} to minimize $\left[y_i - \sum_b g(\hat{p}_b, x_i) \hat{f}(p_b|x_i, z_i) \right]^2$

And you've turned IV into two *generic* machine learning tasks

Learning to love Deep Nets



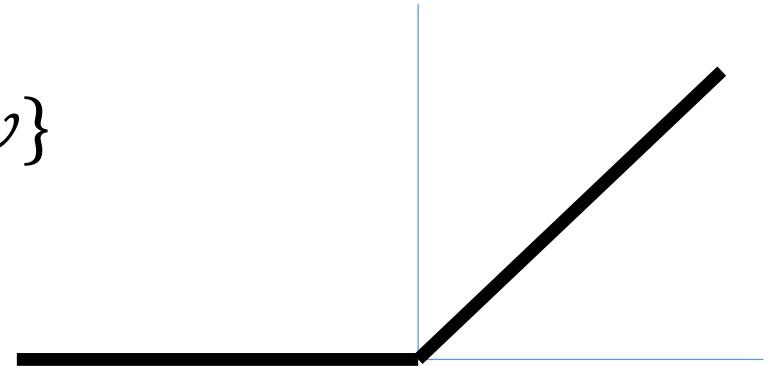
What is a deep net?

$$\hat{y}_i = \sum_k \eta_k(a_{ik}), \quad a_{ik} = \sum_j w_{kj} z_j, \quad z_j = \sum_l h_l^1(b_{il}), \dots$$

And so-on until you get down to a bottom layer $\{f_l(x_i)\}_l$

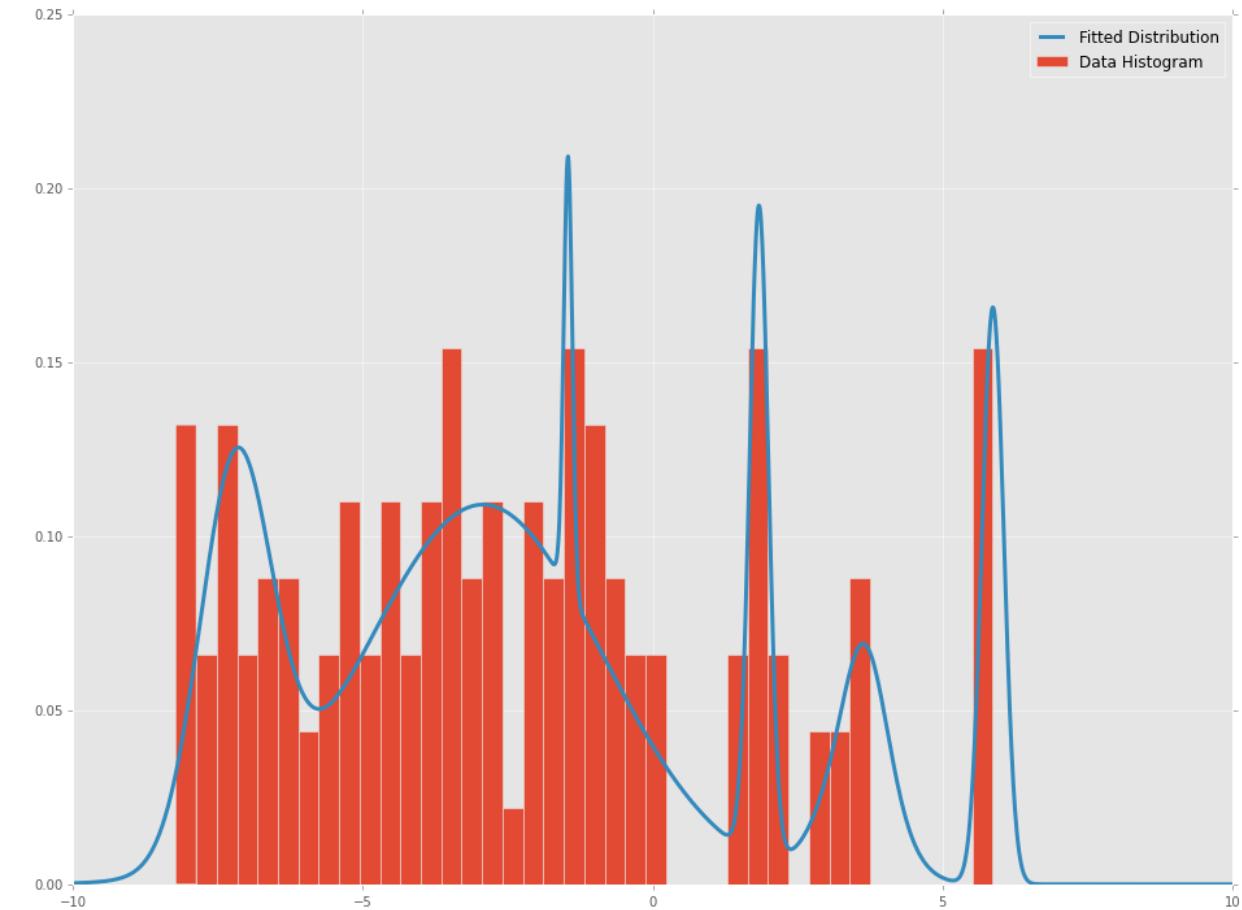
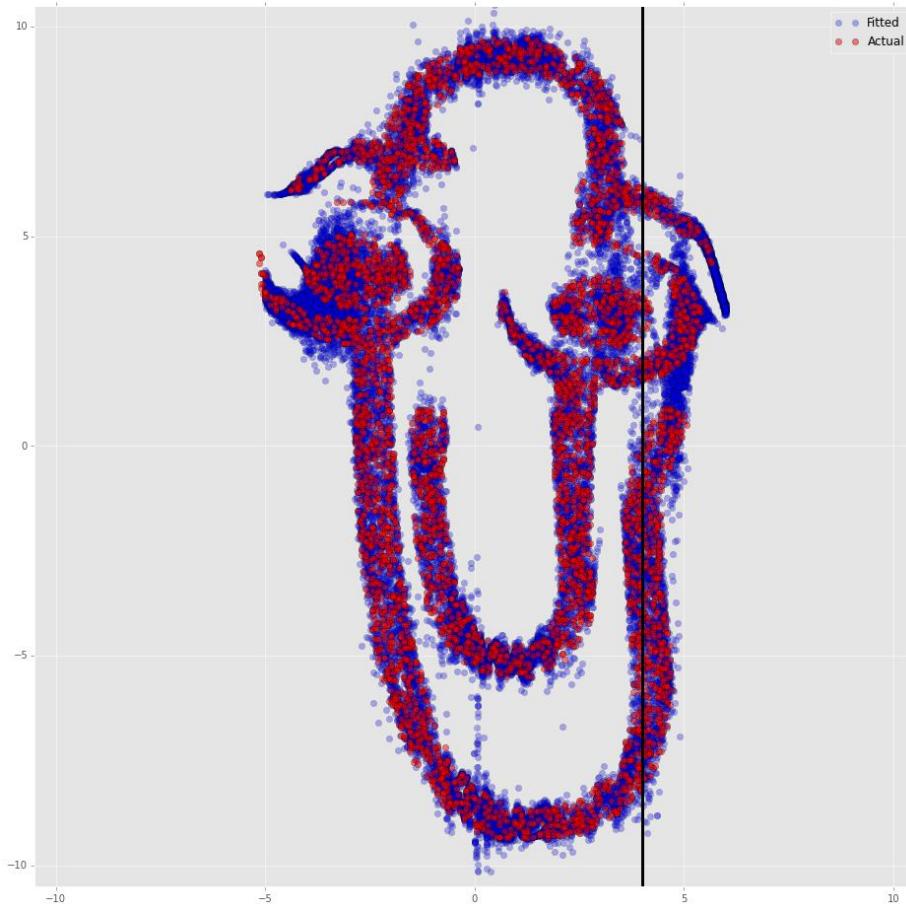
Many different variations here: recursive, convolutional, ...

Apart from the bottom, usually $h(v) = \max\{0, v\}$



e.g., first-stage learning for $F(p|x_i, z_i)$

Bishop 96: Final layer of network parametrizes a mixture of Gaussians



The second stage involves an integral loss function

If p is not discrete or can take many values, we can't just integrate

Brute force just samples from $\hat{F}(p|x_i, z_i)$ and you take gradients on

$$\frac{1}{N} \sum_i \left(y_i - \frac{1}{B} \sum_b g(\hat{p}_{ib}, x_i; \theta) \right)^2, \quad \hat{p}_{ib} \sim \hat{F}(p|x_i, z_i)$$

But this is inefficient

And more generally, it's inefficient even *without* the integral...

Stochastic Gradient Descent

You have loss $L(\mathbf{D}, \theta)$ where $\mathbf{D} = [\mathbf{d}_1 \dots \mathbf{d}_N]$

In the usual GD, you iteratively descend

$$\theta_t = \theta_{t-1} - \mathbf{C}_t \nabla L(\mathbf{D}, \theta_{t-1})$$

In SGD, you instead follow *noisy* but *unbiased* sample gradients

$$\theta_t = \theta_{t-1} - \mathbf{C}_t \nabla L(\{\mathbf{d}_{t_b}\}_{b=1}^B, \theta_{t-1})$$

Why SGD? You get what you need faster

6

Léon Bottou

Table 2. Asymptotic equivalents for various optimization algorithms: gradient descent (GD, eq. 2), second order gradient descent (2GD, eq. 3), stochastic gradient descent (SGD, eq. 4), and second order stochastic gradient descent (2SGD, eq. 5). Although they are the worst optimization algorithms, SGD and 2SGD achieve the fastest convergence speed on the expected risk. They differ only by constant factors not shown in this table, such as condition numbers and weight vector dimension.

	GD	2GD	SGD	2SGD
Time per iteration:	n	n	1	1
Iterations to accuracy ρ :	$\log \frac{1}{\rho}$	$\log \log \frac{1}{\rho}$	$\frac{1}{\rho}$	$\frac{1}{\rho}$
Time to accuracy ρ :	$n \log \frac{1}{\rho}$	$n \log \log \frac{1}{\rho}$	$\frac{1}{\rho}$	$\frac{1}{\rho}$
Time to excess error ε :	$\frac{1}{\varepsilon^{1/\alpha}} \log^2 \frac{1}{\varepsilon}$	$\frac{1}{\varepsilon^{1/\alpha}} \log \frac{1}{\varepsilon} \log \log \frac{1}{\varepsilon}$	$\frac{1}{\varepsilon}$	$\frac{1}{\varepsilon}$

$$\mathcal{E} = \mathcal{E}_{\text{app}} + \mathcal{E}_{\text{est}} + \mathcal{E}_{\text{opt}} \sim \mathcal{E}_{\text{app}} + \left(\frac{\log n}{n} \right)^\alpha + \rho$$

Learning rates

$$\theta_t = \theta_{t-1} - C_t \nabla L(\{d_{t_b}\}_{b=1}^B, \theta_{t-1}) = \theta_{t-1} - C_t \nabla L_t$$

What is C_t ? Hopefully not $c_t \mathbf{I}$...

Ideally, $C_t \rightarrow t^{-1} C$ and $C^{-1} = -\nabla \nabla L(D, \theta^*)$, but that's infeasible

ADAGRAD

(for convex loss)

$$C_t = \text{diag} \left(\sqrt{\sum_t \nabla L_{tj}^2} \right) \Rightarrow \sum_t (err(\theta_t) - err(\theta^*)) = O(\|\theta^*\| \text{tr}(C_t))$$

We use heuristic innovations on this (ADAM, momentum, etc)

SGD for integral loss functions

Our one-observation stochastic gradient is

$$\nabla L(d_i, \theta) = -2 \left(y_i - \int g_\theta(p, x_i) d\hat{F}(p|x_i, z_i) \right) \int g_\theta'(p, x_i) d\hat{F}(p|x_i, z_i)$$

Do SGD by pairing each observation with *two independent* treatment draws

$$\nabla \hat{L}(d_i, \theta) = -2(y_i - g_\theta(\hat{p}_{i1}, x_i)) g'_\theta(\hat{p}_{i2}, x_i), \quad \hat{p}_{ib} \sim \hat{F}(p|x_i, z_i)$$

So long as the draws are independent, $\mathbb{E} \nabla \hat{L}(d_i, \theta) = \mathbb{E} \nabla L(d_i, \theta) = L(\mathbf{D}, \theta)$

Lower variance grad \Rightarrow faster/better convergence

e.g., in most architectures you want to fit F and g in two-stages

And with MC integration, consider:

$$\nabla \hat{L}(\{\mathbf{d}_b\}_{b=1}^B, \theta) = \sum_{b=1}^B (y_b - g_\theta(\hat{p}_{b1}, x_b)) g'_\theta(\hat{p}_{b2}, x_b)$$

$$\nabla \tilde{L}^B(\mathbf{d}, \theta) = (y - B^{-1} \sum_{b=1}^B g_\theta(\hat{p}_{b1}, x)) B^{-1} \sum_{b=1}^B g'_\theta(\hat{p}_{b2}, x)$$

Both involve the same number of operations, but

$$\text{var} \nabla \hat{L}(\{\mathbf{d}_b\}_{b=1}^B, \theta) = \frac{\text{var} \nabla \hat{L}(\mathbf{d}, \theta)}{B} \quad \text{while} \quad \text{var} \nabla \tilde{L}^B(\mathbf{d}, \theta) \approx \frac{E[\text{var} \nabla \tilde{L}^1(\mathbf{d}, \theta)]}{B} + \text{var}(\nabla L(\mathbf{d}, \theta))$$

Aside: we can use SGD more in econ ...

There are a ton of setups where we use simulation to solve

$$\min_{\beta} \sum \left(y_i - \int g(x_i; \theta) dP(\theta | \beta) \right)^2$$

Random coefficient models, or wider class in Pakes and Pollard

Sampling SGD is a perfect fit here

Validation and model tuning

We can do *causal validation* via two OOS loss functions

Leave-out deviance on first stage

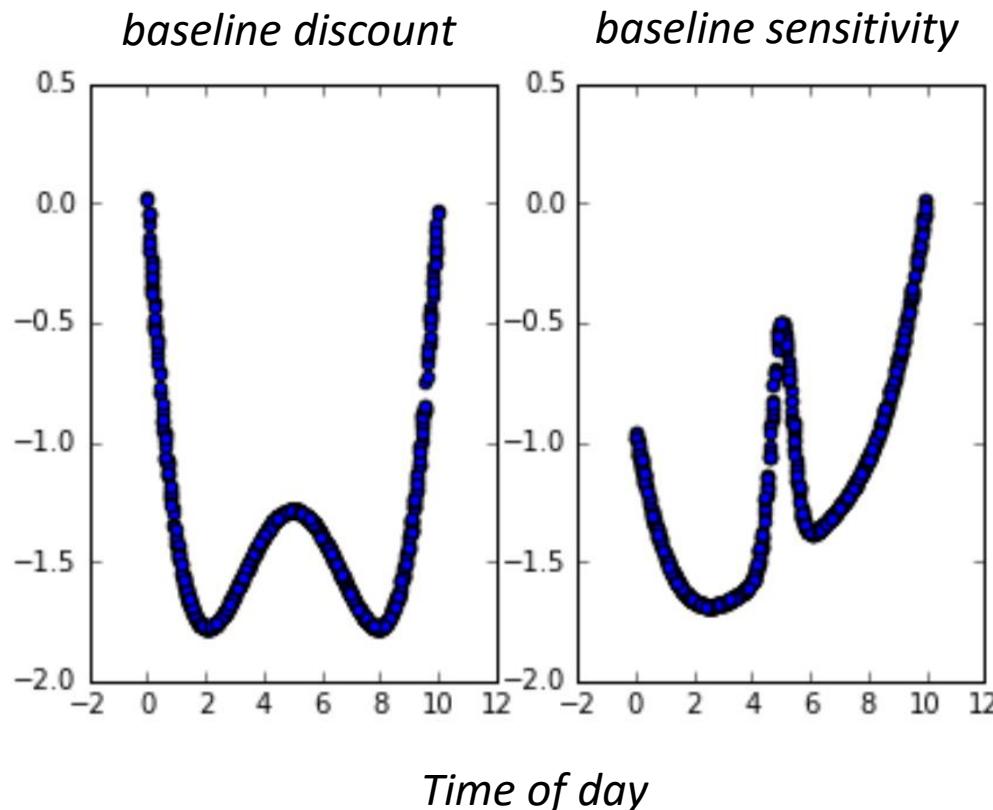
$$\sum_{i \in LO} -\log \hat{f}(p|x_i, z_i)$$

Leave-out loss on second stage (constrained fit on $\mathbb{E}[y|xz]$)

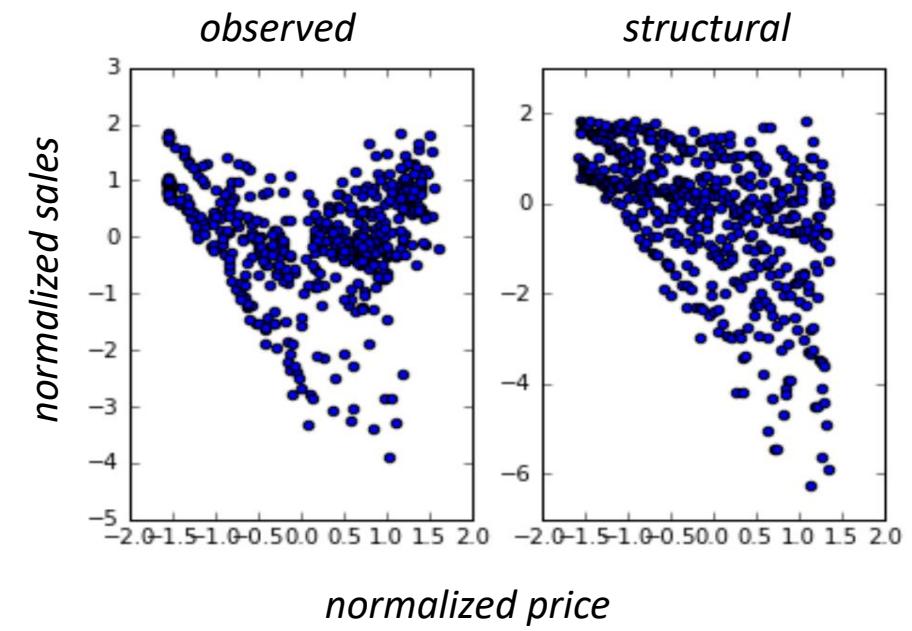
$$\sum_{i \in LO} (y_i - g_\theta(\dot{p}_i, x_i))^2, \quad \dot{p}_i \sim \hat{F}(p|x_i, z_i)$$

You want to minimize both of these (in order).

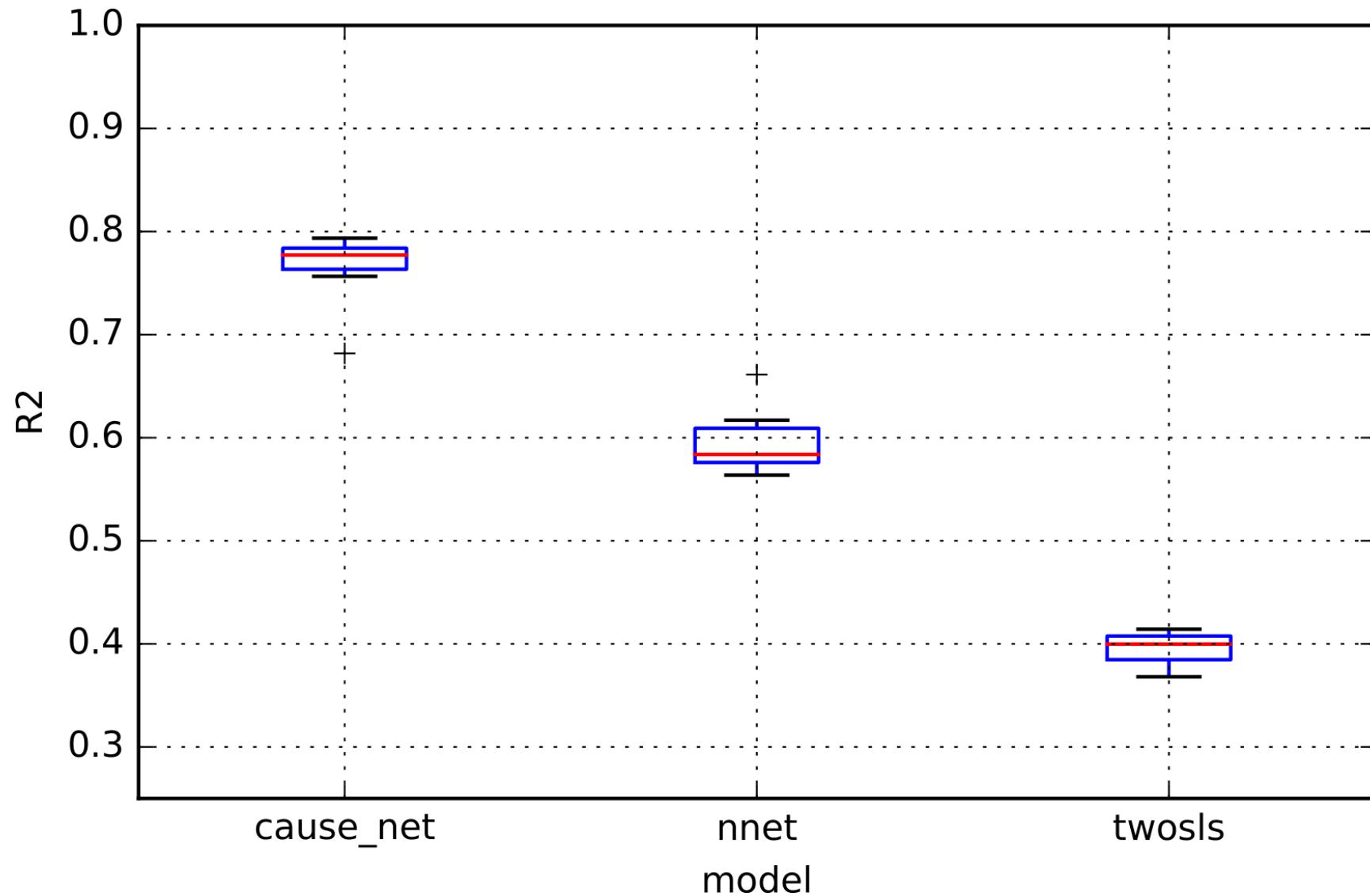
heterogeneous price effects



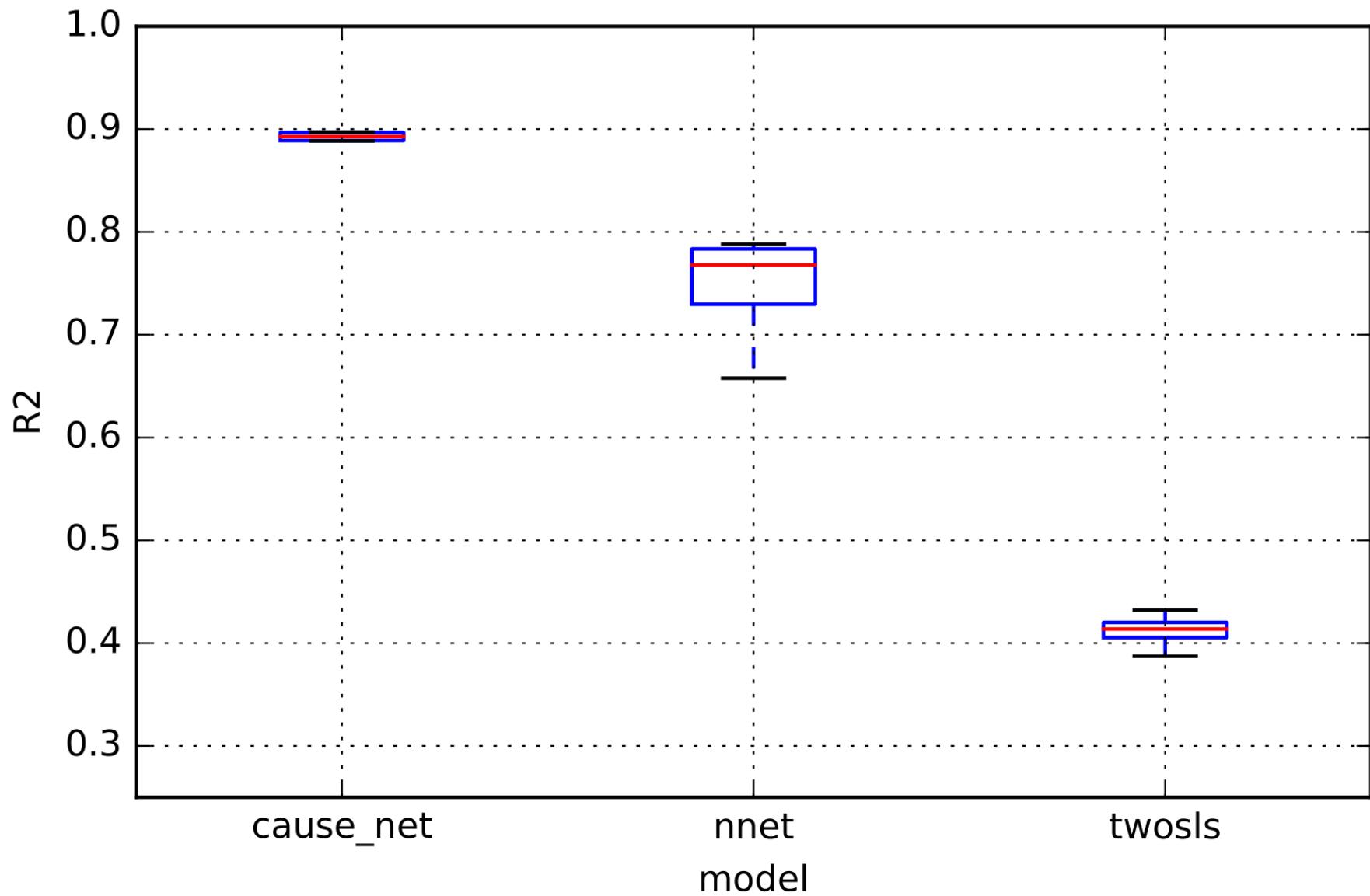
- ‘time’ dependent prices and elasticity
- 7 different customer types multiply discounts, sensitivities, and demand



$N=1000$



$N=5000$



Identification

The truth g_0 is identified if this equation has a unique solution. Let $T[g](x, z) = \int g(p, x)dF(p|x, z)$. Since T is a linear operator on the function space G , uniqueness of the solution is equivalent to the kernel of T being the singleton zero function:

$$\{g \in G : T(g) = 0\} = \{0\}$$

This can equivalently be stated as the “completeness condition” of Newey and Powell (2003): for all measurable g , we have:

$$E[g(t, x)|x, z] = 0 \quad \forall (x, z) \in \text{supp}(X, Z) \Rightarrow g(t, x) = 0$$

where the equivalence comes from the fact that the LHS function is just $T(g)$.

Consistency

Our is just a minimum distance estimator conditional upon draws from the 1st stage

Consistency has $\|\bar{g} - \bar{g}^*\| \rightarrow 0$ after averaging over the instrument distribution

The relevant theory is in Newey and Powell (03) and Chen and Pouzo (12)

Promise to grow both nets at rates $k(n) \rightarrow \infty$, $k/n \rightarrow 0$, then approximation

theorems for neural nets (e.g. Hornik 91) imply density around the truth

For treatment (and 1st stage) continuous, also need # mixture components to grow

Ads Application

Taken from Goldman and Rao (2014)

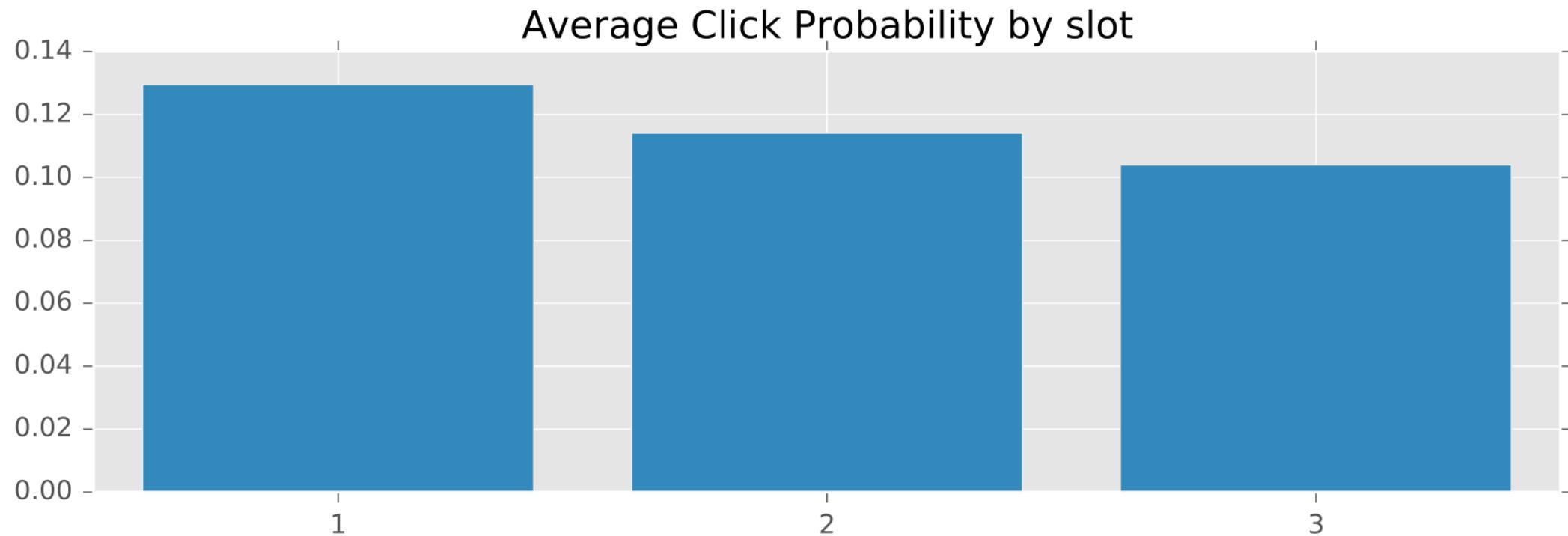
We have 74 mil click-rates over 4 hour increments for 10k search terms

Treatment: ad position 1-3

Instrument: background AB testing (bench of ~ 100 tests)

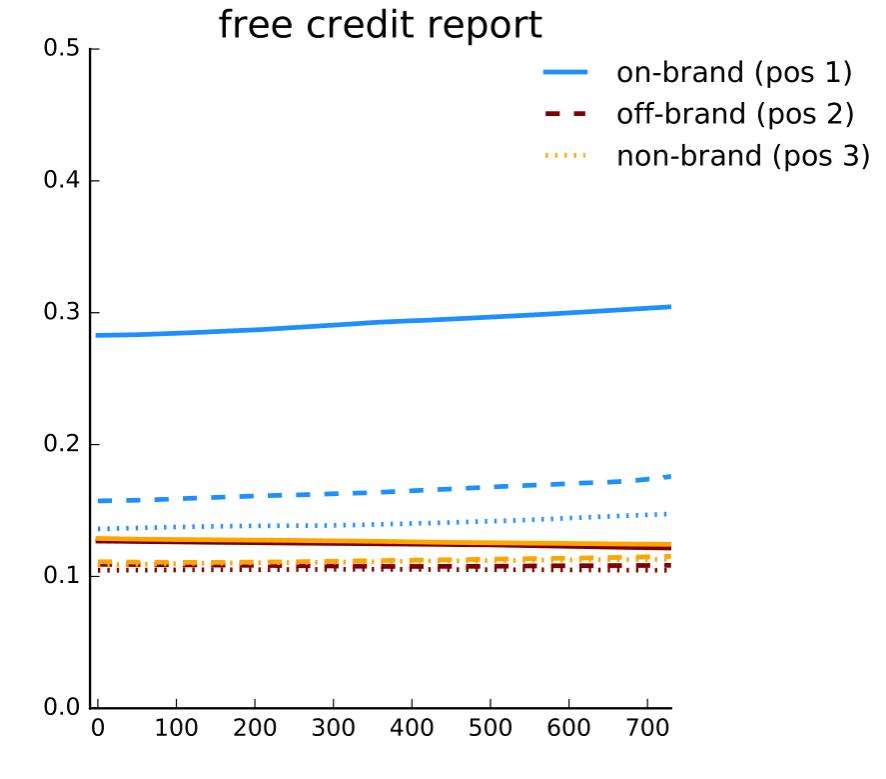
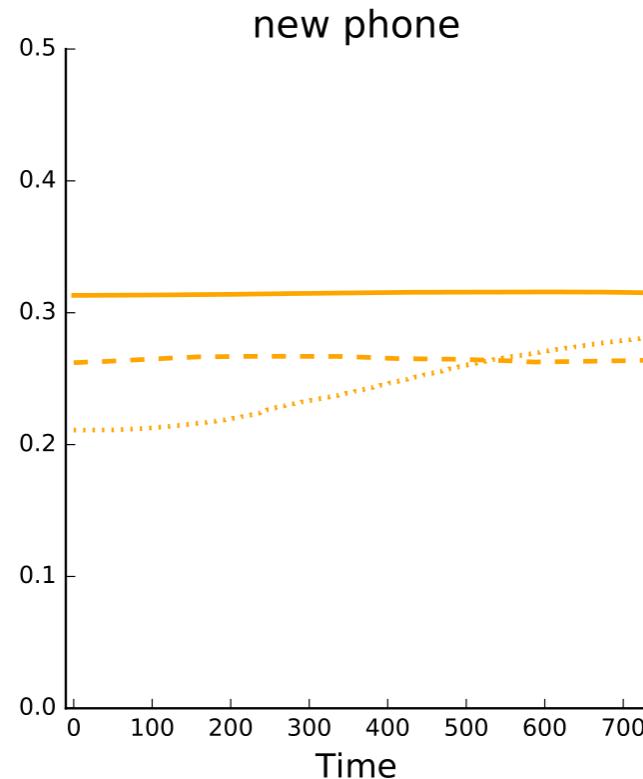
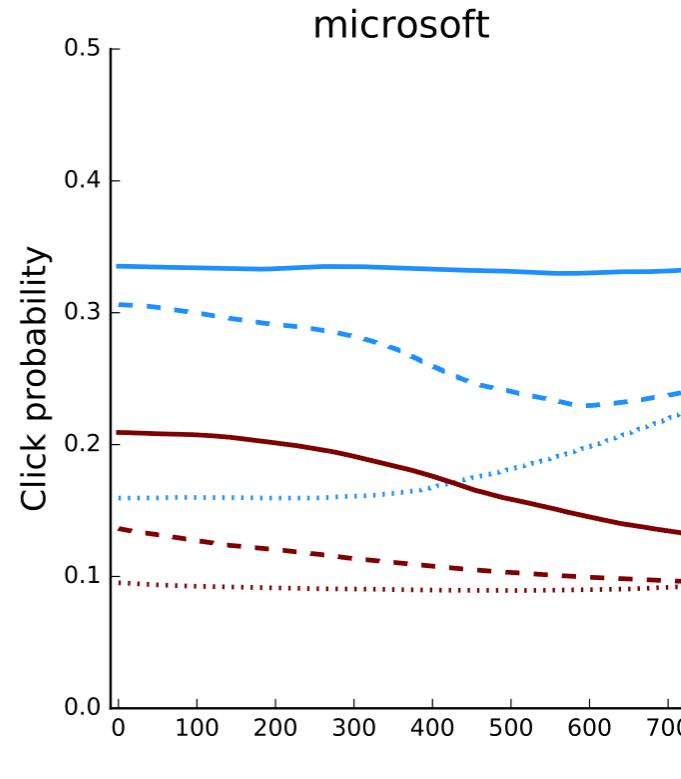
Covariates: advertiser id and ad properties, search text, time period

Average Treatment Effects



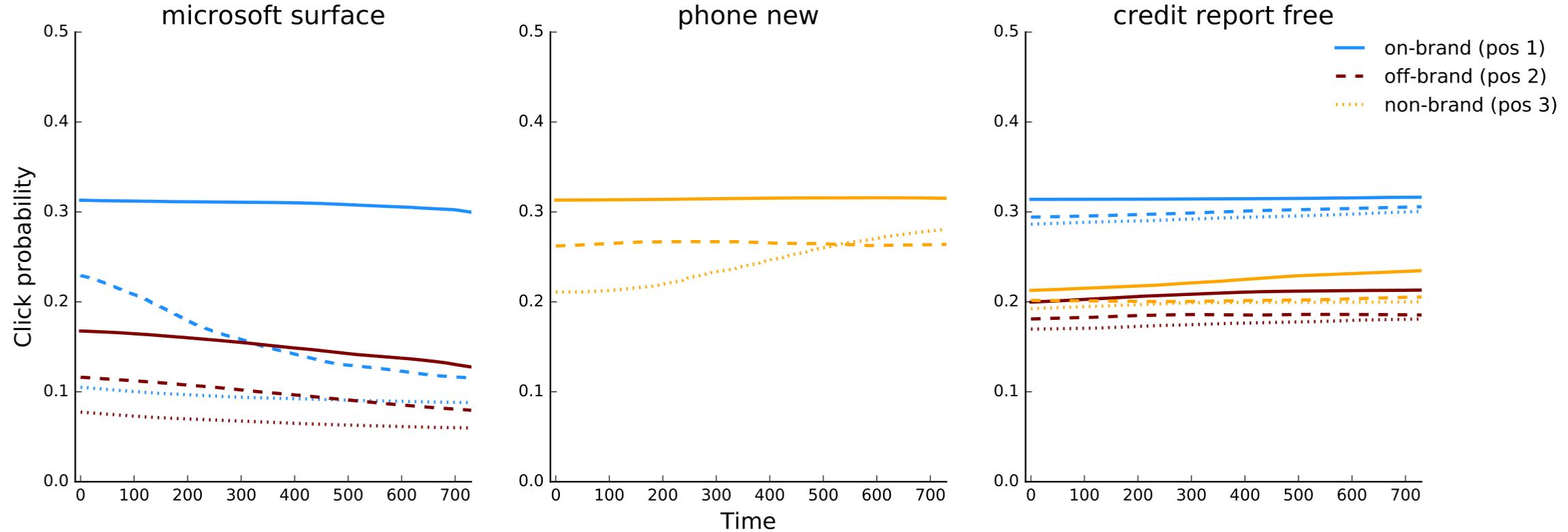
These compare to observed click probabilities of 0.33, 0.1, and 0.05.

Heterogeneous Treatment Effects



The *gaps* between ad-position lines can be interpreted *causally*.

Heterogeneous Treatment Effects



The *gaps* between ad-position lines can be interpreted *causally*.

Inference? Good question

For frequentist results we can do data splitting

Take the trained DNNs, and on *left-out* data do:

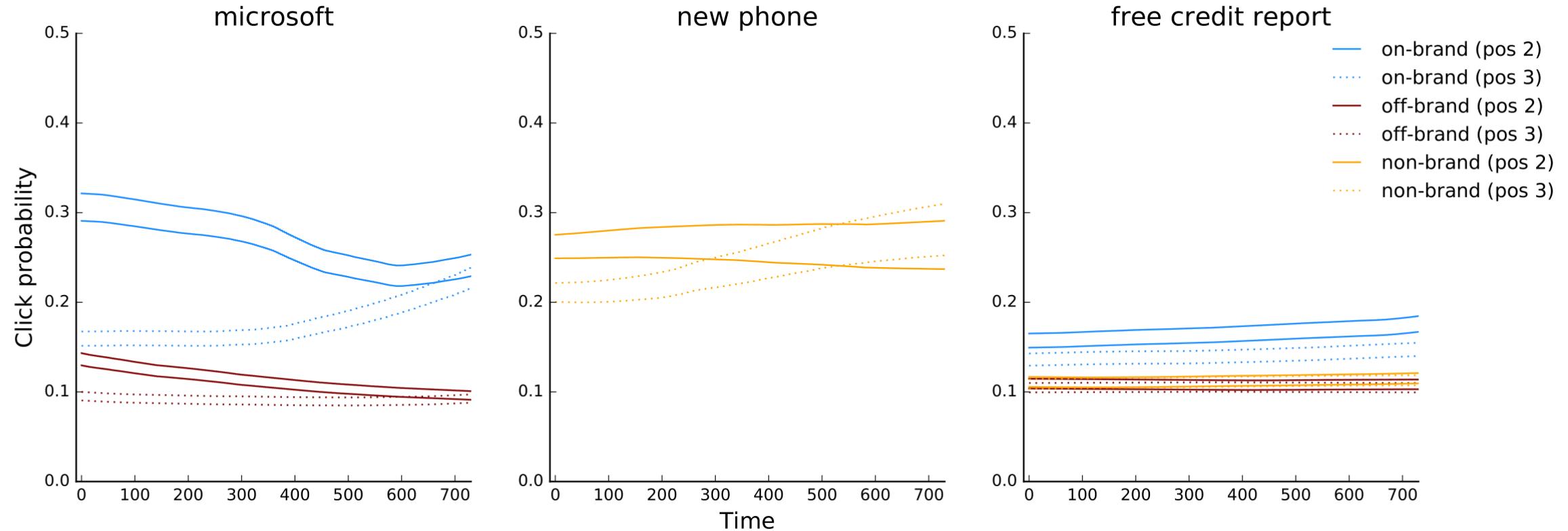
- Calculate conditional averages for each node in the top layer

$$\bar{\eta}_{ik} = E_{\hat{F}(\dot{p}|x_i, z_i)} \eta_k(x_i, \dot{p})$$

- Stack these as a design $\bar{\mathbf{H}} = [\bar{\boldsymbol{\eta}}_1 \cdots \bar{\boldsymbol{\eta}}_L]'$ and use OLS to fit $\mathbf{y} \approx \bar{\mathbf{H}}\boldsymbol{\beta}$
- Calculate *structural* residuals $e_i = y_i - \sum_k \hat{\beta}_k \eta_k(x_i, p_i)$
- Use the sandwich formula $\text{var}(\hat{\boldsymbol{\beta}}) = (\bar{\mathbf{H}}'\bar{\mathbf{H}})^{-1} \bar{\mathbf{H}}' \text{diag}(\mathbf{e}) \bar{\mathbf{H}} (\bar{\mathbf{H}}'\bar{\mathbf{H}})^{-1}$

$$\text{Then, e.g., } \text{var}(\hat{y}(x, p)) = \eta_k(x, p) \text{var}(\hat{\boldsymbol{\beta}}) \eta_k'(x, p)$$

Heterogeneous Treatment Effects



+/- 2 post selection SD

Inference? Good question

Or Bayes: recall the deep net

$$\hat{y}_i = \sum_k \eta_k(a_{ik}), \quad a_{ik} = \sum_j w_{kj} z_j, \quad z_j = \sum_l h_l^1(b_{il}), \dots$$

When training with SGD, we actually use **dropout** for regularization

At each update, calculate gradients against $\tilde{\theta}$ where

$$\tilde{w}_{kj} = w_{kj} \delta_{kj}, \quad \delta_{kj} \sim \text{Bern}(\pi)$$

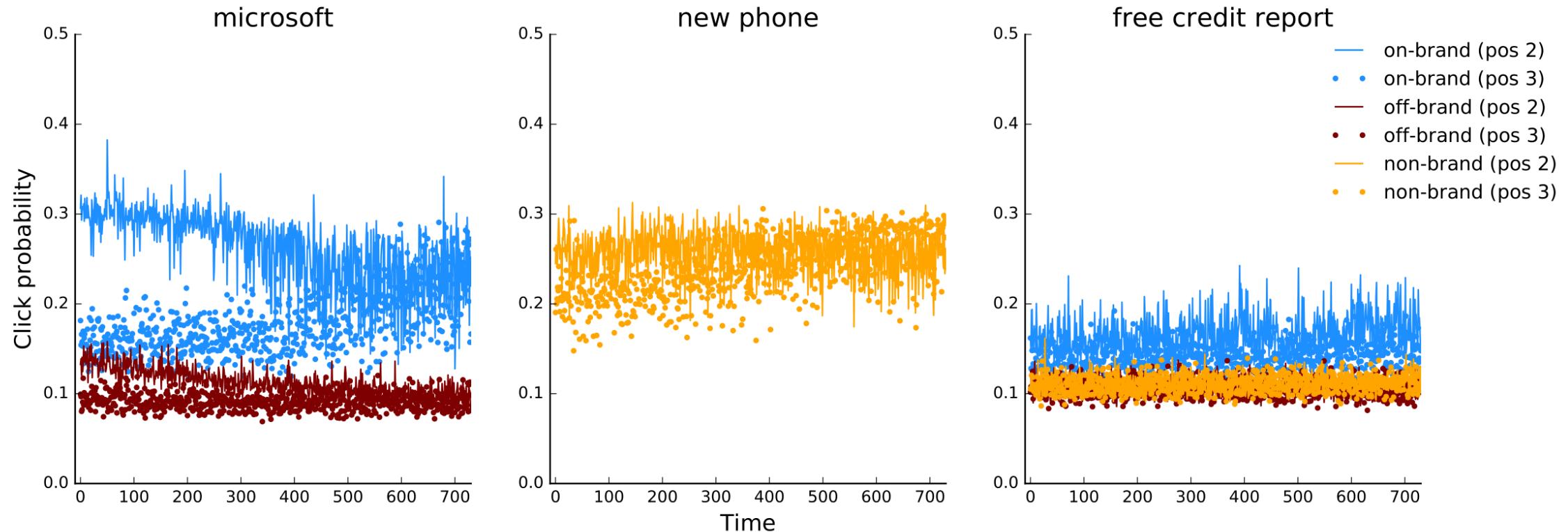
Bayesian inference via dropout

Variational Bayes minimizes

$$KL(q \mid p) = \mathbb{E}_q[\log q(y|x)] - \mathbb{E}[\log p(y|x, D)]$$

Gal and Ghahramani show that *dropout SGD* is optimizing variational distribution q , with uncertainty parametrized by the Bernoulli weights

Posterior Treatment Effect Samples



Each point/dash is an independent draw from the 'posterior'

Economics and Artificial Intelligence

We have a track record pointing ML at questions of science + causation.

We're going to replicate this success *at scale on unstructured data*

We use economic theory to build systems of tasks that can be addressed with deep nets and other state-of-the-art ML.

This is the construction of systems for *Economic Artificial Intelligence*.