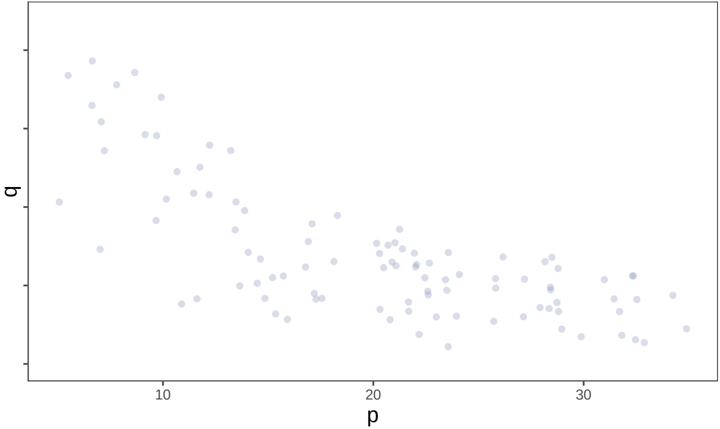
Structural Regularization: Transfer Learning with Theory

Jiaming Mao

Xiamen University

DSE 2025 HKU

Motivation



Example: Consumption Data

Motivation

The data is generated by the following model:

$$\max u_i\left(q_i, q_i^o\right) \text{ subject to } p_i q_i + p_i^o q_i^o \leq I,$$

where o stands for an outside good, I is income, and

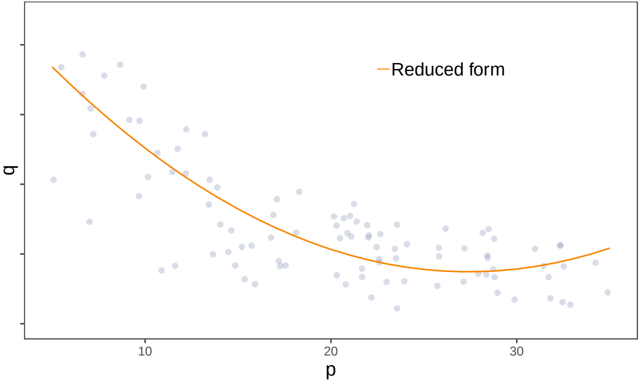
$$u_i\left(q_i, q_i^o\right) = \left[\alpha_i q_i^\rho + \left(1-\alpha_i\right)\left(q_i^o\right)^\rho\right]^{\frac{1}{\rho}},$$

with $\rho = -0.5$, implying an elasticity of substitution $\sigma = 0.67$.

Motivation

We can fit a statistical model to the data:

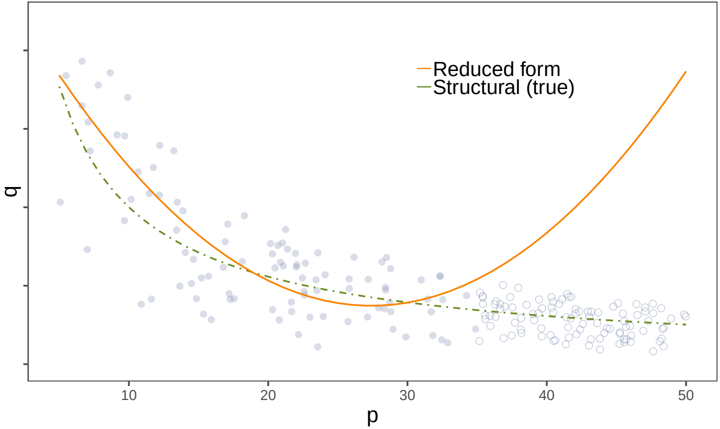
$$q_i = \beta_0 + \beta_1 p_i + \beta_2 p_i^2 + e_i$$



Motivation

- Assuming changes in p are exogenous, the statistical model represents a reduced-form estimate of the demand curve.
- However, the model does not look good when we extrapolate outside the observed range of data ...

Motivation



Motivation

- A major assumption for standard statistical and machine learning models is that the training and the test data are drawn from the same distribution (**domain**).
- Transfer learning problem: applying a model trained on a source domain to a *different* target domain.

Motivation

In contrast to the statistical approach, **structural estimation** estimates a structural model that is believed to describe the mechanism that generates the observed data.

Motivation

Structural models are causal models based on economic theory.

- Describe economic and social phenomena as outcomes of individual behavior in specific economic and social environments.
- More generally: any models that use economic theory to specify the functional form of causal relationships.

Motivation

- Because structural models are based on economic theory, their parameters can be *deep*, i.e., domain-invariant.
- A structural model estimated from a source domain can apply to a different target domain—such as in **counterfactual analyses**—and thus has generalizability or external validity.
- However, validity hinges on the model being correctly specified. In practice, structural models (and economic theory) are often highly stylized, relying on strong, unrealistic assumptions and identification by functional form.

Competing Paradigms

- Disagreement over the role of theory in learning:
    - Wolpin (2013), "The Limits of Inference *without* Theory"
    - Rust (2014), "The Limits of Inference *with* Theory"
- Today, "the division between structural and reduced-form methods has split the economics profession into two camps whose research programs have evolved almost independently despite focusing on similar questions." (Chetty 2009)

Competing Paradigms

- Rust (2014): "Notice the huge difference in world views. The primary concern of Leamer, Manski, Pischke, and Angrist is that we rely too much on assumptions that *could be wrong*, and which could result in incorrect empirical conclusions and policy decisions. Wolpin argues that assumptions and models *could be right*, or at least they may provide reasonable first approximations to reality ... Looking back nearly four decades after the Lucas critique paper, it is fair to ask whether structural models really have succeeded and resulted in significantly more accurate and reliable policy forecasting and evaluation."

This Paper

Is there any way to combine statistical and structural approaches to achieve more *external validity* than (reduced-form) statistical models, while being more *robust* against misspecification than structural models?

This Paper

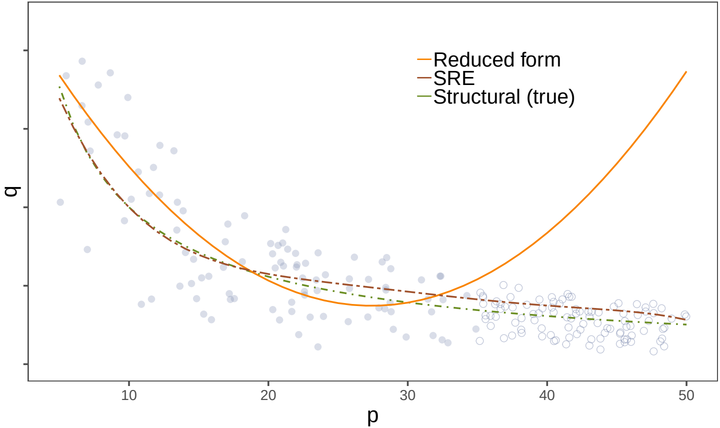
Structural Regularization Estimator (SRE)

Treat a given structural model as the benchmark model and estimate a flexible statistical model with a penalty on deviance from the structural benchmark.

Intuition

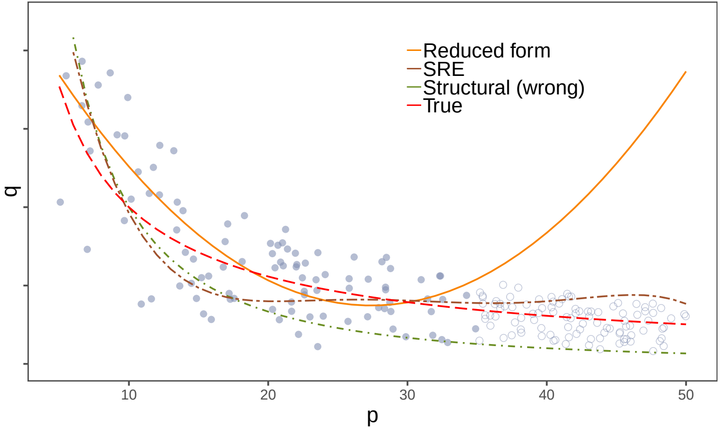
- Theory as *guide* rather than *truth*.
- Theory as prior knowledge.

This Paper



This Paper

What if we have an incorrect structural model that assumes $\rho = 0.5$ in the CES consumer utility function?



Literature

- **Transfer learning with covariate shift:** Zadrozny(2004), Huang et al.(2007), Tibishirani and Barber(2019).
  - focuses on $\mathbb{P}_x^T \neq \mathbb{P}_x^S$, but $\mathbb{P}_{y|x}^T = \mathbb{P}_{y|x}^S$.
- **Combining statistical and structural estimation:** Chetty (2009), Heckman (2010).
- **Robust structural estimation:** Bonhomme and Weidner (2018), Christensen and Connault (2019).

Literature

- **Transfer performance of economic theory:** Bo and Galiani (2021), Andrews et al. (2022).
- **Regularization with theory:** Fessler and Kasy (2019 REStat).
  - Theory as parameter constraints on statistical models.
  - No cross domain, no out-of-range extrapolation. Theoretical restrictions improve in-domain efficiency.

STRUCTURAL REGULARIZATION

Setup

- Given variables $(x, y) \in \mathcal{O}$ with joint distribution $\mathbb{P}_{xy}$. Our goal is to learn a target function $f(x) = \mathbb{E}_{\mathbb{P}_{y|x}}[y|x]$.
- Observed data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, $(x_i, y_i) \in \mathcal{O}' \subseteq \mathcal{O}$, with data-generating probability distribution $\mathbb{P}'_{xy} = \mathbb{P}_{xy}\big|_{(x,y)\in\mathcal{O}'}$.
    - Statistical/ML algorithms built on the random sampling assumption will have difficulty learning f from $\mathcal{D}$.
- Let $\mathcal{M}(x, y, u)$ be a structural model that describes the causal mechanisms that generate $\mathbb{P}_{xy}$, based on our prior knowledge.

Structural Regularization

First Stage

- Estimate $\mathcal{M}$ on $\mathcal{D}$ to obtain $\widehat{\mathcal{M}}$.
- Use $\widehat{\mathcal{M}}$ to generate *synthetic* data $\mathcal{D}^{\mathcal{M}} = \left\{ \left(x_i^{\mathcal{M}}, y_i^{\mathcal{M}} \right) \right\}_{i=1}^{M}$, $\left(x_i^{\mathcal{M}}, y_i^{\mathcal{M}} \right) \in \mathcal{O}$.
- We can also compute $f^{\mathcal{M}}(x) = \mathbb{E}^{\mathcal{M}}[y|x]$—the *implied* conditional expectation of y according to $\widehat{\mathcal{M}}$.
- Structural models are generative models, from which discriminative relationships can be derived (Bishop and Lasserre, 2007).

Structural Regularization

Second Stage

Estimate a flexible statistical model $g\left(x;\theta\right)$ by seeking solution to the following problem:

$$\min_{\theta\in\Theta}\left\{\sum_{i=1}^{N}\left(y_i-g\left(x_i;\theta\right)\right)^2+\lambda\cdot\Omega\left(\theta,\widehat{\theta}^{\mathcal{M}}\right)\right\}$$

- $\Omega\left(\theta,\widehat{\theta}^{\mathcal{M}}\right)$: regularizer that penalizes the distance between θ and $\widehat{\theta}^{\mathcal{M}}$.
- λ controls the tradeoff between goodness of fit and deviance from theory.

Structural Regularization

Second Stage

$\widehat{\theta}^{\mathcal{M}}$ is obtained by fitting g to $\mathcal{D}^{\mathcal{M}}$:

$$\begin{aligned}\widehat{\theta}^{\mathcal{M}} &= \underset{\theta \in \Theta}{\arg\min} \sum_{i=1}^{M} \left(y_i^{\mathcal{M}} - g\left(x_i^{\mathcal{M}}; \theta\right)\right)^2 \\ &= \underset{\theta \in \Theta}{\arg\min} \sum_{i=1}^{M} \left(f^{\mathcal{M}}\left(x_i^{\mathcal{M}}\right) - g\left(x_i^{\mathcal{M}}; \theta\right)\right)^2\end{aligned}$$

- $g\left(x; \widehat{\theta}^{\mathcal{M}}\right)$ represents a statistical approximation to $f^{\mathcal{M}}(x)$.
- $\widehat{\theta}^{\mathcal{M}}$ represents the "structural" or "theoretical" value of θ.

Bayesian Interpretation

- The solution to the SRE problem with ℓ_2 regularizers corresponds to the *maximum à posteriori* (*MAP*) estimates with a Gaussian prior centered around $\widehat{\beta}^{\mathcal{M}}$.

- Bayesian interpretation: theory as prior knowledge.

Bayesian Interpretation

- From a Bayesian perspective, regularization amounts to the use of *informative priors*, where we introduce our knowledge or belief about the target function in the form of priors, and use them to "regulate" the behavior of the hypothesis we choose.
- Theoretical models are formulated based on prior empirical evidence and should naturally serve as priors for analyzing new evidence.

Implementation

Algorithm 1 Structural Regularization with Sample-Splitting

Require: Observed data $\mathcal{D}$

1: samples $\{\mathcal{D}_1, \mathcal{D}_2\} \leftarrow$ Partition($\mathcal{D}, K = 2$)
2: **output** $\widehat{\theta} \leftarrow$ StructuralRegularization($\mathcal{D}_1, \mathcal{D}_2$)

The function Partition randomly partitions a sample into K equal sized parts. The function StructuralRegularization takes in two data samples and uses them to produce the SRE estimates as follows:

1: **procedure:** StructuralRegularization(sample $\mathcal{I}$, sample $\mathcal{J}$)
2: fit the structural model $\mathcal{M}$ on sample $\mathcal{I}$ to obtain $\widehat{\mathcal{M}}$
3: use $\widehat{\mathcal{M}}$ to generate a synthetic data set $\mathcal{D}^{\mathcal{M}}$
4: solve problem (4) on $\mathcal{D}^{\mathcal{M}}$ to obtain $\widehat{\theta}^{\mathcal{M}}$
5: substitute $\widehat{\theta}^{\mathcal{M}}$ into problem (3)
6: solve problem (3) on sample $\mathcal{J}$ for a grid of λ values and find the optimal λ^* by cross-validation
7: **return** $\widehat{\theta}$ as the solution to problem (3) on sample $\mathcal{J}$ at $\lambda = \lambda^*$

Implementation

Algorithm 2 Structural Regularization with Cross-Fitting

1: samples $\{\mathcal{D}_1, \mathcal{D}_2\} \leftarrow \textsc{Partition}(\mathcal{D}, K = 2)$
2: $\widehat{\theta}_1 \leftarrow \textsc{StructuralRegularization}(\mathcal{D}_1, \mathcal{D}_2)$
3: $\widehat{\theta}_2 \leftarrow \textsc{StructuralRegularization}(\mathcal{D}_2, \mathcal{D}_1)$
4: **output** $\widehat{\theta} \leftarrow \frac{1}{2}\left(\widehat{\theta}_1 + \widehat{\theta}_2\right)$

Causal Inference

Setup

- Observed variables: (y, d, v), where y is outcome, d is treatment, v is a set of covariates.
- We are interested in the ATE of d on y, denoted by τ. Allow τ to be fully nonlinear and heterogeneous, i.e. $\tau = \tau(d, v)$. Then

$$\tau(d, v) = \frac{\partial}{\partial d}\mathbb{E}\left[y^d \middle| v \right],$$

where y^d is the *potential outcome* of y under treatment d.

Causal Inference

Under the **unconfoundedness** assumption of $d \perp\!\!\!\perp y^d \big| v$,

$$\tau(d, v) = \frac{\partial}{\partial d}\mathbb{E}\left[y^d \middle| v\right] = \frac{\partial}{\partial d}\mathbb{E}\left[y \middle| d, v\right] = \frac{\partial}{\partial d}\mathbb{E}\left[y \middle| x\right],$$

where $x := (d, v)$.

- Under unconfoundedness, the causal inference problem reduces procedurally to a regression problem.

Causal Inference

Under **unmeasured confounding**, assume the availability of a valid instrument z such that θ is identified via moment conditions $\mathbb{E}\left[\left(y-g\left(x;\theta\right)\right)\middle| z\right]=0$.

Solve the following problem in the second stage:

$$\min_{\theta\in\Theta}\left\{\overline{m}\left(\theta\right)'W\overline{m}\left(\theta\right)+\lambda\cdot\Omega\left(\theta,\widehat{\theta}^{\mathcal{M}}\right)\right\},$$

where $\overline{m}\left(\theta\right)\doteq\frac{1}{N}\sum_{i=1}^{N}m_i\left(\theta\right)$, $m_i\left(\theta\right):=z_i\left(y_i-g\left(x_i;\theta\right)\right)$ are the moment functions and W is a weight matrix.

Experiments

1. First-price auction
2. Dynamic entry-exit
3. Demand Estimation with IV

First-Price Auction

Baseline Data Generating Model

- N risk-neutral bidders
- Independent private value $v_i \sim^{i.i.d} F(.)$
- Each bidder knows her own v_i and the common distribution $F(.)$, but not values of others.
- Nash equilibrium bidding strategy:

$$b\left(v\right) = v - \frac{1}{F\left(v\right)^{n-1}} \int_0^{v_i} F(x)^{n-1} dx$$

First-Price Auction

Target function: $\mathbb{E}\left[b_{\max}\middle| N\right]$

- Statistical model:

$$b_{\max} = g\left(N;\beta\right) + e$$

  - g : polynomial with degree selected by AIC.

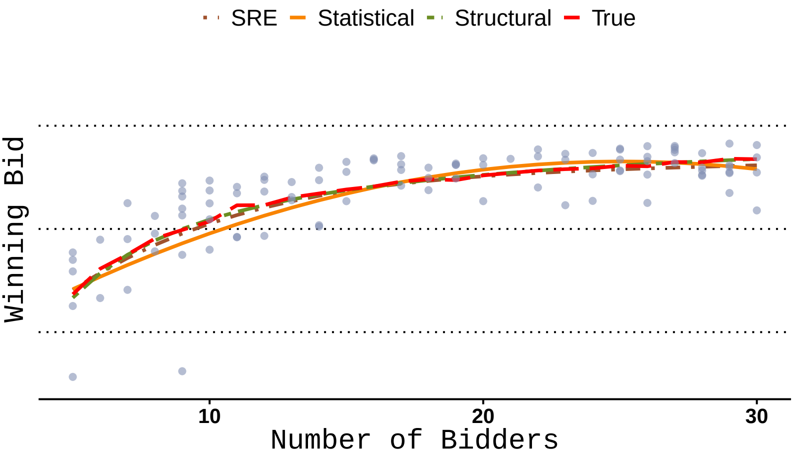
- Structural estimation: Guerre et al. (2000)

First-Price Auction

Experiment	True Mechanism	Structural Model
1	$v_i \overset{\text{i.i.d.}}{\sim} U(0,1), b_i = b(v_i)$	$v_i \overset{\text{i.i.d.}}{\sim} U(0,1)$ $b_i = b(v_i)$
2	$v_i \overset{\text{i.i.d.}}{\sim} \text{Beta}(2,5), b_i = b(v_i)$	
3	$v_i \overset{\text{i.i.d.}}{\sim} U(0,1), b_i = \eta_i \cdot b(v_i)$	

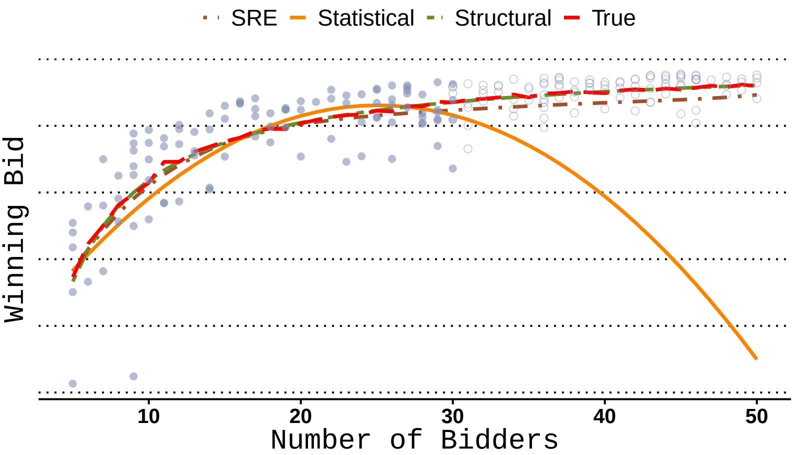
[a] $b(v_i)$ is the equilibrium bid function. $\eta_i \overset{\text{i.i.d.}}{\sim} \text{TN}(0, 0.25, 0, \infty)$.

First-Price Auction



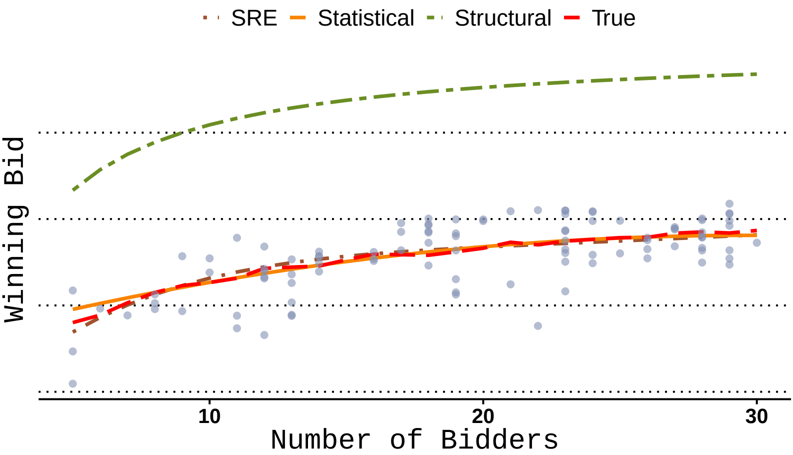
Experiment 1: in-domain

First-Price Auction



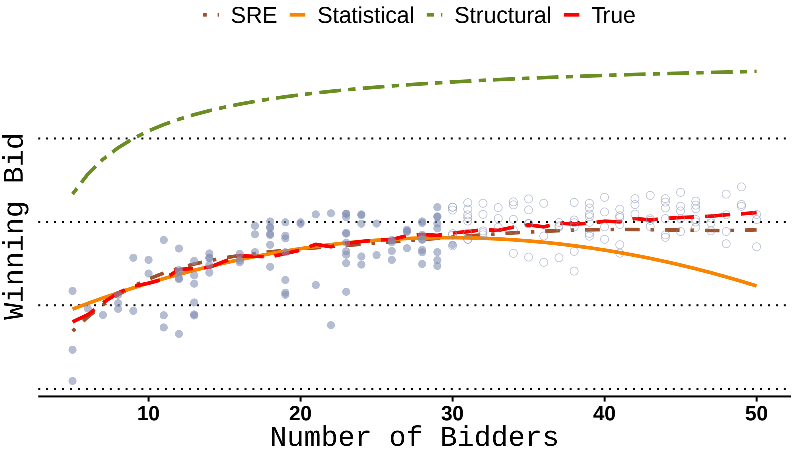
Experiment 1: out-of-domain

First-Price Auction



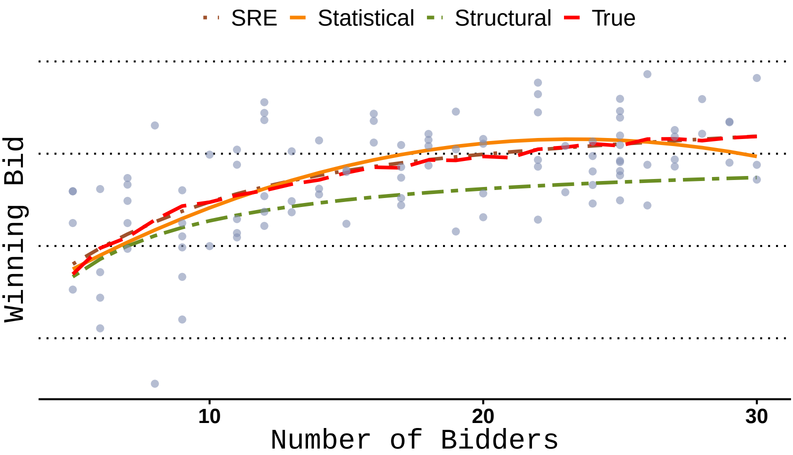
Experiment 2: in-domain

First-Price Auction



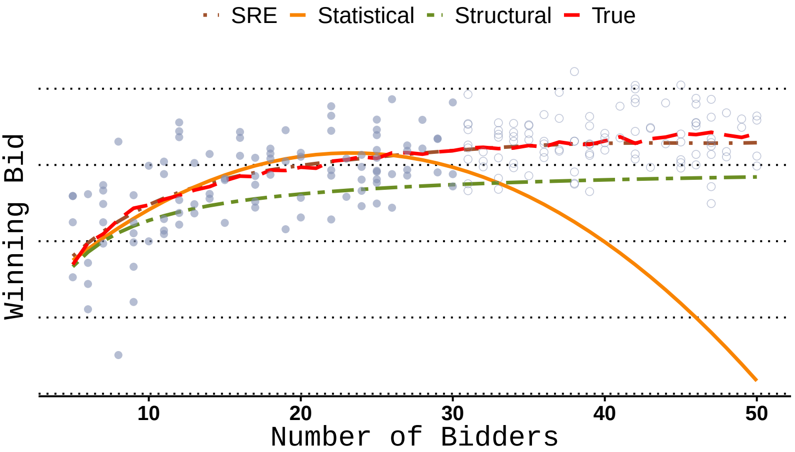
Experiment 2: out-of-domain

First-Price Auction



Experiment 3: in-domain

First-Price Auction



Experiment 3: out-of-domain

First-Price Auction

		In-Domain			Out-of-Domain		
Experiment	Estimator	Bias	Var	MSE	Bias	Var	MSE
1	Statistical	0.0037	0.0003	0.0003	0.1256	118.8697	118.9157
	Structural	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	SRE	0.0035	0.001	0.0001	0.0431	0.2690	0.2734
2	Statistical	0.1251	0.1129	0.1426	3.5506	572.1151	595.3736
	Structural	9.8255	0.0000	98.5233	12.6631	0.0000	160.6319
	SRE	0.2772	0.0883	0.2399	0.3359	12.7092	12.8710
3	Statistical	0.1455	0.1987	0.2307	0.5285	18.2734	18.5692
	Structural	7.7095	0.0000	60.4870	9.8956	0.0000	98.0982
	SRE	0.2787	0.1991	0.3215	0.4517	817.5363	817.8250

Notes: results are based on 100 simulation trials. Reported are the mean bias, variance, and MSE, averaged over the number of bidders n. Since the structural model predicts $\mathbb{E}[b^*|n] = (n-1)/(n+1)$, its predictions have zero variance.

Dynamic Entry and Exit

Baseline Data Generating Model

- N agents in a market. In each period, incumbents decide whether to remain in the market and potential entrants decide whether to enter. The return to operating in the market in time t is R_t (exogenous time-varying payoff).

DYNAMIC ENTRY AND EXIT

Baseline Data Generating Model

- Let entry status be indicated by $(0, 1)$. The time-t flow utility of agent i who is in state j in time $t-1$ and state k in time t is:

$$u_{i,t}^{j,k} = \pi\left(R_t, c^{j,k}\right) + e_{i,t}^{k},$$

where $c^{j,k}$ is the cost of transitioning from state j to k, and $e_{i,t} = (e_{i,t}^0, e_{i,t}^1) \sim^{i.i.d.}$ Gumbel $(0, 1)$ are idiosyncratic shocks.

Dynamic Entry and Exit

Baseline Data Generating Model

- At the beginning of each period t, each agent receives an idiosyncratic shock $e_{i,t}$ and chooses her entry status $d_{i,t}$ by solving the following problem:

$$d_{i,t} = \arg\max_{k} \left\{ \pi_t^{j,k} + e_{i,t} + \beta \mathbb{E}_t \overline{V}_{t+1}^k \right\},$$

where

$$\overline{V}_t^j = \mathbb{E}_e \left[V_{i,t}^j \left(e_{i,t} \right) \right]$$

$$V_{i,t}^j \left(e_{i,t} \right) = \max_{k} \left\{ \pi_t^{j,k} + e_{i,t} + \beta \mathbb{E}_t \overline{V}_{t+1}^k \right\}$$

Dynamic Entry and Exit

Let y_t denote the percentage of entries in period t.

Target function: $E\left[y_t \mid R_t\right]$

- Statistical model:

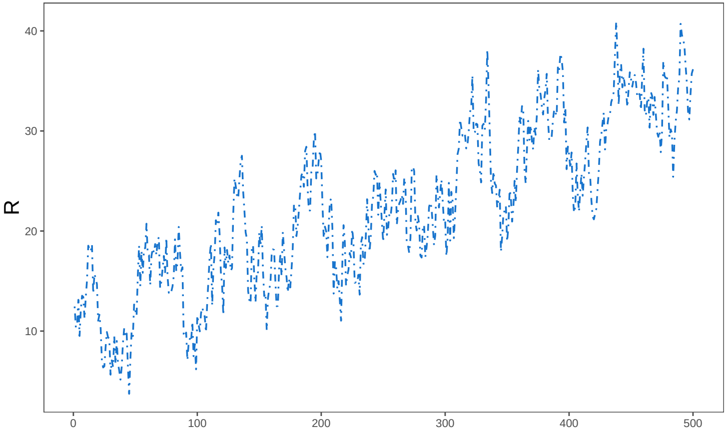
$$y_t = \text{ARMAX}\left(R_t\right) + e$$

- Structural estimation: Hotz-Miller CCP (Arcidiacono and Miller, 2011)

Dynamic Entry and Exit

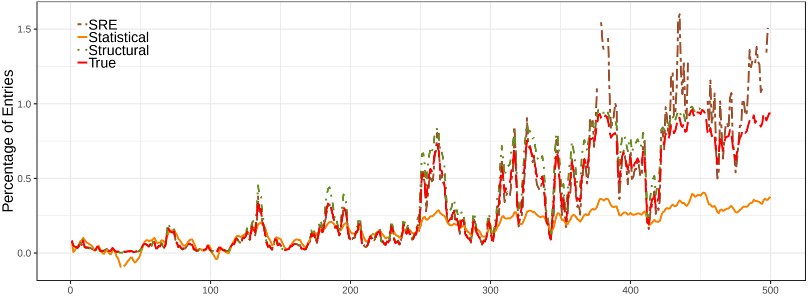
Experiment	True Mechanism	Structural Model
1	Rational Expectations	
2	Adaptive Expectations	Rational Expectations
3	Myopic	

Dynamic Entry and Exit



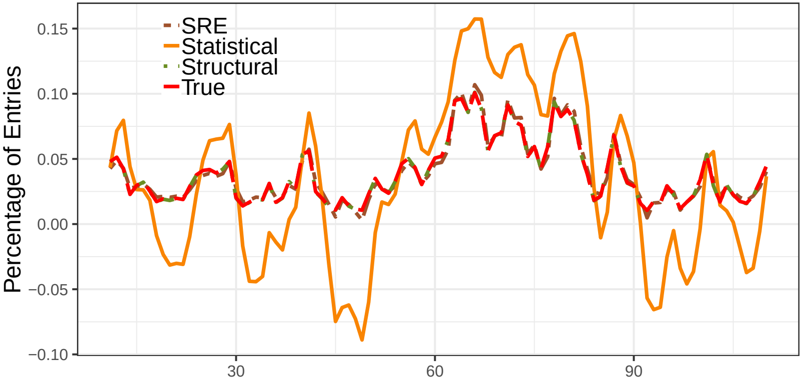
Exogenous Operating Profits

Dynamic Entry and Exit



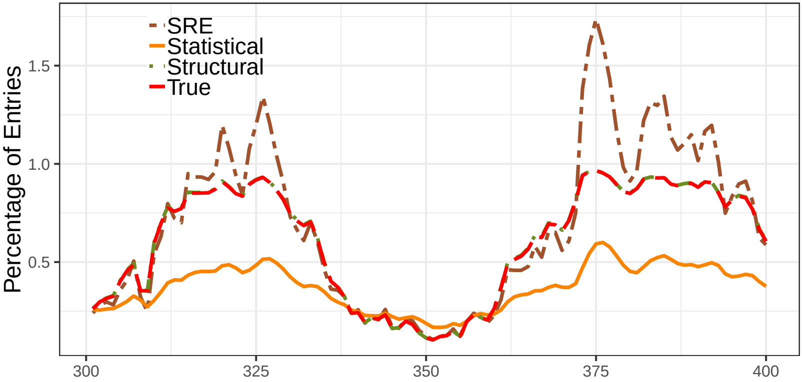
Experiment 1

Dynamic Entry and Exit



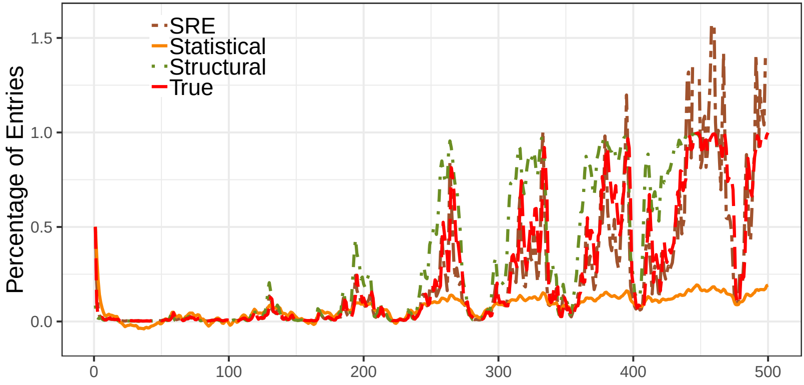
Experiment 1: in-domain

Dynamic Entry and Exit



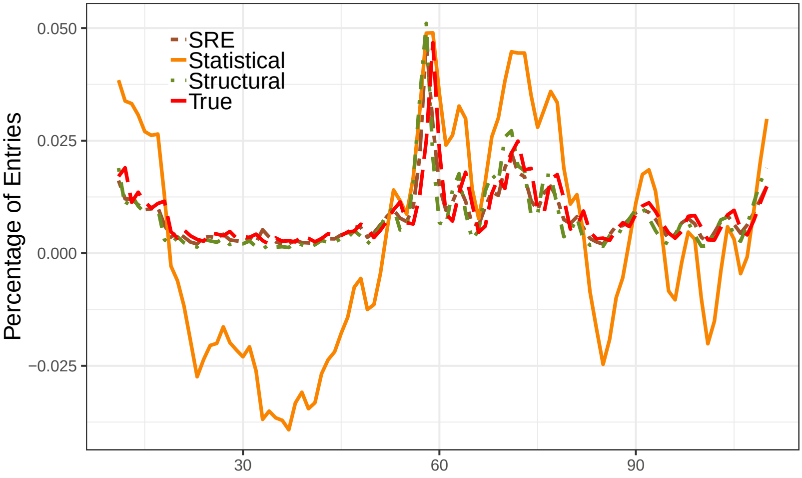
Experiment 1: out-of-domain

Dynamic Entry and Exit



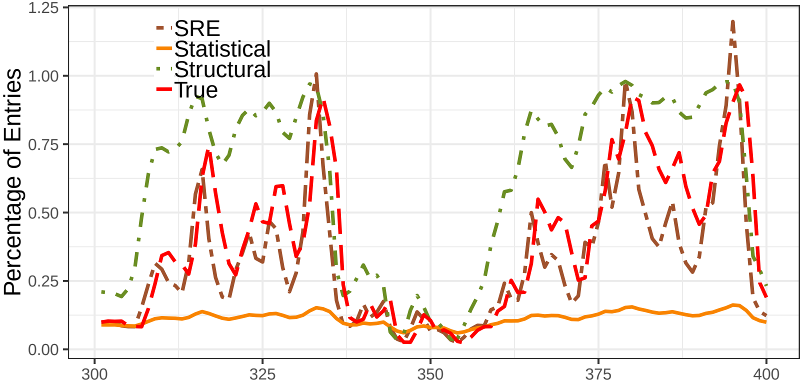
Experiment 2

Dynamic Entry and Exit



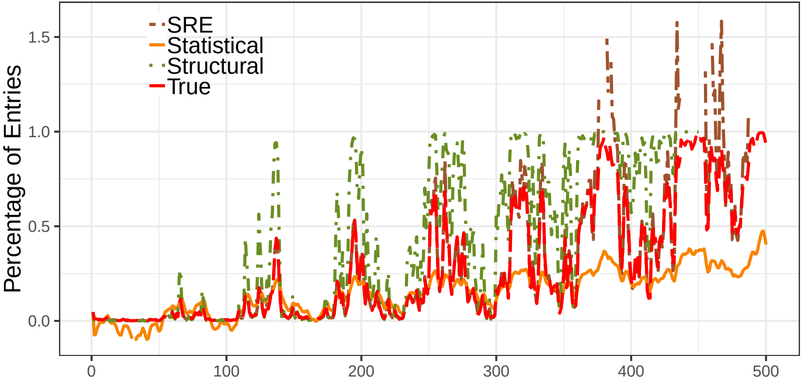
Experiment 2: in-domain

Dynamic Entry and Exit



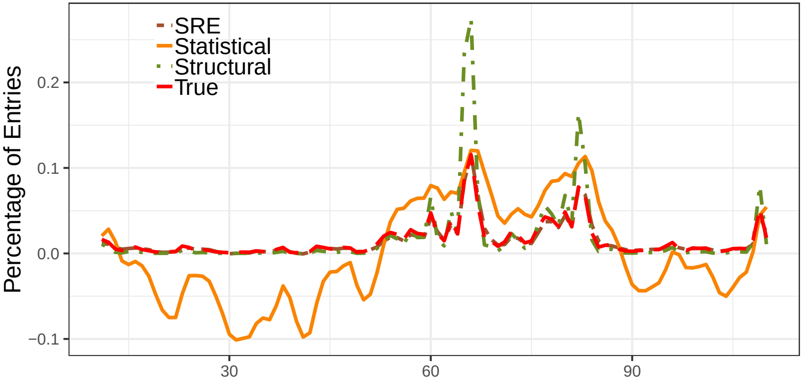
Experiment 2: out-of-domain

Dynamic Entry and Exit



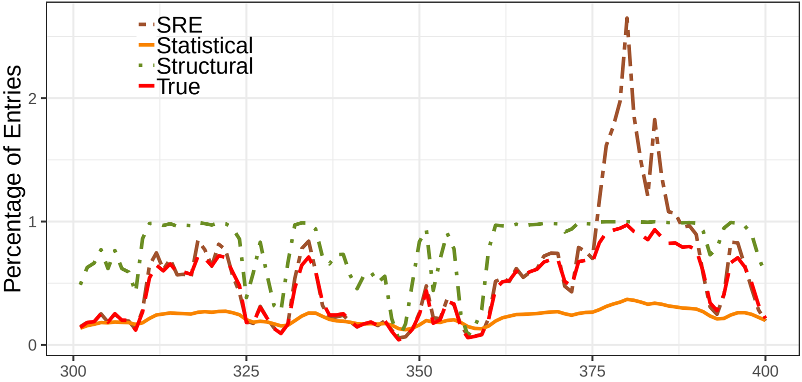
Experiment 3

Dynamic Entry and Exit



Experiment 3: in-domain

Dynamic Entry and Exit



Experiment 3: out-of-domain

Dynamic Entry and Exit

Experiment	Estimator	In-Domain			Out-of-Domain		
		Bias	Var	MSE	Bias	Var	MSE
1	Statistical	0.0321	0.0018	0.0033	0.1347	0.0027	0.0400
	Structural	0.0012	0.0009	0.0009	0.0180	0.0143	0.0148
	SRE	0.0014	0.0009	0.0009	0.0353	0.0310	0.0392
2	Statistical	0.0323	0.0046	0.0085	0.1258	0.6969	0.7371
	Structural	0.0074	0.0002	0.0023	0.1621	0.0259	0.0731
	SRE	0.0037	0.0001	0.0004	0.0472	0.0506	0.0668
3	Statistical	0.0329	0.0015	0.0031	0.1558	0.0047	0.0583
	Structural	0.0073	0.0025	0.0026	0.2111	0.0335	0.0986
	SRE	0.0010	0.0004	0.0004	0.0736	0.0661	0.1091

Notes: results are based on 100 simulation trials. Reported are the mean bias, variance, and MSE, averaged over time t.

Demand Estimation

- We revisit the demand estimation problem under a different setting.
- Instead of observing consumer demand under exogenously varying prices, the prices we observe are set by a monopolist and are hence endogenous.
- We are interested in learning the true demand curve.
- This exercise focuses on comparing the *in-domain* performance of different estimators.

Demand Estimation

Baseline Data Generating Model

- Consider M geographical markets in which a product is sold.
- Market-specific aggregate demand

$$q_m = Q^d(p_m) = \alpha - \beta \cdot p_m + \epsilon_m, \tag{1}$$

where (p_m, q_m) are equilibrium price and quantity in market m.

Demand Estimation

Baseline Data Generating Model

- The monopoly firm incurs marginal cost c_m for operating in market m. Optimal pricing

$$p_m = \arg\max_{p>0} \left\{ (p - c_m) \, Q^d (p) \right\} = c_m + \frac{1}{\beta} q_m \tag{2}$$

- Assume we observe a cost-shifter z_m (e.g. transportation costs) such that

$$c_m = a + b \cdot z_m,$$

then z_m can serve as an instrument for p_m for identifying $Q^d(p)$.

Demand Estimation

- We conduct 4 experiments in this exercise.
- The baseline model is used to generate data for experiment 1 and 3. For experiment 2 and 4, we assume the monopoly firm fails to set optimal prices or does not have complete monopoly power, so that

$$p_m = c_m + \frac{\lambda}{\beta} q_m,$$

  where $\lambda \in (0, 1)$.
  - The firm earns a lower markup than a optimal price-setting monopoly.

Demand Estimation

Reduced-form estimation

- Use z_m as intrument for p_m and estimate $Q^d(p)$ by 2SLS.
- In experiment 1 and 2, we fit the correct $Q^d(p)$ model (1). In experiment 3 an 4, we fit a misspecified model:

$$\log q_m = \alpha - \beta \cdot \log p_m + \epsilon_m$$

Demand Estimation

Structural estimation

- We fit a structural model featuring demand function (1) and price-setting function (2).
- From (1) and (2), we obtain

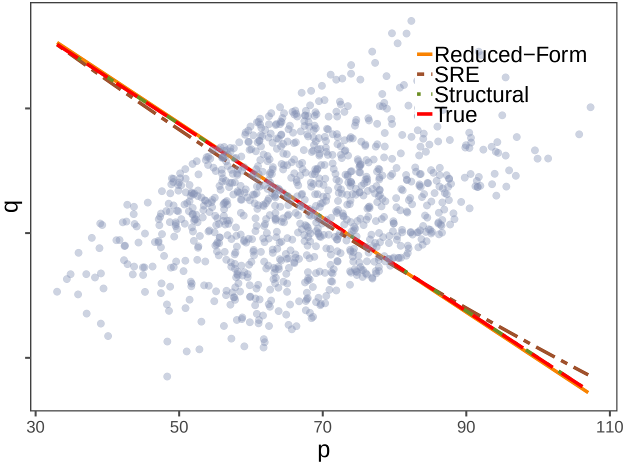
$$p_m = a + b \cdot z_m + \frac{1}{\beta} q_m,$$

from which we can solve directly for $\left(\widehat{a}, \widehat{b}, \widehat{\beta}\right)$. Substituting $\widehat{\beta}$ into (1) obtains $\widehat{\alpha} = \frac{1}{M} \sum_{m=1}^{M} \left(q_m + \widehat{\beta} p_m\right)$.

Demand Estimation

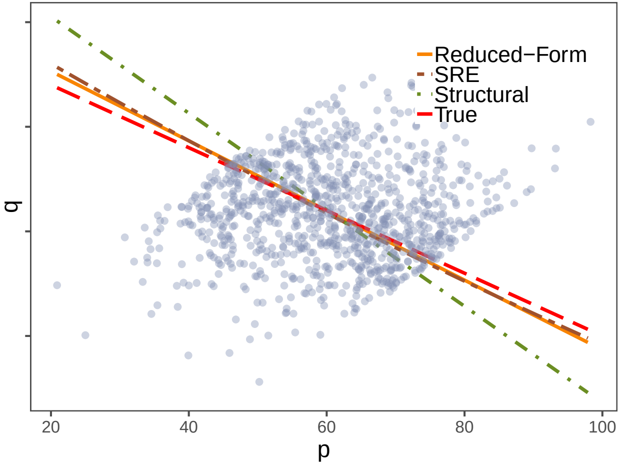
Experiment	True Mechanism	Reduced-Form	Structural Model
1	linear demand, optimal monopoly pricing	linear demand	linear demand, optimal monopoly pricing
2	linear demand, non-optimal monopoly pricing	linear demand	
3	linear demand, optimal monopoly pricing	log-log demand	
4	linear demand, non-optimal monopoly pricing	log-log demand	

Demand Estimation



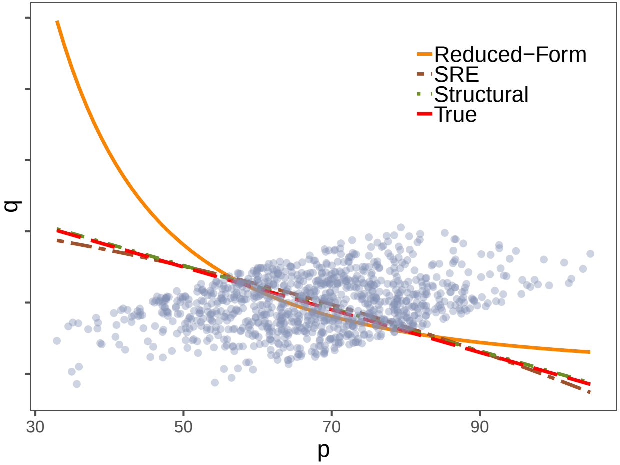
Experiment 1

Demand Estimation



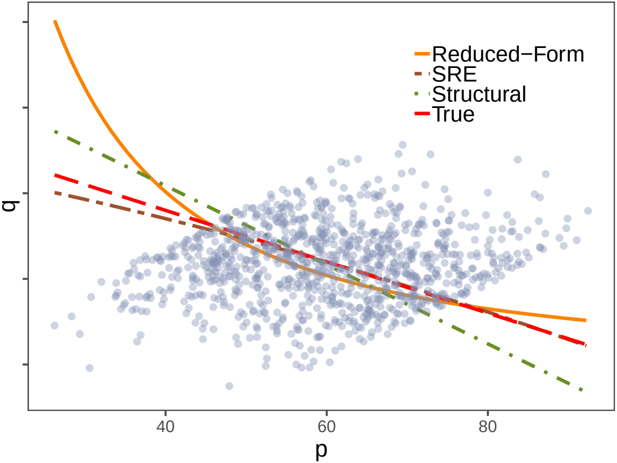
Experiment 2

Demand Estimation



Experiment 3

Demand Estimation



Experiment 4

Demand Estimation

	Reduced-Form			Structural			SRE		
Experiment	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
1	0.2720	7.6780	7.7863	0.0770	0.9102	0.9161	0.9899	11.1375	13.3879
2	0.2884	5.1712	5.2821	12.3218	1.4233	203.8835	0.2783	11.4235	11.5223
3	25.9431	174.9081	2601.8750	0.1167	0.9648	0.9784	0.9703	13.0669	15.2066
4	11.8060	22.3152	423.3862	12.3212	1.4277	203.8519	0.3721	12.3088	12.5519

Notes: results are based on 100 simulation trials. Reported are the mean bias, variance, and MSE, averaged over p.

Empirical Application: Risk Preference

- We study certainty equivalents (CEs) for binary lotteries.
- Data: samples from 44 domains (subject pools), compiled by Andrews et al. (2022).
- Observations take the form $(\bar{z}, \underline{z}, p; y)$.
  - $\bar{z}$ and $\underline{z}$ denote the possible prizes of the lottery.
  - p : the probability of the larger prize.
  - y : the reported certainty equivalent by a given subject.

Empirical Application: Risk Preference

Structural model derived from cumulative prospect theory (CPT):

$$U\left(\bar{z}, \underline{z}, p; y\right) = w\left(p\right) v\left(\bar{z}\right) + \left(1 - w\left(p\right)\right) v\left(\underline{z}\right),$$

where

$$v\left(z\right) = \begin{cases} z^{\alpha} & \text{if } z \geq 0 \\ -\left(-z\right)^{\beta} & \text{if } z < 0 \end{cases},$$

and

$$w\left(p\right) = \frac{\delta p^{\gamma}}{\delta p^{\gamma} + \left(1 - p\right)^{\gamma}}.$$

Empirical Application: Risk Preference

- Based on the structural model, the CE is given by

$$y = v^{-1}\left(w\left(p\right)v\left(\overline{z}\right) + \left(1 - w\left(p\right)\right)v\left(\underline{z}\right)\right)$$

- Alternatively, we can predict the CE using statistical models:

$$y = f\left(\overline{z}, \underline{z}, p\right) + \varepsilon$$

Empirical Application: Risk Preference

Table: Transfer Performance

Model	Transfer Error	Normalized Error
Structural Model	$[1.50, 30.77]$	$[1.05, 2.88]$
SRE	$[1.06, 22.60]$	$[1.01, 2.04]$
Random Forest	$[3.54, 58.62]$	$[1.01, 6.51]$
Neural Net	$[3.82, 132.43]$	$[1.01, 13.85]$

The decision framework: Domain and Model

- $Y = f(X;\theta) + \epsilon$, $E[\epsilon|X] = 0$.
- Domain A, the true model is θ_A, X is in range $[\underline{X}_A, \bar{X}_A]$.
- Observes sample $\{X_i, Y_i\}$ from A.
- Domain B, the true model is θ_B, X in range $[\underline{X}_B, \bar{X}_B]$.

The decision framework: Domain and Model

- The true model in θ_A is identified, in region $[\underline{X}_A, \bar{X}_A]$.
- However, we are ambiguous on the 'location' of the true model θ_B.
- All economic theory implies that θ_B is close to θ_A in some sense (formalized later)

Loss and Decision Criteria

- The mini-max criteria takes into account of the ambiguity

$$\theta^{min-max} = \arg\min_{\theta} \max_{\theta_B \in \Theta^{trans}} E_{P_B}[(f(X;\theta) - f(X;\theta_B))^2]$$

- Θ^{trans} is a set of model to characterize ‘transfer learning’.
- We are ambiguous on the location of θ_B.

Assumptions and Intuitive Model

- The performance of θ_A can inform us about the location of θ_B.
- Assumption on Transfer Learning: The performance of θ_A in the target domain is not too bad:

$$\int_{\underline{X}_B}^{\bar{X}_B} (f(x;\theta_A) - f(x;\theta_B))^2 dQ_{B,x} \leq (\delta^{ext})^2$$

- In other words, while we are ambiguous about θ_B, we know it centers around θ_A.

The Ideal Estimator

- From identification perspective, we only know $f(x;\theta_A)$ up to region $[\underline{X}_A, \bar{X}_A]$—If we use non-parametric method.
- Outside $[\underline{X}_A, \bar{X}_A]$, we don't know.
- However, structural model may some extrapolation power outside $[\underline{X}_A, \bar{X}_A]$.
- Roughly speaking: there may exist a local neighborhood $\mathcal{B}_h([\underline{X}_A, \bar{X}_A])$, where
  1. On $\mathcal{B}_h([\underline{X}_A, \bar{X}_A])$, statistical method (say serial) is closer to the true θ_A.
  2. On $\mathcal{B}_h([\underline{X}_A, \bar{X}_A])^c$, statistical method loss power, structural model is better approximation.

The Ideal Estimator

- Our best knowledge about the θ_A is that it centers around the following θ^{mix}:

$$\begin{aligned} f(x;\theta^{mix}) = & f(x,\theta_A)1(x \in \mathcal{B}_h([\underline{X}_A, \bar{X}_A])) \\ & + f(x,\theta^{struc}))1(x \in \mathcal{B}_h([\underline{X}_A, \bar{X}_A])^c) \end{aligned}$$

  - The first part is identified by statistical model.
  - θ^{struc} is the best restricted-model.

The Ideal Estimator

- It is not hard to see the min-max decision will lead to θ^{mix} as desired model.
- However, it is infeasible since the shape of the neighborhood $\mathcal{B}_h$ is unknown.
- However, the size of $\mathcal{B}_h$ can be measured by the overall performance of structural model in the source domain.

Summary

- We propose a way to combine statistical method with structural method.
- Still working on:
  - Alternative result with non-uniform λ
  - Convergence Result