Machine Learning and Economic Models

Review of Recent Works by Drew Fudenberg and Annie Liang

Renjie Zhong

August 2023

Outline

Introduction

Motivation: Learning the Relationship between x and y

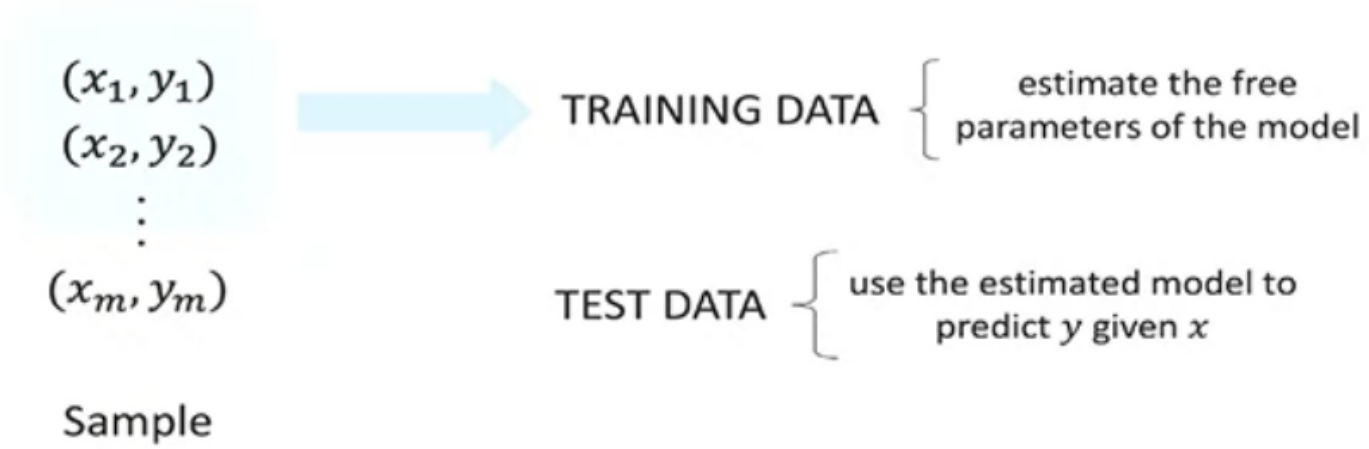
- A standard approach (in both economics and CS) for learning f:
  1. Start with a family of mappings $\mathcal{F}$ (often parametric)
  2. Use data to choose between them (e.g. by training/estimatingfree parameters) $\Rightarrow$ arrive at an estimation of f
- ML: choose $\mathcal{F}$ to be as large as possible subject to data limitations and feasibility of identifying optima $\Rightarrow$ black-box algorithm
- EM: choose $\mathcal{F}_\Theta$ where the parameter $\theta \in \Theta$ has meaning, and the imposed structure is interpretable $\Rightarrow$ a narrative/story

Focus and Contribution

- Focus on evaluating and improving how well a model predicts outcomes.
- Beyond predictive accuracy:
  1. completeness: how much it improves predictions over a naive baseline, relative to how much improvement is possible (performance test)
  2. restrictiveness: how well it matches arbitrary hypothetical data (relevance/casuality test)
  3. transferrability: how well a model trained on data from one domain will perform in a new domain (extrapolation test)
- an axiomatic foundation and estimator
- Applications: certainty equivalents, initial play in games

Elements in Prediction Problem: One Domain

- $X \in \mathcal{X}$ is an observable feature vector that is used to make predictions
- $Y \in \mathcal{Y}$ is an outcome-the thing we are trying to predict.
- Any $f : \mathcal{X} \to \mathcal{Y}$ is a (predictive) mapping e.g., a mapping from lotteries into certainty equivalents.
- We consider a parametric economic models $\mathcal{F}_\Theta = \{f_\theta\}_{\theta \in \Theta}$



Motivating Example: Predicting Certainty Equivalents

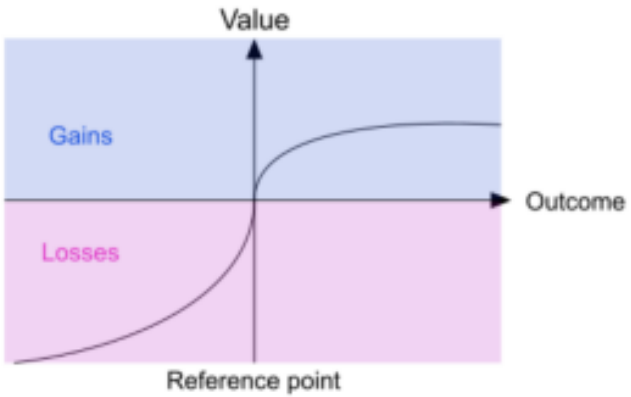
- Subject is offered a risky lottery $(\bar{z}, \underline{z}, p)$, which is X in this prediction problem:

  $\bar{z}$ with probability p

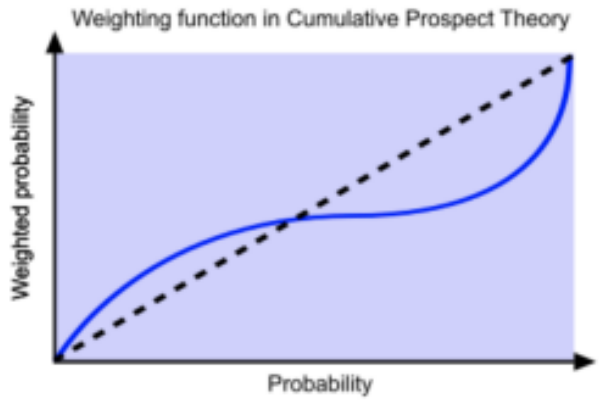
  $\underline{z}$ with probability $1 - p$

- Predicting Certainty Equivalents CE, which is Y in this prediction problem
- Models: $\mathcal{F}_\Theta$, parameter set $\Theta = \{\alpha, \beta, \delta, \gamma\}$

$$CE = w(p) \times v(\bar{z}) + (1 - w(p)) \times v(\underline{z})$$



parameters α, β



parameters δ, γ

Completeness

Prediction Performance Evaluation

- Loss function: $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ gives the error assigned to a prediction of y' when the realized outcome is y

  e.g. $\ell(y', y) = (y' - y)^2$ or $\ell(y', y) = \mathbf{1}(y' \neq y)$ respectively.

- Given $(x, y) \sim P$, $\mathcal{E}_P(f) = \mathbb{E}_P[\ell(f(x), y)]$: expected error of prediction rule f on a new observation

Prediction Performance Evaluation

- Errors to our interest:
  1. reference point: $\mathcal{E}_P(f_{\text{base}})$

     a baseline model $f_{\text{base}} : \mathcal{X} \to \mathcal{Y}$ suited to the prediction problem, e.g. the lottery's expected value is a natural baseline in the motivating example
  2. ideal point: $\mathcal{E}_P(f^*) = \mathbb{E}_P[\ell(f^*(x), y)]$

     ideal prediction rule $f^*(x) = \underset{y' \in \mathcal{Y}}{\operatorname{argmin}} \mathbb{E}_P[\ell(y', y) \mid x]$
  3. ideal prediction in $\mathcal{F}_\Theta$: $\mathcal{E}_P(f^*_\Theta)$.

     $f^*_\Theta = \underset{f \in \mathcal{F}_\Theta}{\operatorname{argmin}} \mathcal{E}_P(f)$ minimizes the expected prediction error in the parametric class $\mathcal{F}_\Theta$
- Completeness: how much it improves predictions over a naive baseline, relative to how much improvement is possible

Completeness

- Completeness: $\kappa = \frac{\mathcal{E}_P(f_{\text{base}}) - \mathcal{E}_P(f^*_\Theta)}{\mathcal{E}_P(f_{\text{base}}) - \mathcal{E}_P(f^*)}$
  1. normalization

     $0 \Rightarrow$ no better than thebaseline, $1 \Rightarrow$ a "fully complete"model
  2. Measuring "units" of completeness as percentage improvements in prediction error facilitates comparison across settings with different loss functions.

Estimating Completeness from Finite Data

- $\mathcal{F} = \{f_{\text{base}}\}$, $\mathcal{F} = \mathcal{F}_\Theta$, and $\mathcal{F} = \mathcal{X}^\mathcal{Y}$ respectively return the desired prediction errors $\mathcal{E}_P(f_{\text{base}})$, $\mathcal{E}_P(f^*_\Theta)$, and $\mathcal{E}_P(f^*)$
- 10-fold cross-validated out-of-sample error.
  1. Split data into $K = 10$ equally sized disjoint subsets $Z_1, \ldots, Z_K$ In each iteration $1 \leq i \leq K$, one as test data $Z^i_{\text{test}} \equiv Z_i$ and remaining as training data $Z^i_{\text{train}} \equiv \cup_{j \neq i} Z_j$
  2. Select a mapping from F that minimizes the in-sample performance: $e(f, Z^i_{\text{train}}) = \frac{1}{|Z^i_{\text{train}}|} \sum_{(x,y) \in Z^i_{\text{train}}} \ell(f(x), y)$
  3. Evaluate how well the chosen mapping performs out of sample: $\text{CV}_i = e(f_i, Z^i_{\text{test}})$
  4. Average over out-of-sample errors: $\text{CV}\left(\mathcal{F}, \{Z_i\}_{i=1}^K\right) = \frac{1}{K} \sum_{i=1}^K \text{CV}_i$

Theoretical gurantees

TABLE 2
OUR RESULTS IN THE SUBSEQUENT APPLICATIONS

	Error	Completeness (%)
Baseline	$\hat{\mathcal{E}}_{\text{base}}$	0
Economic model	$\hat{\mathcal{E}}_{\Theta}$	$100 \times (\hat{\mathcal{E}}_{\text{base}} - \hat{\mathcal{E}}_{\Theta})/(\hat{\mathcal{E}}_{\text{base}} - \hat{\mathcal{E}}_{\text{best}})$
Irreducible error	$\hat{\mathcal{E}}_{\text{best}}$	100

- The empirical quantities $\hat{\mathcal{E}}_{\text{base}}$,$\hat{\mathcal{E}}_{\Theta}$, and $\hat{\mathcal{E}}_{\text{best}}$ are consistent estimators for $\mathcal{E}_P\left(f_{\text{base}}\right)$, $\mathcal{E}_P\left(f^*_{\Theta}\right)$, and $\mathcal{E}_P\left(f^*\right)$, respectively (Hastie, Tibshirani, and Friedman 2009)
- The empirical estimate of completeness is also a consistent estimator

Testing CPT

- We evaluate CPT on data from Bruhin et al (2001): 179 certainty equivalents for each of 25 binary lotteries

TABLE 4
CPT Is Nearly Complete for Prediction of Our Data

	Error	Completeness (%)
Baseline	104.63 (10.14)	0
CPT	67.78 (8.37)	94 (2.0)
Irreducible error	65.58 (8.11)	100

- Estimate CPT, and evaluate its mean-squared error for predicting the certainty equivalent Y given the lottery X.
- Benchmark. Because we have a large number of reports per lottery, can estimate $E[Y|X]$ (i.e., the best predictor).

Restrictiveness

Prediction Performance: Model and Narrative

- Good Prediction ✓ Then?

TABLE 4
CPT Is Nearly Complete for Prediction of Our Data

	Error	Completeness (%)
Baseline	104.63	0
	(10.14)	
CPT	67.78	94
	(8.37)	(2.0)
Irreducible error	65.58	100
	(8.11)	

- Is the model a good description of how people perceive/interact? or just flexible enough to mimic potential functions?

  CPT well describes the structure in perception of risk

  CPT is flexible enough to mimic most functions from binary lotteries to certainty equivalents.

Restrictiveness

- We'd like to distinguish between when a model is precisely tailored to capture real regularities from when it is simply unrestrictive.
- Our approach:
  1. Generate synthetic data sets
  2. See how well the model performs on each of these
  3. An unrestrictive model performs well on all data sets
  4. Define restrictiveness to be the (normalized) average error for predicting these synthetic data sets

Discription

- Admissible set $\mathcal{F}$: mappings $f : \mathcal{X} \to \mathcal{Y}$ that obey some basic background constraints.

  e.g. In the lottery example, may impose the constraint that subjects prefer more money to less (certainty equivalents obey FOSD)

  Our restrictiveness measure tells us how restrictive the model is beyond these background constraints

- $d(f, f')$: an appropriate measure for "how different" predictions are under f and f'

  e.g. expected squared distance: $d(f, f') = \mathbb{E}\left[(f(X) - f'(X))^2\right]$

- Approximation error to each generated mapping f:

  $d(\mathcal{F}_\Theta, f) \equiv \min_{f' \in \mathcal{F}_\Theta} d(f', f)$.

  Its expected error is $\mathbb{E}[d(\mathcal{F}_\Theta, f)]$

Restrictiveness

- The restrictiveness of the model $\mathcal{F}_\Theta$ wrt the admissible set $\mathcal{F}$ is:

$$r := \frac{\mathbb{E}[d(\mathcal{F}_\Theta, f)]}{\mathbb{E}[d(f_{\text{base}}, f)]}$$

  where the expectation is with respect to a uniform distribution on the admissible set $\mathcal{F}$

- So r is the model's normalized approximation error to a random admissible mapping f

  measure is unitless (thanks to the normalization)

  ranges from 0 (completely unrestrictive) to 1 (no better than f_{base} does)

- High restrictiveness: precisely identify regularities in real behavior

- Restrictive models are desirable, but we also want the model to fit real data

Completeness and Restrictiveness

- Completeness is computed from real data while restrictiveness from synthetic data
- For certain "paired" choices of loss ℓ and distance d, $\kappa + r = 1$

  "paired": $d(f, f_{\text{best}}) = \mathcal{E}_P(f) - \mathcal{E}_P(f_{\text{best}})$, where $(X, Y) \sim P$
- Prefer models that have high completeness (good fit to real data) and high restrictiveness (poor fit to synthetic data).

Estimator for Restrictiveness

- I skip the axiomatic foundation for approximation error and discrepency function.

  nonnegativity, symmetry, monotonicity, rescaling of units, linearity

  $\Rightarrow e(\mathcal{F}_\Theta, \mathcal{F}, d) = \mathbb{E}_{f\sim \mathrm{Unif}(\mathcal{F})}[c \cdot d(\mathcal{F}_\Theta, f)] \quad \forall \mathcal{F}_\Theta, \mathcal{F}, d$
- Restrictiveness: $r(\mathcal{F}_\Theta) := \frac{\mathbb{E}_{f\sim \mathrm{Unif}(\mathcal{F})}[d(\mathcal{F}_\Theta, f)]}{\mathbb{E}_{f\sim \mathrm{Unif}(\mathcal{F})}[d(f_{\text{base}}, f)]}$
- Approach:
  1. Sample M times uniformly at random from admissible set $\mathcal{F}$
  2. Use sample averages in place of expectations
- Estimator: $\hat{r}_M := \frac{\frac{1}{M}\sum_{m=1}^{M} d(\mathcal{F}_\Theta, f_m)}{\frac{1}{M}\sum_{m=1}^{M} d(f_{\text{base}}, f_m)}$
- Confidence intervals for r can also be constructed in the standard way

Comparison

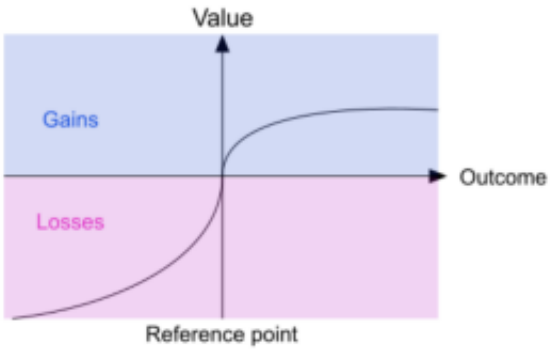
- Cumulative Prospect Theory, henceforth CPT(α, δ, γ) :
  1. utility of lottery $(\bar{z}, \underline{z}, p)$ is $w(p) \times v(\bar{z}) + (1 - w(p)) \times v(\underline{z})$, where

     $v(z) = z^{\alpha}$ is a value function over money

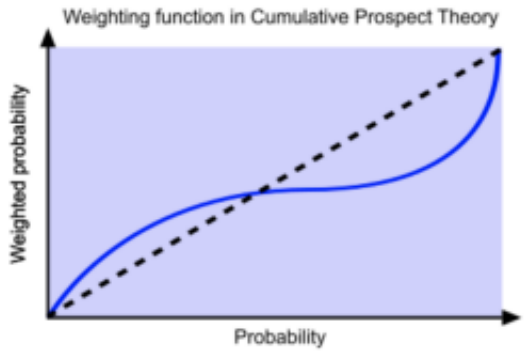
     $w(p) = \frac{\delta p^{\gamma}}{\delta p^{\gamma} + (1-p)^{\gamma}}$ is a probability weighting function
- Disappointment Aversion (Gul, 1991), henceforth DA(α, η) :

  same as above, except that the probability weighting function is

$\tilde{w}(p) = \frac{p}{1+(1-p)\eta}$

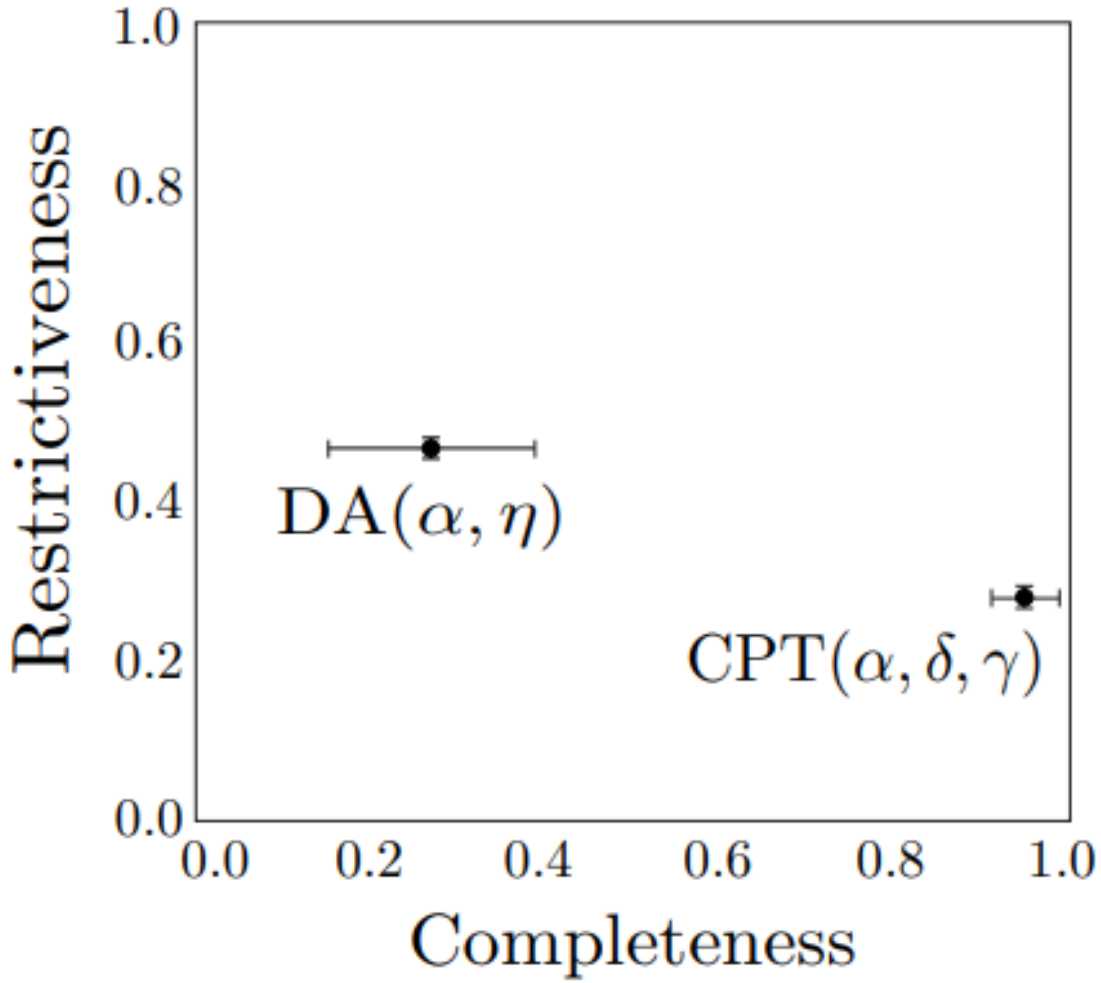


parameters α, β

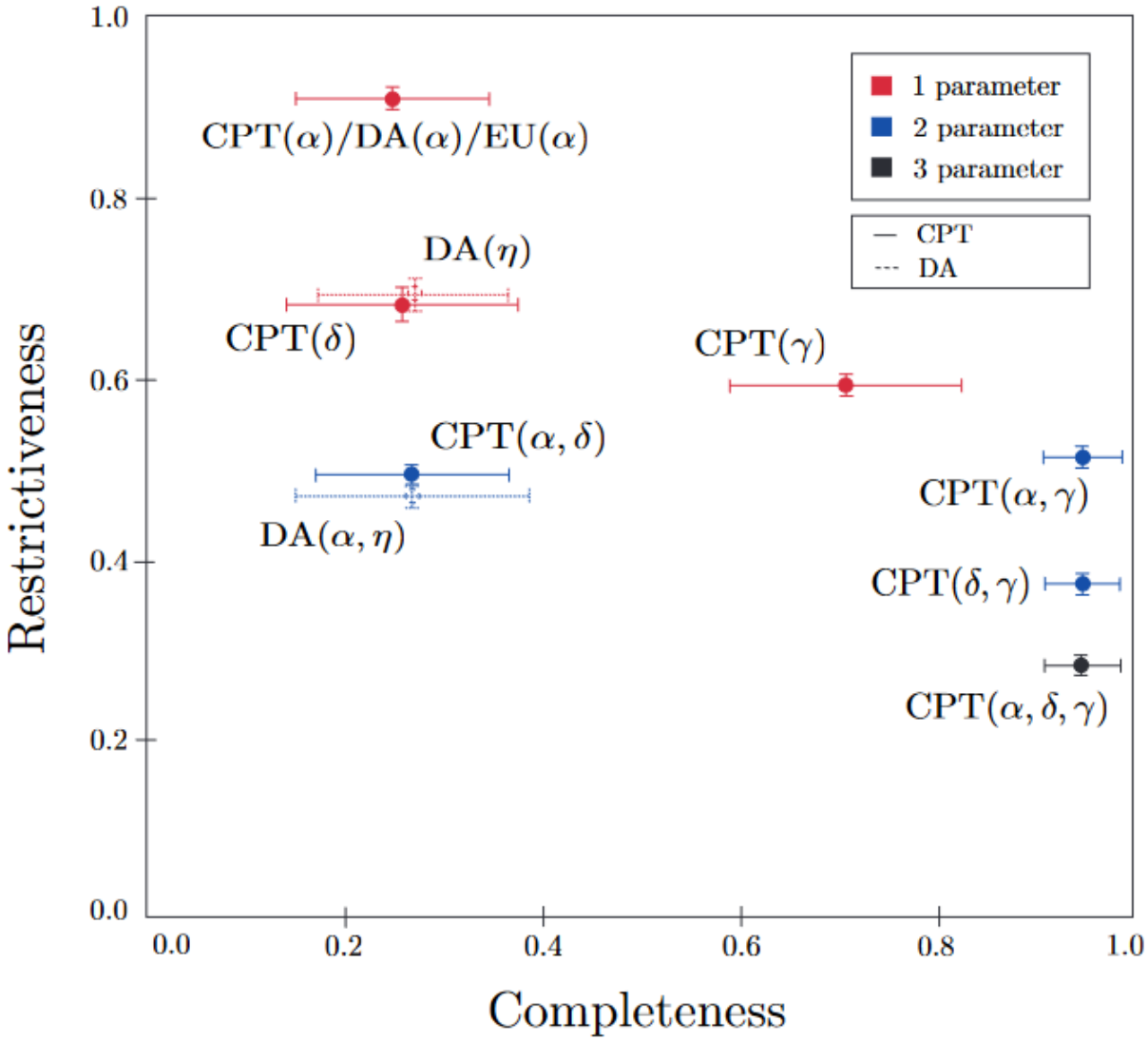


parameters δ, γ

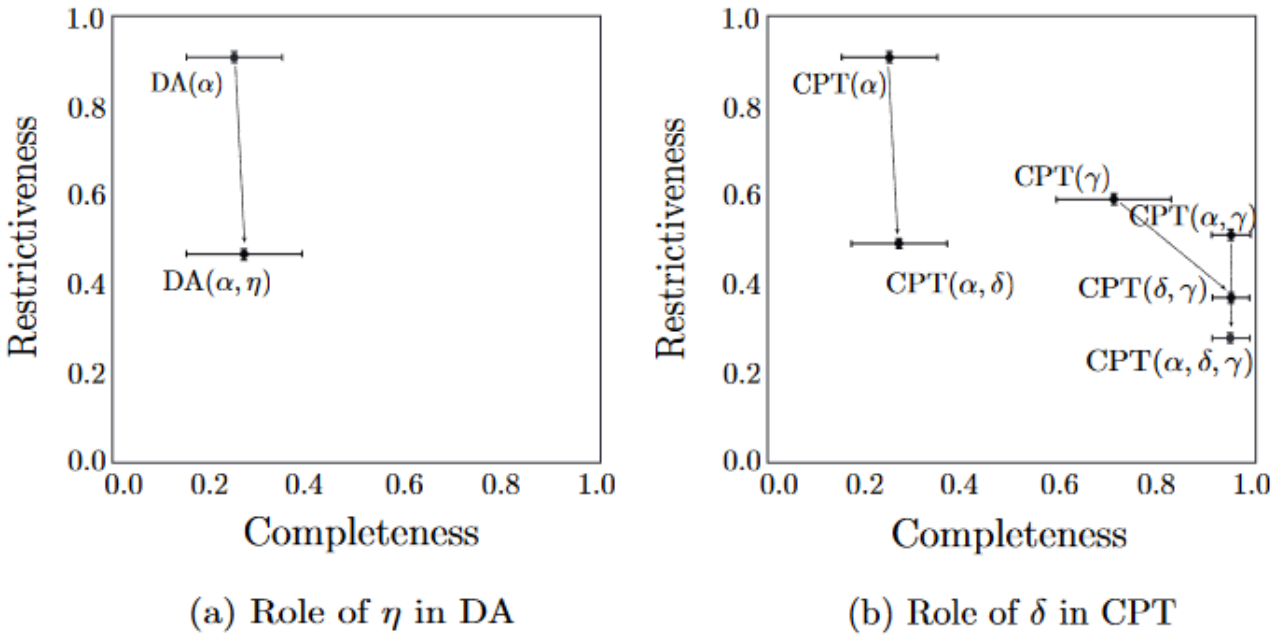
Comparison


1.0
0.8
0.6
0.4
0.2
0.0
Restrictiveness
DA(α, η)
CPT(α, δ, γ)
0.0
0.2
0.4
0.6
0.8
1.0
Completeness

Comparison


1.0
0.8
0.6
0.4
0.2
0.0
Restrictiveness
0.0
0.2
0.4
0.6
0.8
1.0
Completeness
1 parameter
2 parameter
3 parameter
CPT
DA
CPT(α)/DA(α)/EU(α)
DA(η)
CPT(δ)
CPT(γ)
CPT(α, δ)
DA(α, η)
CPT(α, γ)
CPT(δ, γ)
CPT(α, δ, γ)

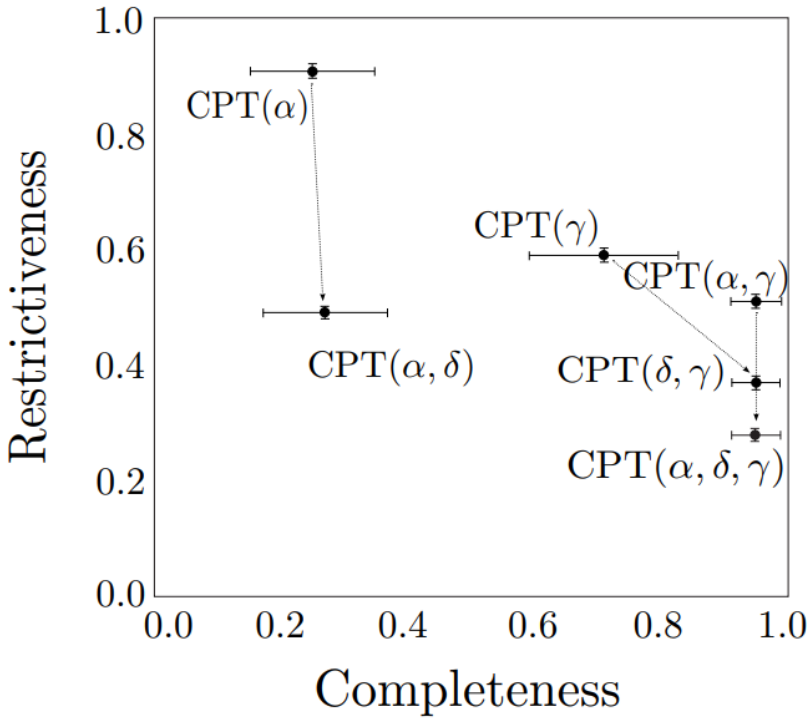
The Value of Parameters



(a) Role of η in DA (b) Role of δ in CPT

- $\tilde{w}(p) = \frac{p}{1+(1-p)\eta}$, η interpreted as disappointment aversion
- $w(p) = \frac{\delta p^\gamma}{\delta p^\gamma + (1-p)^\gamma}$, δ governs elevation of probability weighting function
- not effective: large drop in restrictiveness and small gain in completeness

The Value of Parameters

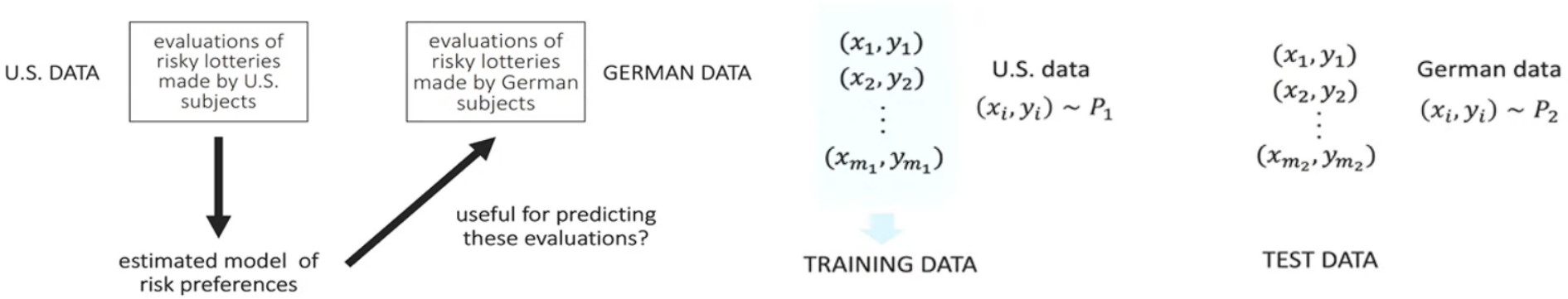


- $w(p) = \frac{\delta p^{\gamma}}{\delta p^{\gamma} + (1-p)^{\gamma}}$, γ governs curvature of probability weighting function
- effective: capture real risk preferences.

Transferability

Elements in Prediction Problem: Across Domains

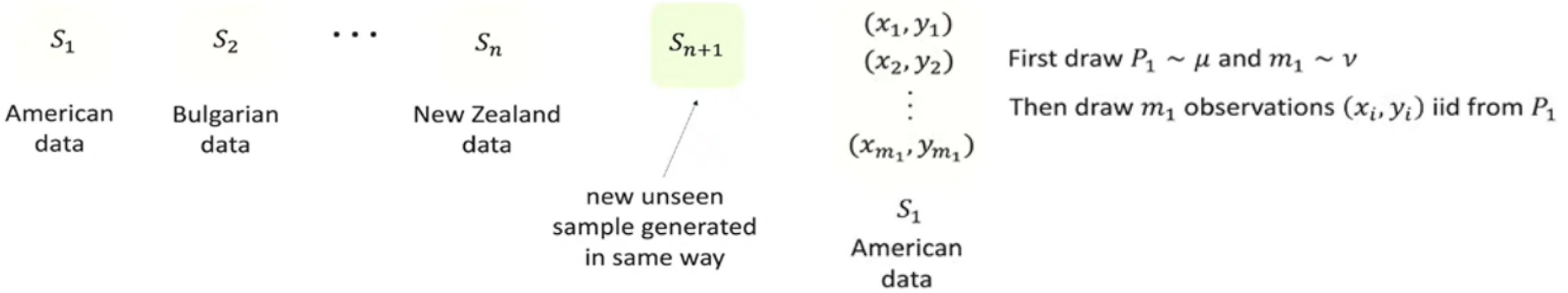
- Sometimes what we're interested in may be instead how well the estimatedmodel will perform on data from a different distribution:



- In this presentation, focusing on IID Baseline.

Elements in Prediction Problem: Across Domains

- The analyst has access to metadata $\mathrm{M} = (S_1, \ldots, S_n)$ consisting of n samples, to predict S_{n+1}
- Assume that across samples $S_1, S_2, S_3, \ldots$ independently generated in this way:
  1. the joint distribution P governing (x, y) is drawn independently from some distribution $\mu \in \Delta(\Delta(\mathcal{X} \times \mathcal{Y}))$
  2. the sample size m is drawn independently from some distribution $\nu \in \Delta(\mathbb{N})$



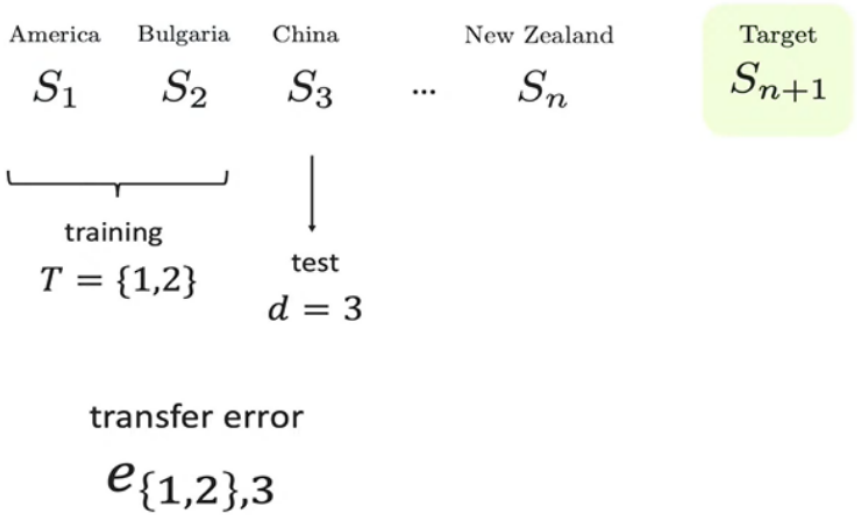
Prediction Across Domains: Evaluations

- $e_{T,d}$: the transfer error when we estimate the model's parameters on $\cup_{t\in T} S_t$ and test on S_d

  $T \subseteq \{1,\ldots,n\}$: training samples

  $d \in \{1,\ldots,n,n+1\}$: test sample

- We want to provide confidence intervals for $e_{T,n+1}$, i.e., the transfer error on the target sample



Prediction Across Domains: Evaluations

- Consider the pooled sample of all transfer errors $e_{T,d}$ where we vary over
  1. all subsets $T \subseteq \{1, \ldots, n\}$ of fixed size n_T
  2. all choices of $d \in \{1, \ldots, n\} \backslash T$
- For any quantile $\tau \in [0, 1]$, let e_τ denote the τ-th quantile of this pooled sample

Confidence Interval

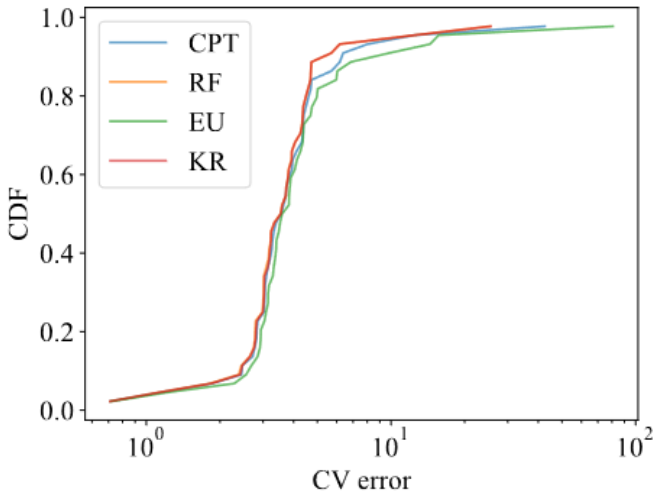
For any $\tau \in (0.5, 1)$,

$$\mathbb{P}\left(e_{\tilde{T},n+1} \notin [e_{1-\tau}, e_\tau]\right) \leq 4\left(1 - \frac{n-n_T}{n-n_T+1}\tau\right)$$

- So $[e_{1-\tau}, e_\tau]$ is a level- $\left(4\left(\frac{n-n_T}{n-n_T+1}\tau\right) - 3\right)$ confidence interval for the transfer error on the target sample.

Prediction Across Domains: Evaluations

- Metadata: 44 samples of reported certainty equivalents across differentsubject pools (from 14 papers)
- Economic models: Expected Utility and CPT
- Machine Learning: kernelized ridge regression and random forest
- The black box's cross-validated Errors is (slightly) lower thanthat of the economic models for most of our subject pools



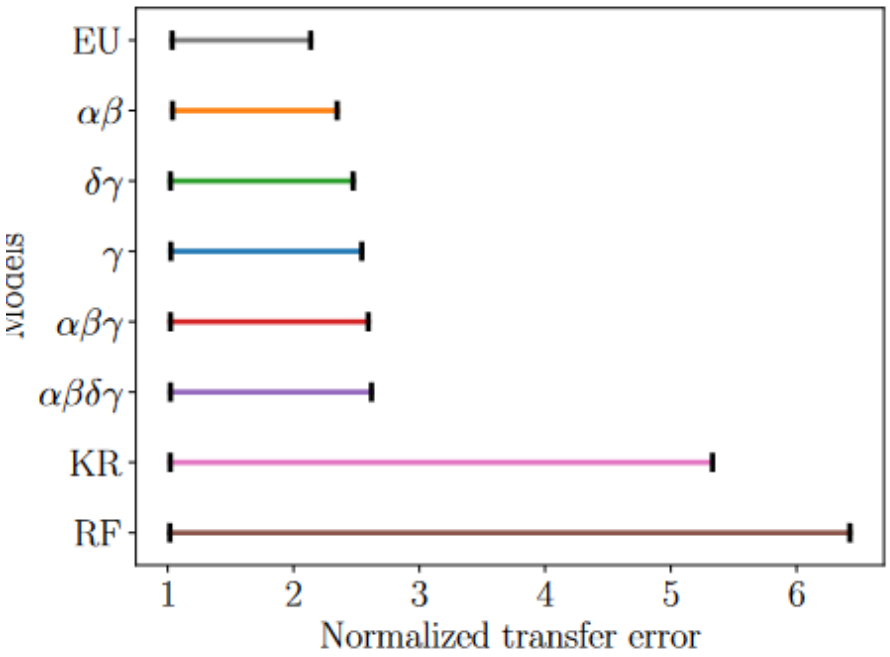
Comparison

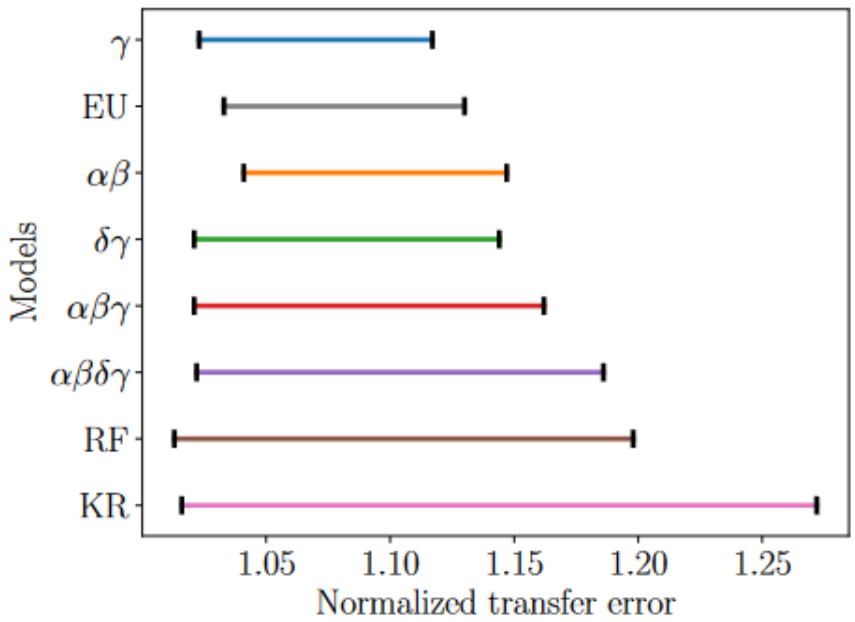
- Black boxes are very flexible and hence learn idiosyncratic details that do not generalize across subject pools
- Economic models transfer better supports the intuition that economic models can recover regularities that are general across a variety of domains
- With $P \in \Delta(\mathcal{X} \times \mathcal{Y})$, Variation across domains may come from two sources:
  1. Common P_X, different $P_{Y|X}$ ("model shift")

     same lotteries but different reported certainty equivalents
  2. Different P_X, common $P_{Y|X}$ ("covariate shift")

     different samples involve different lotteries
- Black boxes seem to transfer worse in the latter instead of in the former.

Prediction Across Domains: Evaluations





Discussions

Takeaway

- When a theory fits the data well (completeness), it matters whether this is:
  1. because the theory captures important regularities in the data
  2. because the theory is so flexible that the only constraints it imposes are basic background constraints that we already know (restrictiveness)
- It also matters whether it captures **fundamental regularities** that apply in a wide variety of domains (transferrability)

Discussions

- more criterions, more applications, more foundation problems
- With enough data from different domains, do black boxes winout again? (transferrability)
- If our goal is purely to predict as well as possible in the newdomain, what is the best approach? (extrapolation ability)
- Empirical values especially for reduced-form model (like in political formal model) (restrictiveness)
- + Model-based inference (subjective model) and decision outcome

Discussions

- ls that where economics is heading, or does it have properties that make it fundamentally different?
  1. more data-poor (counterfactual predictions inenvironments where we do not yet have data)
  2. the importance of parameter estimates for informing welfare analysis
  3. if “qualitative economic intuition”is real and valuable