Evaluating and Improving Economic Models

Drew Fudenberg

September 2021

Introduction

- Three related papers that use machine learning as a complement to theoretical modeling, rather than a substitute for it.
- Focus on evaluating and improving how well a model predicts outcomes.
- Predictive accuracy is only one of many criteria that matter for selecting theories: we also value e.g. parsimony, portability, and causal explanations.
- Our work is intended to clarify some of the tradeoffs and to help focus efforts to develop better theories

Key Concepts:

- Completeness compares how well a model predicts to the "best possible" predictions.
- Restrictiveness measures a theory's ability to match arbitrary hypothetical data: A very unrestrictive theory will be complete on almost any data, so the fact that it is complete on the actual data is not very instructive.
- Algorithmic experimental design is a way to select which experiments to run.

Setting

- $X \in \mathcal{X}$ 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 $f : \mathcal{X} \rightarrow \mathcal{Y}$ is a (predictive) mapping e.g., a mapping from lotteries into certainty equivalents.
- We consider a parametric economic models $\mathcal{F}_\Theta = \{f_\theta\}_{\theta \in \Theta}$.

Example: Predicting Certainty Equivalents

- Subject is offered a risky lottery:

$$\begin{array}{ll} \overline{z} & \text{with probability } p \\ \underline{z} & \text{with probability } 1-p \end{array}$$

  where $\overline{z} > \underline{z} > 0$ (gains domain).

- Ask each subject for their certainty equivalent–the dollar amount x such that the subject would be indifferent between the lottery or x dollars for sure.
- Here the features are $X = (\overline{z}, \underline{z}, p)$.
- Want to predict the average (over subjects) certainty equivalent in a lottery, so the outcomes Y are real numbers.
- One parametric model is CARA utility, $u(x) = x^{\alpha}$.
- Cumulative Prospect Theory (CPT) adds 2 parameters to capture non-linear probability weighting.

Completeness

“Measuring the Completeness of Economic Models,”
Fudenberg, Kleinberg, Liang, and Mullainathan, *JPE* forthcoming.

- More reason to look for ways to improve a model that predicts poorly than one that predicts well.

- But what constitutes “good" performance?

- Our view is that the answer depends on how well the outcome could possibly be predicted given the specified features.

- Decompose prediction error into
  1. Intrinsic noise given the measured features, the irreducible error.
  2. Regularities in the data that the model does not capture.

- Irreducible error is an upper bound for how well any model (based on the measured features) could possibly do.

- A benchmark at the other end is the performance of a baseline model, such as “guess the outcome at random.”

- We use these to define a model’s completeness.

Definitions

- 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 $\ell(y', y) = \mathbb{1}(y' \neq y)$ respectively.

- The expected error of prediction rule f on a new observation $(x, y) \sim P$ is
$$\mathcal{E}_P(f) = \mathbb{E}_P[\ell(f(x), y)]. \tag{1}$$

- The prediction rule in the parametric class $\mathcal{F}_\Theta$ that minimizes the expected prediction error is
$$f^*_\Theta = \operatorname*{argmin}_{f \in \mathcal{F}_\Theta} \mathcal{E}_P(f).$$
The expected error of this “best" rule in $\mathcal{F}_\Theta$ is $\mathcal{E}_P(f^*_\Theta)$.

- The ideal prediction rule is

$$f^*(x) = \operatorname*{argmin}_{y' \in \mathcal{Y}} \mathbb{E}_P \left[\ell(y', y) \mid x \right].$$

- If the conditional distribution $Y \mid X$ is not degenerate, then even this ideal prediction rule does not predict perfectly.

- The irreducible error is

$$\mathcal{E}_P(f^*) = \mathbb{E}_P \left[\ell(f^*(x), y) \right].$$

  The irreducible error is a lower bound on the error when predicting Y using the features in X.

- We also fix a baseline model $f_{\text{base}} : \mathcal{X} \to \mathcal{Y}$ suited to the prediction problem. For example, in the prediction of certainty equivalents, the lottery's expected value is a natural baseline.

The completeness of the parametric model class $\mathcal{F}_\Theta$ is

$$\frac{\mathcal{E}_P(f_{\mathsf{base}}) - \mathcal{E}_P(f^*_\Theta)}{\mathcal{E}_P(f_{\mathsf{base}}) - \mathcal{E}_P(f^*)}.$$

- Completeness is a normalized measure of the reduction in error. A model with completeness 0 does no better than the baseline, while a “fully complete” model with completeness 1 removes all but the irreducible error.

- Measuring “units" of completeness as percentage improvements in prediction error facilitates comparison across settings with different loss functions.

Discussion

- Completeness depends on the baseline f_{base}: without a baseline error rate, it's hard to evaluate the magnitude of a model's error.
- Completeness is defined for a fixed feature set $\mathcal{X}$, which we generally interpret as the measured features in the data. If we vary $\mathcal{X}$ by adding new measured features, the predictive performance of the original model remains the same, but the predictive optimum weakly improves.
- In practice estimate the model error $\mathcal{E}_P(f^*_\Theta)$ using tenfold cross validation: Split the data into ten parts, train model on nine and test on the last.
- When there is sufficient data per feature vector, estimate best possible loss by Table Lookup: For each x, learn the best prediction of y on the training data (e.g. learn average y for each x if loss is MSE).

Theoretical Guarantees

- We report completeness as

$$100 \times \left(\widehat{\mathcal{E}}_{\textsf{naive}} - \widehat{\mathcal{E}}_{\Theta}\right) / \left(\widehat{\mathcal{E}}_{\textsf{naive}} - \widehat{\mathcal{E}}_{\textsf{best}}\right)$$

  where $\widehat{\mathcal{E}}_{\textsf{naive}}$, $\widehat{\mathcal{E}}_{\Theta}$, and $\widehat{\mathcal{E}}_{\textsf{best}}$ denote the estimated quantities
- The estimated losses are consistent estimators for the theoretical values, and and the empirical estimate of completeness is a consistent estimator for completeness).
- These estimates are good approximations for the theoretical quantities when the analyst has many observations for each distinct $x \in \mathcal{X}$.
- This is often the case in applications to lab data.
- And seems to be the case in the applications we consider
- (*Summary* Both bootstrapped and analytic standard errors are small, and our "lookup tables" do better than bagged decision trees.)

Testing CPT

- We evaluate CPT on data from Bruhin et al [2010]: 179 certainty equivalents for each of 25 binary lotteries.

- Estimate CPT, and evaluate its mean-squared error for predicting the certainty equivalent Y given the lottery X.

- Use the expected value of the lottery as the baseline model.

- Because we have a large number of reports per lottery, have good estimates of $\mathbb{E}[Y \mid X]$.

CPT Predicts Very Well

	(Mean Squared) Error
Expected Payoff	103.81
CPT	67.38
Best Possible	65.58

- CPT almost minimizes error in this prediction task
- To get better predictions in these 2-outcome lotteries we would need more information, e.g. about individual characteristics such as financial literacy, education, etc.
- Paper also discusses application to mixture models.

Other Domains/Populations

- We repeat our analysis for the completeness of CPT on two additional data sets from Bruhin et al. The three experiments all used the same experimental design, although there was some variation in the set of lotteries.
- The raw mean-squared error of CPT varies substantially across the three data sets.
- This might suggest that CPT is a better model for certain subject populations, but the completeness of CPT turns out to be very stable across all three data sets, and is lower bounded by 92%.
- This shows the usefulness of benchmarks for interpreting raw prediction errors.

Application #2: Initial Play in Games

"Predicting and Understanding Initial Play," Fudenberg and Liang, *AER* [2019] provides the data here and also the illustrations of using machine learning; I also report some results from the the "Completeness" paper.

- Non-equilibrium models are better predictors than Nash equilibrium of the choices that people make when they first encounter a new game.
- But how much of the predictable regularity do they capture?
- We consider prediction of the action chosen by the row player in a given instance of play of a 3×3 normal-form game using three subsamples of a data set from Fudenberg and Liang [2019]: 23,137 total observations of initial play from 486 3×3 matrix games.
- The available features are the 18 entries of the payoff matrix, and a prediction rule is any map $f : \mathbb{R}^{18} \to \{a_1, a_2, a_3\}$ from 3×3 payoff matrices to row player actions.

- Evaluate error using the *misclassification rate*

  $$\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\left(f(g_i)\neq a_i\right).$$

  This is the fraction of observations where the predicted action was not the observed action.

- Baseline: guess uniformly at random for all games, expected misclassification rate of $2/3$.

- Evaluate a prediction rule based on the *Poisson Cognitive Hierarchy Model* (PCHM), which was the best-performing of the models we considered.

- The PCHM achieves 76% of the achievable reduction, which is good, but leaves room for improvements that capture additional regularities.

Using ML to improve theories

- FL [2019] trained a bagged decision tree algorithm to predict play in these games, and found it predicted better.

- The games where play was predicted correctly by our algorithm but not by PCHM all had an action whose average payoffs closely approximated the level-1 action, but which led to lower variation in possible payoff. Players were more likely to choose this than the level-1 action.

- One explanation is that players act as if they are risk averse. This led us to add a single parameter α to the level-1 model, so that the utility of dollar payoffs z is z^{α}.

- Predicted as well as the the decision trees! Machine learning helped us discover an interpretable and portable extension of an existing model.

Algorithmic Experimental Design

Approach:

- Teach an algorithm to recognize games where the model performs poorly.
- Randomly generate games.
- Use algorithm to predict performance of model on the randomly generated games.
- Keep the cases where the model is predicted to perform poorly.
- Then run new experiments on these algorithmically generated games to look for new regularities.
- And learned how to improve predictions with a ML- theory hybrid that used ML to predict which of two theories to use in a given game.

Restrictiveness and Flexibility

"How Flexible is that Functional Form", Fudenberg, Gao, and Liang.

- CPT does a good job of predicting certainty equivalents for 2-outcome lotteries.
- Is this because it is a good description of how people perceive risk? or
- Is CPT flexible enough to mimic most functions from binary lotteries to certainty equivalents?
- We'd like to distinguish between when a model is precisely tailored to capture real regularities from when it is simply unrestrictive.

Measuring Restrictiveness

- Loosely speaking, our idea is to measure the restrictiveness of a model as 1 minus its expected completeness a range of “synthetic data“ that is generated at random.

- To simplify this we imagine that each synthetic data set has infinite number of observations, and so reveals the ideal prediction rule f^*.

- We then see how well the model class $\mathcal{F}_\Theta$ can approximate each f using a discrepancy function d.

- When the loss function is mean squared error, as in our application to certainty equivalents, d is the expected squared distance between f^* and the closest element of $\mathcal{F}_\Theta$.

More Details

Step 1: Define an “admissible set" $\mathscr{F}$ of mappings $f : \mathcal{X} \to \mathcal{Y}$ that obey some basic background constraints, such as that people prefer more money to less.

Step 2: Choose a baseline f_{base} from the model $\mathcal{F}_\Theta$ and evaluate its approximation error to the randomly drawn mappings.

Step 3: Sample mappings uniformly at random from $\mathscr{F}$ and evaluate how well the model $\mathcal{F}_\Theta$ approximates these mappings.

- The model's approximation error to each generated mapping f is $d(\mathcal{F}_\Theta, f) \equiv \min_{f' \in \mathcal{F}_\Theta} d(f', f)$.
- Its expected error is $\mathbb{E}[d(\mathcal{F}_\Theta, f)]$.

Why Uniform?

- Our measure uses a uniform distribution over the admissible mappings.

- The paper discusses a generalized version with respect to other distributions on $\mathscr{F}$ (analyst's prior).

- We prefer the uniform distribution in our applications: It is transparent and easy to interpret, and avoids cherry-picking;

Restrictiveness

The **restrictiveness** of the model $\mathcal{F}_\Theta$ wrt the admissible set $\mathscr{F}$ 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 $\mathscr{F}$.

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

- Restrictiveness ranges from zero (completely unrestrictive) to 1 (approximates admissible mappings no better than f_{base} does).
- It is unitless (thanks to the normalization), and hence insensitive to rescaling.

Related Literature

- Koopmans & Reiersol [1950] introduce a binary notion of “observationally restrictive," where “unrestrictive” means completely vacuous. This is typically determined analytically.
- Representation theorems (e.g. in decision theory) characterize empirical content of models. We don’t have such theorems for most functional forms used in applied work.]
- We provide a quantitative measure that can be numerically computed.
- Our approach differs from the literature on model selection (AIC, BIC, VC dimension), whose objective is to avoid overfitting—these measure typically prefer more complex models when the sample is large. We we assume an intrinsic preference for parsimonious/restrictive models (even with infinite data).

- Closest analog is Selten [1991], which proposed measuring the flexibility of a model by the fraction of possible data sets that it can **exactly explain**.

- Our focus on approximate fit can lead to very different conclusions.
    - For example, consider the set $\{0, 1/n, ...(n-1)/n, 1\}$ as a model for the unit interval.
    - This model has measure zero, so it is extremely restrictive according to Selten's measure no matter the value of n.
    - For large n this model would be very unrestrictive according to our measure with the standard squared distance d.
    - And Selten's measure is generally difficult to evaluate without the guidance of prior analytical results.

In the Paper

- **Axiomatic foundation** for our restrictiveness measure

- **Estimators** for restrictiveness and completeness and characterizations of their asymptotic distributions
  - Allows us to construct **confidence intervals**

- Three **applications**:
  1. Certainty equivalents — lab data
  2. Initial play in games — lab data
  3. Takeup of microfinance in Indian villages — field data

Estimating Restrictiveness

- Randomly sample M times from the admissible set $\mathscr{F}$, and for each sampled $f_m \in \mathcal{F}$, compute $d(\mathcal{G}, f_m)$ and $d(f_{\text{base}}, f_m)$.

- Then
$$\hat{r} := \frac{\frac{1}{M}\sum_{m=1}^{M} d(\mathcal{G}, f_m)}{\frac{1}{M}\sum_{m=1}^{M} d(f_{\text{base}}, f_m)}$$
is an estimator for restrictiveness $r = r(\mathcal{G}, \mathscr{F})$.

- Under some regularity conditions, show the estimator is asymptotically normal, and show how to estimate its standard deviation.

Back to the CPT

Setting

The data: 25 binary lotteries $(\overline{z}, \underline{z}, p)$ over positive prizes, with 179 reported certainty equivalents per lottery

$\overline{z}$	$\underline{z}$	p	$f(\overline{z}, \underline{z}, p)$
20	0	0.25	17.04
40	10	0.95	39.45
$\vdots$	$\vdots$	$\vdots$	$\vdots$
150	50	0.05	73.99

Define the admissible set $\mathscr{F}$ to include all mappings that satisfy:

- First-order stochastic dominance (people prefer more money to less).
- Certainty equivalents fall in the range of the outcomes.

Models

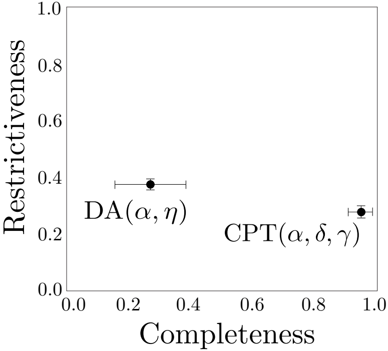
Cumulative Prospect Theory, henceforth CPT(α, δ, γ):

- utility of lottery $(\overline{z}, \underline{z}, p)$ is $w(p) \times v(\overline{z}) + (1 - w(p)) \times v(\underline{z})$, 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 $\tilde{w}(p) = \frac{p}{1+(1-p)\eta}$

Comparison of Models

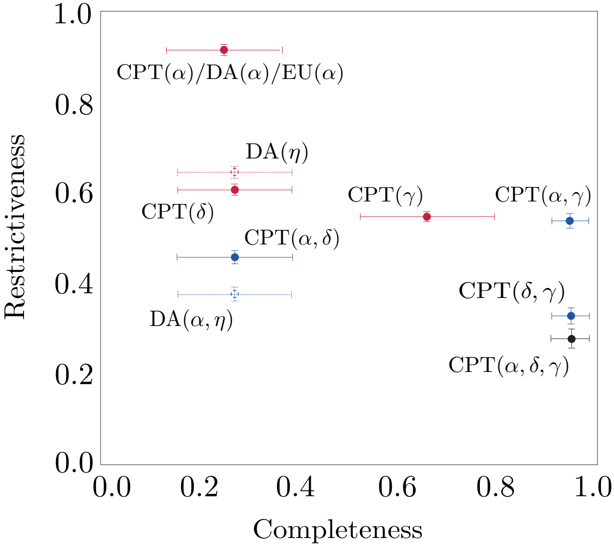


- CPT(α, δ, γ) is nearly complete but not very restrictive.
- This flexibility is not revealed by a simple count of the number of free parameters!
- DA(α, η) is more restrictive than CPT(α, δ, γ), but substantially less predictive of the real data.

The Value of a Parameter

Look at lower-parameter specifications of CPT and DA, e.g.

- Allow probability weighting but suppose that the agent is risk-neutral
- Shut down probability weighting but allow for risk aversion

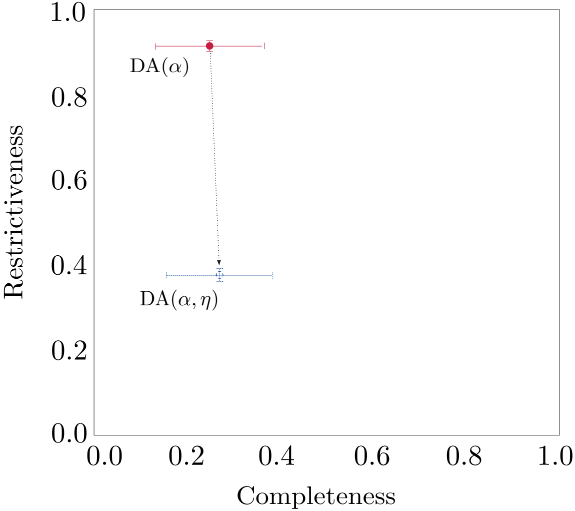

Restrictiveness
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Completeness
CPT(α)/DA(α)/EU(α)
DA(η)
CPT(δ)
CPT(γ)
CPT(α, γ)
CPT(α, δ)
DA(α, η)
CPT(δ, γ)
CPT(α, δ, γ)

Role of η in DA

Disappointment Aversion:

- $v(z) = z^\alpha$ is a value function over money
- $w(p) = \frac{p}{1+(1-p)\eta}$ is a probability weighting function
- η interpreted as disappointment aversion

Role of Disappointment Aversion Parameter η in DA

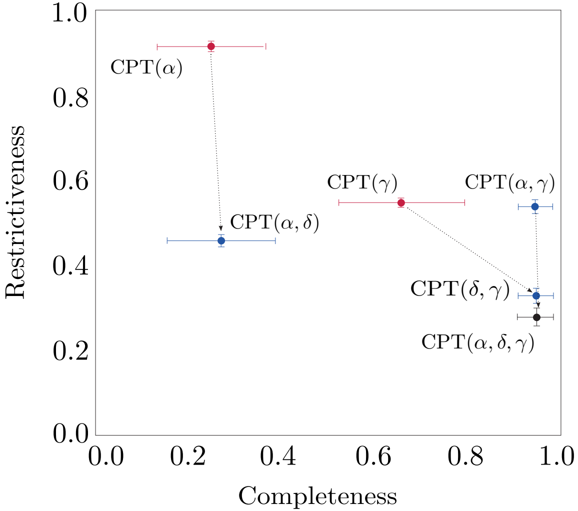


Role of Probability Weighting Parameter δ in CPT

Cumulative Prospect Theory:

- $v(z) = z^{\alpha}$ is a value function over money
- $w(p) = \frac{\delta p^{\gamma}}{\delta p^{\gamma} + (1-p)^{\gamma}}$ is a probability weighting function
- δ governs elevation of probability weighting function

Role of Probability Weighting Parameter δ in CPT

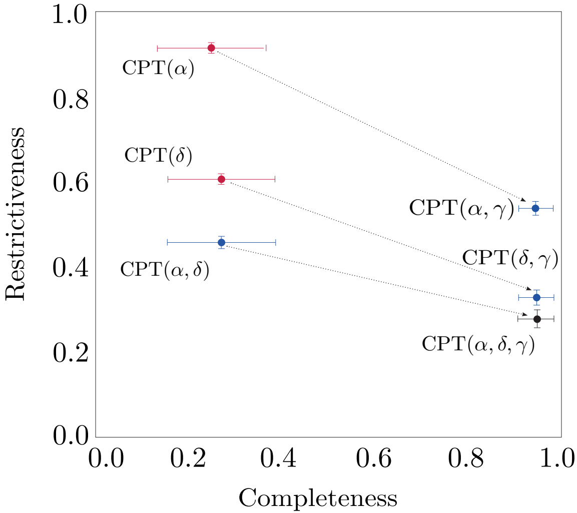


Role of Probability Weighting Parameter γ in CPT

Cumulative Prospect Theory:

- $v(z) = z^{\alpha}$ is a value function over money
- $w(p) = \frac{\delta p^{\gamma}}{\delta p^{\gamma} + (1-p)^{\gamma}}$ is a probability weighting function
- γ governs curvature of probability weighting function

Role of Probability Weighting Parameter γ in CPT



What We Learn

Not all parameters are equally effective.

- Adding the disappointment aversion parameter η or the curve elevation parameter δ

  $\longrightarrow$ large drop in restrictiveness

  $\longrightarrow$ small gain in completeness

- In contrast the curve slope parameter γ substantially increases completeness with similar or smaller decreases in restrictiveness. This suggests it captures real risk preferences.

Conclusion

- The completeness of a model compares how well it predicts to the “best possible” predictions.
- No point in looking for predictive improvements using the same features when the current theory is nearly complete.
- The same model (e.g. CPT) might have very different mean squared error on different data sets but be equally complete in all of them.
- Machine learning help uncover regularities that can be incorporated into new theories.
- ML can also help focus experimental effort on the most informative treatments.

- Restrictiveness helps distinguish theories that fit because they captures important regularities in the data from theories can provide a good approximation of any and all behavior.

- Can also help us understand the role of specific parameters: prefer parameters that add a lot of completeness (better fit to real data) w/o a big loss of restrictiveness (i.e. w/o also allowing a better fit to the synthetic data).

- We provide a practical, algorithmic approach for evaluating restrictiveness.

Thank You

Different Lottery Domain: Three Outcomes

- Evaluate CPT on a set of 18 three-outcome gain-domain lotteries from Bernheim and Sprenger [2020].
- Impose same background constraints as before: FOSD and range restriction.
- The restrictiveness of CPT on this set of three-outcome lotteries is 0.57

  $\longrightarrow$ CPT is about twice as restrictive on three-outcome lotteries as on binary lotteries.
- Besides imposing FOSD, CPT imposes the property of “rank dependence" for lotteries with more than two outcomes.
- We view the increase in restrictiveness as a quantification of the additional constraints implied by this property.