How Flexible is that Functional Form? Quantifying the Restrictiveness of Theories

Drew Fudenberg[1] Wayne Gao[2] Annie Liang[3]

[1]MIT

[2]UPenn [3]Northwestern

Introduction

When a model does a good job of fitting available data, we'd like to interpret this as evidence that the model is a good one.

Introduction

When a model does a good job of fitting available data, we'd like to interpret this as evidence that the model is a good one.

But another possibility is that the model is simply so flexible it would have fit any data.

- At an extreme: the model may not be falsifiable.

Introduction

When a model does a good job of fitting available data, we'd like to interpret this as evidence that the model is a good one.

But another possibility is that the model is simply so flexible it would have fit any data.

- At an extreme: the model may not be falsifiable.

To distinguish between these two explanations, need to know how **restrictive** the model is.

Our Contribution

- We propose a **measure** of the restrictiveness of an economic model that is easily computed across a variety of applications.

Our Contribution

- We propose a **measure** of the restrictiveness of an economic model that is easily computed across a variety of applications.
- Provide an **axiomatic foundation** for the measure

Our Contribution

- We propose a **measure** of the restrictiveness of an economic model that is easily computed across a variety of applications.
- Provide an **axiomatic foundation** for the measure
- Provide an **estimator** for the measure and characterization of its asymptotic distribution
  - Allows us to construct **confidence intervals**

Our Contribution

- We propose a **measure** of the restrictiveness of an economic model that is easily computed across a variety of applications.
- Provide an **axiomatic foundation** for the measure
- Provide an **estimator** for the measure and characterization of its asymptotic distribution
  - Allows us to construct **confidence intervals**
- Three **applications**:
  1. Certainty equivalents — lab data
  2. Initial play in games — lab/MTurk data
  3. Takeup of microfinance in Indian villages — field data

Motivating Example

Predicting Certainty Equivalents

Prediction problem:

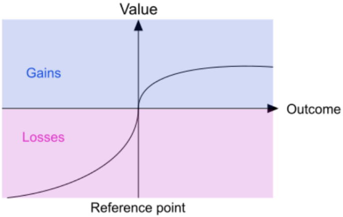
- Subject is offered a risky lottery:

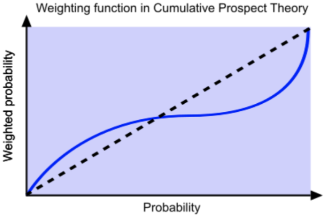
$$\begin{array}{ll} \overline{z} & \text{with probability } p \\ \underline{z} & \text{with probability } 1-p \end{array}$$

- What is the dollar amount x such that the subject would be indifferent between the lottery versus x dollars **for sure**?

Behavioral Model: Cumulative Prospect Theory

Cumulative Prospect Theory:





parameters α, β

parameters δ, γ

Testing CPT

- We evaluate CPT on data from Bruhin et al (2001): 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.

Testing CPT

- We evaluate CPT on data from Bruhin et al (2001): 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.
- **Benchmark.** Because we have a large number of reports per lottery, can estimate $\mathbb{E}[Y \mid X]$ (i.e., the best predictor).
  - Lower bound for what is achievable by CPT.

CPT Predicts Very Well

	(Mean Squared) Error
Best Possible	65.58

CPT Predicts Very Well

	(Mean Squared) Error
Expected Payoff	103.81
Best Possible	65.58

CPT Predicts Very Well

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

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

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

  ↪ CPT well describes the structure in perception of risk

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

  ↪ CPT well describes the structure in perception of risk

  or

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

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

  ↪ CPT well describes the structure in perception of risk

  or

  ↪ CPT is 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** versus when it is simply **unrestrictive**.

Our Approach

Our approach for measuring model restrictiveness:

- Generate synthetic data sets
- See how well the model performs on each of these
- An unrestrictive model performs well on all data sets

Our Approach

Our approach for measuring model restrictiveness:

- Generate synthetic data sets
- See how well the model performs on each of these
- An unrestrictive model performs well on all data sets
- Define restrictiveness to be the **(normalized) average error** for predicting these synthetic data sets
    - Ranges between zero and 1

Approach

Setting

- X belonging to a finite set $\mathcal{X}$ is an observable feature vector
  - in example: description of the lottery $X = (\overline{z}, \underline{z}; p, 1-p)$
- Y belonging to $\mathcal{Y} \subseteq \mathbb{R}^k$ is an outcome
  - in example, Y is a reported certainty equivalent by a subject

Setting

- X belonging to a finite set $\mathcal{X}$ is an observable feature vector
  - in example: description of the lottery $X = (\overline{z}, \underline{z}; p, 1-p)$
- Y belonging to $\mathcal{Y} \subseteq \mathbb{R}^k$ is an outcome
  - in example, Y is a reported certainty equivalent by a subject
- Any $f : \mathcal{X} \to \mathcal{Y}$ is a (predictive) mapping.
  - e.g., a mapping from lotteries into certainty equivalents

Setting

- X belonging to a finite set $\mathcal{X}$ is an observable feature vector
  - in example: description of the lottery $X = (\overline{z}, \underline{z}; p, 1 - p)$
- Y belonging to $\mathcal{Y} \subseteq \mathbb{R}^k$ is an outcome
  - in example, Y is a reported certainty equivalent by a subject
- Any $f : \mathcal{X} \to \mathcal{Y}$ is a (predictive) mapping.
  - e.g., a mapping from lotteries into certainty equivalents
- Let $d(f, f')$ be an appropriate measure for "how different" predictions are under f and f'.
  - e.g. expected squared distance: $d(f, f') = \mathbb{E}[(f(X) - f'(X))^2]$

Setting

- X belonging to a finite set $\mathcal{X}$ is an observable feature vector
  - in example: description of the lottery $X = (\overline{z}, \underline{z}; p, 1-p)$
- Y belonging to $\mathcal{Y} \subseteq \mathbb{R}^k$ is an outcome
  - in example, Y is a reported certainty equivalent by a subject
- Any $f : \mathcal{X} \to \mathcal{Y}$ is a (predictive) mapping.
  - e.g., a mapping from lotteries into certainty equivalents
- Let $d(f, f')$ be an appropriate measure for "how different" predictions are under f and f'.
  - e.g. expected squared distance: $d(f, f') = \mathbb{E}[(f(X) - f'(X))^2]$
- Let $\mathcal{G} = \{g_\theta\}_{\theta \in \Theta}$ be a parametric economic model.
  - e.g., $\mathcal{G}$ is CPT with $\theta = (\alpha, \beta, \delta, \gamma)$.

Approach to Measuring Restrictiveness

Step 1: Define an "admissible set" $\mathscr{F}$ of 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

Approach to Measuring Restrictiveness

Step 1: Define an "admissible set" $\mathscr{F}$ of 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

Step 2: Sample mappings uniformly at random from $\mathscr{F}$ and evaluate how well the model $\mathcal{G}$ approximates these mappings.

- The model's approximation error to each generated mapping f is $d(\mathcal{G}, f) \equiv \min_{f' \in \mathcal{G}} d(f', f)$.

Approach to Measuring Restrictiveness

Step 1: Define an "admissible set" $\mathscr{F}$ of 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

Step 2: Sample mappings uniformly at random from $\mathscr{F}$ and evaluate how well the model $\mathcal{G}$ approximates these mappings.

- The model's approximation error to each generated mapping f is $d(\mathcal{G}, f) \equiv \min_{f' \in \mathcal{G}} d(f', f)$.

Step 3: Choose a baseline mapping f_{base} from the model $\mathcal{G}$ and evaluate its approximation error to randomly drawn mappings.

- e.g., in lottery example, let f_{base} predict the lottery's expected payoff.

Restrictiveness

The **restrictiveness** of the model $\mathcal{G}$ wrt the admissible set $\mathscr{F}$ is

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

where the expectations are with respect to a uniform distribution on the admissible set $\mathscr{F}$.

Restrictiveness

The **restrictiveness** of the model $\mathcal{G}$ wrt the admissible set $\mathscr{F}$ is

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

where the expectations are 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.

- ranges from zero (completely unrestrictive) to 1 (approximates admissible mappings no better than f_{base} does)

Restrictiveness

The **restrictiveness** of the model $\mathcal{G}$ wrt the admissible set $\mathscr{F}$ is

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

where the expectations are 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.

- ranges from zero (completely unrestrictive) to 1 (approximates admissible mappings no better than f_{base} does)
- measure is unitless (thanks to the normalization)

Restrictiveness

The **restrictiveness** of the model $\mathcal{G}$ wrt the admissible set $\mathscr{F}$ is

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

where the expectations are 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.

- ranges from zero (completely unrestrictive) to 1 (approximates admissible mappings no better than f_{base} does)
- measure is unitless (thanks to the normalization)

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

Complementary Measure: Completeness

Use the definition of **completeness** from Fudenberg, Kleinberg, Liang, and Mullainathan (forthcoming JPE).

Complementary Measure: Completeness

Use the definition of **completeness** from Fudenberg, Kleinberg, Liang, and Mullainathan (forthcoming JPE).

- Let P^* denote the distribution of (X, Y) in the **real** data.

Complementary Measure: Completeness

Use the definition of **completeness** from Fudenberg, Kleinberg, Liang, and Mullainathan (forthcoming JPE).

- Let P^* denote the distribution of (X, Y) in the **real** data.
- Choose a loss function $\ell(Y, Y')$, e.g., $\ell(Y, Y') = (Y - Y')^2$

Complementary Measure: Completeness

Use the definition of **completeness** from Fudenberg, Kleinberg, Liang, and Mullainathan (forthcoming JPE).

- Let P^* denote the distribution of (X, Y) in the **real** data.
- Choose a loss function $\ell(Y, Y')$, e.g., $\ell(Y, Y') = (Y - Y')^2$
- The **error** of mapping f is $e(f) := \mathbb{E}_{P^*}[\ell(f(X), Y)]$

Complementary Measure: Completeness

Use the definition of **completeness** from Fudenberg, Kleinberg, Liang, and Mullainathan (forthcoming JPE).

- Let P^* denote the distribution of (X, Y) in the **real** data.
- Choose a loss function $\ell(Y, Y')$, e.g., $\ell(Y, Y') = (Y - Y')^2$
- The **error** of mapping f is $e(f) := \mathbb{E}_{P^*}[\ell(f(X), Y)]$
- The best achievable error is $\min_{f \in \mathscr{F}^*} e(f)$, where $\mathscr{F}^*$ denotes the set of all mappings $f : \mathcal{X} \to \mathcal{Y}$

Complementary Measure: Completeness

Use the definition of **completeness** from Fudenberg, Kleinberg, Liang, and Mullainathan (forthcoming JPE).

- Let P^* denote the distribution of (X, Y) in the **real** data.
- Choose a loss function $\ell(Y, Y')$, e.g., $\ell(Y, Y') = (Y - Y')^2$
- The **error** of mapping f is $e(f) := \mathbb{E}_{P^*}[\ell(f(X), Y)]$
- The best achievable error is $\min_{f \in \mathscr{F}^*} e(f)$, where $\mathscr{F}^*$ denotes the set of all mappings $f : \mathcal{X} \to \mathcal{Y}$
- The completeness of model $\mathcal{G}$ is

$$\kappa(\mathcal{G}) := \frac{e(f_{\text{base}}) - \min_{f \in \mathcal{G}} e(f)}{e(f_{\text{base}}) - \min_{f \in \mathscr{F}^*} e(f)}.$$

  i.e., the fraction of the achievable reduction in error (relative to the baseline mapping) that is achieved by the model

Complementary Measure: Completeness

Use the definition of **completeness** from Fudenberg, Kleinberg, Liang, and Mullainathan (forthcoming JPE).

- Let P^* denote the distribution of (X, Y) in the **real** data.
- Choose a loss function $\ell(Y, Y')$, e.g., $\ell(Y, Y') = (Y - Y')^2$
- The **error** of mapping f is $e(f) := \mathbb{E}_{P^*}[\ell(f(X), Y)]$
- The best achievable error is $\min_{f \in \mathscr{F}^*} e(f)$, where $\mathscr{F}^*$ denotes the set of all mappings $f : \mathcal{X} \to \mathcal{Y}$
- The completeness of model $\mathcal{G}$ is

$$\kappa(\mathcal{G}) := \frac{e(f_{\text{base}}) - \min_{f \in \mathcal{G}} e(f)}{e(f_{\text{base}}) - \min_{f \in \mathscr{F}^*} e(f)}.$$

  i.e., the fraction of the achievable reduction in error (relative to the baseline mapping) that is achieved by the model
- For certain "paired" choices of loss ℓ and distance d, restrictiveness $=$ 1 - expected completeness of the model on synthetic data.

Restrictiveness and Completeness

Restrictiveness:

- Ranges from zero to 1
- Computed from synthetic data
- Larger values mean that the model imposes more restrictions.

Completeness:

- Ranges from zero to 1
- Computed from real data
- Larger values implies a model that predicts real data better.

Prefer models that have high completeness (good fit to real data) and high restrictiveness (poor fit to synthetic data).

Related Literature

Koopmans & Reiersol (1950) introduce (binary) notion of "observationally restrictive," here we ask **how** restrictive

Related Literature

Koopmans & Reiersol (1950) introduce (binary) notion of "observationally restrictive," here we ask **how** restrictive

Our measures relate to but differ from large literature on model selection (Cox, 1961; Cox, 1962)

- objective of these approaches is typically to avoid overfitting
- typically trade off some notion of **fit to real data** against some notion of **complexity**, e.g. AIC combines
  - log-likelihood of the data (similar to completeness)
  - number of parameters (similar to restrictiveness)

Related Literature

Koopmans & Reiersol (1950) introduce (binary) notion of "observationally restrictive," here we ask **how** restrictive

Our measures relate to but differ from large literature on model selection (Cox, 1961; Cox, 1962)

- objective of these approaches is typically to avoid overfitting
- typically trade off some notion of **fit to real data** against some notion of **complexity**, e.g. AIC combines
  - log-likelihood of the data (similar to completeness)
  - number of parameters (similar to restrictiveness)
- penalty on complexity shrinks as sample grows large
- we assume an intrinsic preference for parsimonious models (even with infinite data!)

Related Literature

Our measures is also related to complexity measures from CS (e.g., VC dimension), but our measure is substantially easier to compute

Representation theorems (e.g. in decision theory) characterize empirical content of models

- don't have such theorems for many models, and often not for functional forms used in applied work

Plan for Rest of Talk

1. Axiomatic foundation for our restrictiveness measure
2. Estimators for restrictiveness and completeness
3. Return to first application (certainty equivalents)
4. Brief summary of remaining two applications (initial play in games and takeup of microfinance)

Axiomatic Foundation

Approximation Error

- Recall $\mathscr{F}^*$ is the set of all mappings $f : \mathcal{X} \to \mathcal{Y}$.

Approximation Error

- Recall $\mathscr{F}^*$ is the set of all mappings $f : \mathcal{X} \to \mathcal{Y}$.
- Endow $\mathscr{F}^*$ with the Lebesgue σ-algebra and a σ-finite measure μ, which can be interpreted as the analyst's prior.

Approximation Error

- Recall $\mathscr{F}^*$ is the set of all mappings $f : \mathcal{X} \to \mathcal{Y}$.
- Endow $\mathscr{F}^*$ with the Lebesgue σ-algebra and a σ-finite measure μ, which can be interpreted as the analyst's prior.
- We define an *approximation error* e that takes as input
  - the model $\mathcal{G} \subseteq \mathscr{F}^*$
  - a Lebesgue-measurable set of admissible mappings $\mathscr{F} \subseteq \mathscr{F}^*$
  - a distance $d : \mathscr{F}^* \times \mathscr{F}^* \to \mathbb{R}$

  and outputs a real number.

Approximation Error

- Recall $\mathscr{F}^*$ is the set of all mappings $f : \mathcal{X} \to \mathcal{Y}$.
- Endow $\mathscr{F}^*$ with the Lebesgue σ-algebra and a σ-finite measure μ, which can be interpreted as the analyst's prior.
- We define an *approximation error* e that takes as input
    - the model $\mathcal{G} \subseteq \mathscr{F}^*$
    - a Lebesgue-measurable set of admissible mappings $\mathscr{F} \subseteq \mathscr{F}^*$
    - a distance $d : \mathscr{F}^* \times \mathscr{F}^* \to \mathbb{R}$

  and outputs a real number.

$e(\mathcal{G}, \mathscr{F}, d)$ is model $\mathcal{G}$'s approximation error to the admissible set $\mathscr{F}$, where the quality of the approximation is measured using d

Axioms

Axiom 1 (Nonnegativity). For every model $\mathcal{G}$, admissible set $\mathscr{F}$, and distance d, $e(\mathcal{G}, \mathscr{F}, d) \geq 0$.

Approximation error is always nonnegative.

Axioms

Axiom 1 (Nonnegativity). For every model $\mathcal{G}$, admissible set $\mathscr{F}$, and distance d, $e(\mathcal{G}, \mathscr{F}, d) \geq 0$.

Approximation error is always nonnegative.

Axiom 2 (Monotonicity). Fix any admissible set $\mathscr{F}$. If the models $\mathcal{G}_1$ and $\mathcal{G}_2$ satisfy

$$d(\mathcal{G}_1, f) \geq d(\mathcal{G}_2, f) \quad \forall f \in \mathscr{F}$$

then $e(\mathcal{G}_1, \mathscr{F}, d) \geq e(\mathcal{G}_2, \mathscr{F}, d)$.

If $\mathcal{G}_1$ can approximate every admissible mapping better than $\mathcal{G}_2$, then $\mathcal{G}_1$ has lower approximation error.

Axioms

Axiom 3 (Rescaling of Units).

(a) Fix any model $\mathcal{G}$, admissible set $\mathscr{F}$, and distance d. Then

$$e(\mathcal{G}, \mathscr{F}, \alpha \cdot d) = \alpha \cdot e(\mathcal{G}, \mathscr{F}, d) \quad \forall\, \alpha \in \mathbb{R}_+$$

Rescaling of the units of d is inherited by the approximation error measure.

Axioms

Axiom 3 (Rescaling of Units).

(a) Fix any model $\mathcal{G}$, admissible set $\mathscr{F}$, and distance d. Then

$$e(\mathcal{G}, \mathscr{F}, \alpha \cdot d) = \alpha \cdot e(\mathcal{G}, \mathscr{F}, d) \quad \forall\, \alpha \in \mathbb{R}_+$$

Rescaling of the units of d is inherited by the approximation error measure.

(b) Fix any admissible set $\mathscr{F}$ and distance d. If $\mathcal{G}_1$ and $\mathcal{G}_2$ satisfy

$$d(\mathcal{G}_1, f) = \alpha \cdot d(\mathcal{G}_2, f) \quad \forall\, f \in \mathscr{F},$$

then $e(\mathcal{G}_1, \mathscr{F}, d) = e(\mathcal{G}_2, \mathscr{F}, \alpha \cdot d)$.

Scaling the distance between $\mathcal{G}$ and each mapping f leads to the same value of approximation error as scaling the units of d.

Axioms

Axiom 4 (Linearity). For any countable sequence of disjoint measurable sets $\mathscr{F}_1, \mathscr{F}_2, \ldots$ whose union $\mathscr{F} \equiv \cup_{i=1}^{\infty} \mathscr{F}_i$ has strictly positive measure,

$$e(\mathcal{G}, \mathscr{F}, d) = \sum_{i=1}^{\infty} \frac{\mu(\mathscr{F}_i)}{\mu(\mathscr{F})} \cdot e(\mathcal{G}, \mathscr{F}_i, d) \quad \forall \mathcal{G}, d.$$

Axioms

Axiom 4 (Linearity). For any countable sequence of disjoint measurable sets $\mathscr{F}_1, \mathscr{F}_2, \dots$ whose union $\mathscr{F} \equiv \cup_{i=1}^{\infty} \mathscr{F}_i$ has strictly positive measure,

$$e(\mathcal{G}, \mathscr{F}, d) = \sum_{i=1}^{\infty} \frac{\mu(\mathscr{F}_i)}{\mu(\mathscr{F})} \cdot e(\mathcal{G}, \mathscr{F}_i, d) \quad \forall \mathcal{G}, d.$$

Consider constraining the admissible set $\mathscr{F}$ to a subset $\mathscr{F}_1$ or its complement $\mathscr{F}_2$.

- Ex ante approximation error $e(\mathcal{G}, \mathscr{F}, d)$ is a convex combination of the ex post approximation errors $e(\mathcal{G}, \mathscr{F}_1, d)$ or $e(\mathcal{G}, \mathscr{F}_2, d)$.
- Each ex post subset contributes to the ex ante approximation error in proportion to its measure.

Axioms

Axiom 5 (Symmetry). Fix any admissible set $\mathscr{F}$ and any bijection τ from $\mathscr{F}$ to itself. Consider two models $\mathcal{G}_1$ and $\mathcal{G}_2$ where

$$d(\mathcal{G}_1, f) = d(\mathcal{G}_2, \tau(f)) \quad \forall f \in \mathscr{F}.$$

Then $e(\mathcal{G}_1, \mathscr{F}, d) = e(\mathcal{G}_2, \mathscr{F}, d)$.

Permuting the various discrepancies between the model and the admissible mappings f does not affect the overall approximation error.

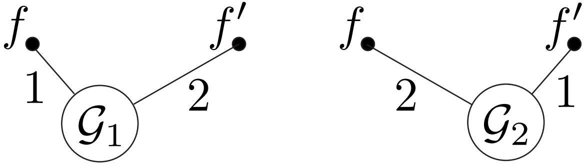
Axioms

Axiom 5 (Symmetry). Fix any admissible set $\mathscr{F}$ and any bijection τ from $\mathscr{F}$ to itself. Consider two models $\mathcal{G}_1$ and $\mathcal{G}_2$ where

$$d(\mathcal{G}_1, f) = d(\mathcal{G}_2, \tau(f)) \quad \forall f \in \mathscr{F}.$$

Then $e(\mathcal{G}_1, \mathscr{F}, d) = e(\mathcal{G}_2, \mathscr{F}, d)$.

Permuting the various discrepancies between the model and the admissible mappings f does not affect the overall approximation error.



Characterization of Approximation Error

Proposition

An approximation error e satisfies Axioms 1-5 if and only if

$$e(\mathcal{G}, \mathscr{F}, d) = \mathbb{E}_{f \sim Unif(\mathscr{F})}\left[c \cdot d(\mathcal{G}, f)\right] \quad \forall \mathcal{G}, \mathscr{F}, d$$

for a positive constant c, where Unif($\mathscr{F}$) denotes the uniform distribution on $\mathscr{F}$.

Characterization of Approximation Error

Proposition

An approximation error e satisfies Axioms 1-5 if and only if

$$e(\mathcal{G}, \mathscr{F}, d) = \mathbb{E}_{f \sim Unif(\mathscr{F})}\left[c \cdot d(\mathcal{G}, f)\right] \quad \forall \mathcal{G}, \mathscr{F}, d$$

for a positive constant c, where Unif($\mathscr{F}$) denotes the uniform distribution on $\mathscr{F}$.

- Dropping A5 (Symmetry) returns a broader class of restrictiveness measures with respect to different distributions.

Characterization of Approximation Error

Proposition

An approximation error e satisfies Axioms 1-5 if and only if

$$e(\mathcal{G}, \mathscr{F}, d) = \mathbb{E}_{f \sim Unif(\mathscr{F})}\left[c \cdot d(\mathcal{G}, f)\right] \quad \forall \mathcal{G}, \mathscr{F}, d$$

for a positive constant c, where Unif($\mathscr{F}$) denotes the uniform distribution on $\mathscr{F}$.

- Dropping A5 (Symmetry) returns a broader class of restrictiveness measures with respect to different distributions.
- We prefer the uniform distribution in our applications:
  - Principle of indifference
  - Avoids cherry-picking
  - Transparent and easy to interpret

Restrictiveness

Approximation error:

$$e(\mathcal{G}, \mathscr{F}, d) = \mathbb{E}_{f \sim \mathsf{Unif}(\mathscr{F})} [c \cdot d(\mathcal{G}, f)]$$

Our restrictiveness measure normalizes this approximation error relative to the approximation error of a baseline mapping f_{base}.

$$r(\mathcal{G}, \mathscr{F}, d) := \frac{\mathbb{E}_{f \sim \mathsf{Unif}(\mathscr{F})}[d(\mathcal{G}, f)]}{\mathbb{E}_{f \sim \mathsf{Unif}(\mathscr{F})}[d(f_{\text{base}}, f)]}$$

Generalization of Selten (1991)

- Selten proposed measuring the flexibility of a model by the fraction of possible data sets that it can **exactly explain**.

Generalization of Selten (1991)

- Selten proposed measuring the flexibility of a model by the fraction of possible data sets that it can **exactly explain**.
- This is equivalent to $1 - e(\mathcal{G}, \mathscr{F}, d)$ where $\mathscr{F}$ is unrestricted and

$$d(f, f') = \begin{cases} 0 & \text{if } f = f' \\ 1 & \text{if } f \neq f' \end{cases}$$

Generalization of Selten (1991)

- Selten proposed measuring the flexibility of a model by the fraction of possible data sets that it can **exactly explain**.
- This is equivalent to $1 - e(\mathcal{G}, \mathscr{F}, d)$ where $\mathscr{F}$ is unrestricted and

$$d(f, f') = \begin{cases} 0 & \text{if } f = f' \\ 1 & \text{if } f \neq f' \end{cases}$$

- We focus on approximate fit, which broadens the applicability of the measure, and can lead to very different conclusions:
  - 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 when d is squared distance.

Estimators

Estimators

Restrictiveness:
$$r(\mathcal{G}) := \frac{\mathbb{E}_{f\sim\text{Unif}(\mathscr{F})}[d(\mathcal{G}, f)]}{\mathbb{E}_{f\sim\text{Unif}(\mathscr{F})}[d(f_{\text{base}}, f)]}$$

- Sample M times uniformly at random from admissible set $\mathscr{F}$.
- Use sample averages in place of expectations

Estimators

Restrictiveness: $$r(\mathcal{G}) := \frac{\mathbb{E}_{f \sim \mathrm{Unif}(\mathscr{F})}[d(\mathcal{G}, f)]}{\mathbb{E}_{f \sim \mathrm{Unif}(\mathscr{F})}[d(f_{\text{base}}, f)]}$$

- Sample M times uniformly at random from admissible set $\mathscr{F}$.
- Use sample averages in place of expectations

Completeness: $$\kappa(\mathcal{G}) := \frac{e(f_{\text{base}}) - \min_{f \in \mathcal{G}} e(f)}{e(f_{\text{base}}) - \min_{f \in \mathscr{F}^*} e(f)}$$

Use tenfold cross-validation.

- Split data into $K = 10$ folds. Train free parameters of the model on $K - 1$ folds, test estimated model on the final fold
- Average over the K possible choices of the test fold.

Estimators

Restrictiveness:

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

- Sample M times uniformly at random from admissible set $\mathscr{F}$.
- Use sample averages in place of expectations

Completeness:

$$\kappa(\mathcal{G}) := \frac{e(f_{\text{base}}) - \min_{f\in\mathcal{G}} e(f)}{e(f_{\text{base}}) - \min_{f\in\mathscr{F}^*} e(f)}$$

Use tenfold cross-validation.

- Split data into $K = 10$ folds. Train free parameters of the model on $K - 1$ folds, test estimated model on the final fold
- Average over the K possible choices of the test fold.

Estimators are asymptotically normal, see paper for their asymptotic variance (which we use to compute standard errors).

Back to Application 1:

Certainty Equivalents

Setting

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

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	
40	10	0.95	
⋮	⋮	⋮	
150	50	0.05	

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	15.96
40	10	0.95	18.58
⋮	⋮	⋮	⋮
150	50	0.05	83.71

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
⋮	⋮	⋮	⋮
150	50	0.05	73.99

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

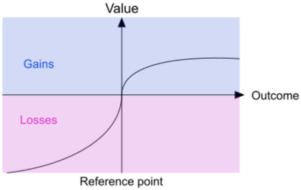
Define the admissible set $\mathscr{F}$ to include all mappings that satisfy:

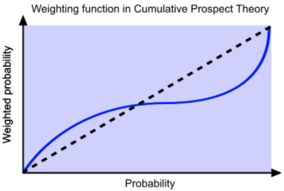
- first-order stochastic dominance (people like more money)
- certainty equivalents fall in the range of the outcomes

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





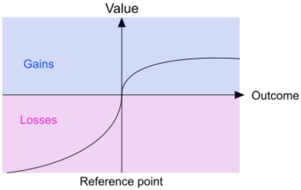
Models

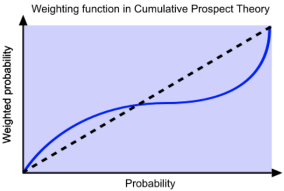
Cumulative Prospect Theory, henceforth CPT(α, δ, γ):

- 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

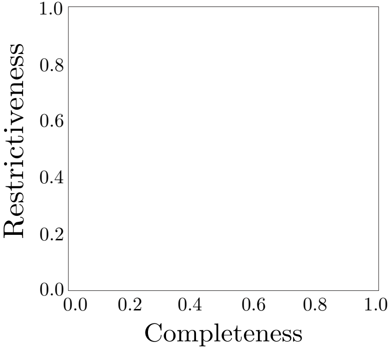
Disappointment Aversion (Gul, 1991), henceforth DA(α, η):

- same as above, except that the probability weighting function is $\tilde{w}(p) = \frac{p}{1+(1-p)\eta}$

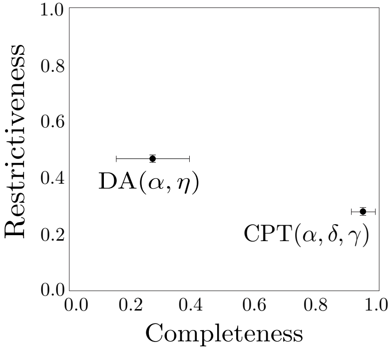




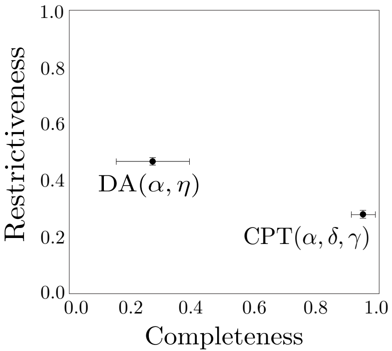
Comparison of Models



Comparison of Models

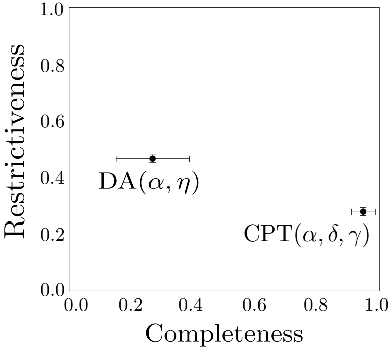


Comparison of Models



- CPT(α, δ, γ) is nearly complete but not very restrictive.

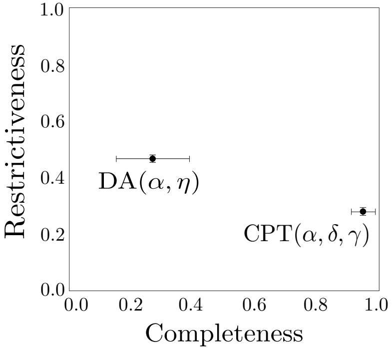
Comparison of Models



- 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!

Comparison of Models



- CPT(α, δ, γ) is nearly complete but not very restrictive.

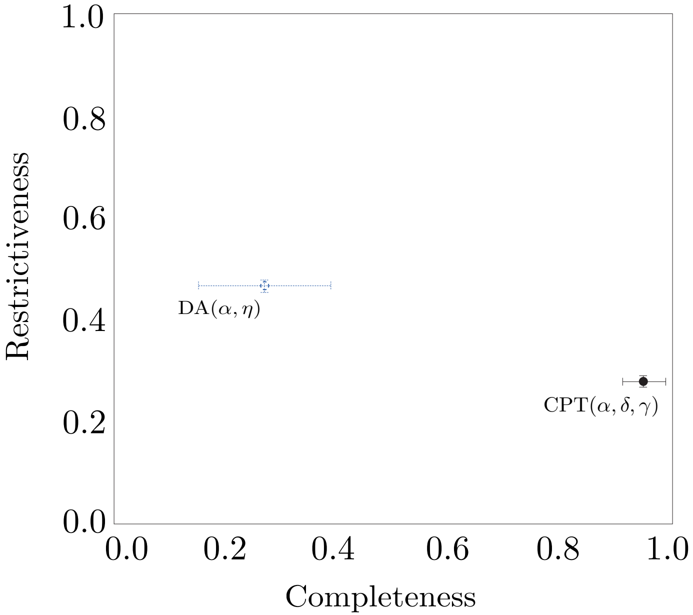
  This flexibility is not revealed by a simple count of the number of free parameters!

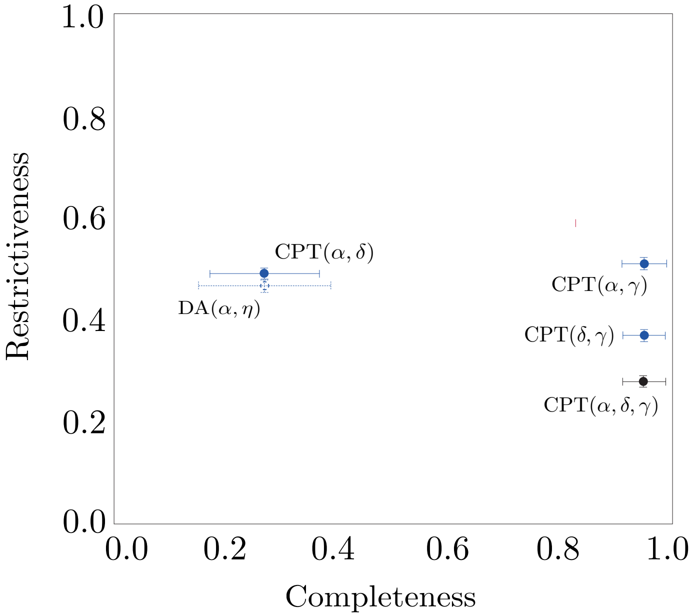
- DA(α, η) is more restrictive than CPT(α, δ, γ), but substantially less predictive of the real data.

The Value of a Parameter

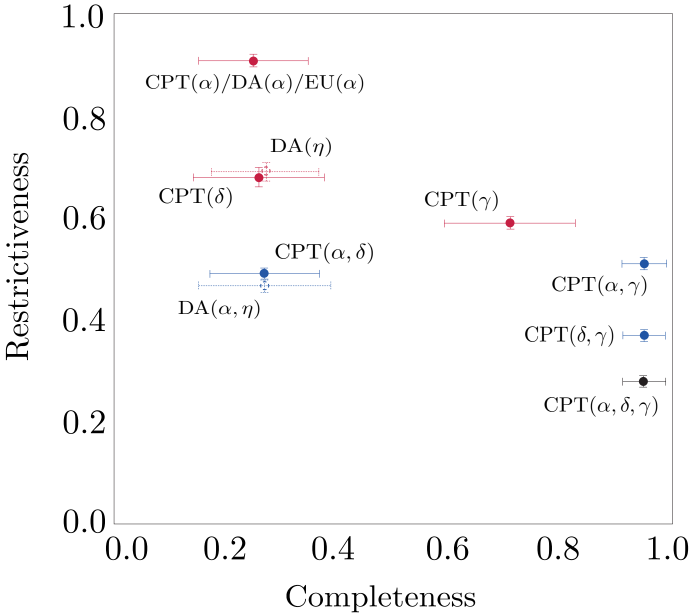
Can further evaluate the value of a parameter by looking at lower-parameter specifications of CPT and DA, e.g.

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



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
CPT(α, δ)
DA(α, η)
CPT(α, γ)
CPT(δ, γ)
CPT(α, δ, γ)



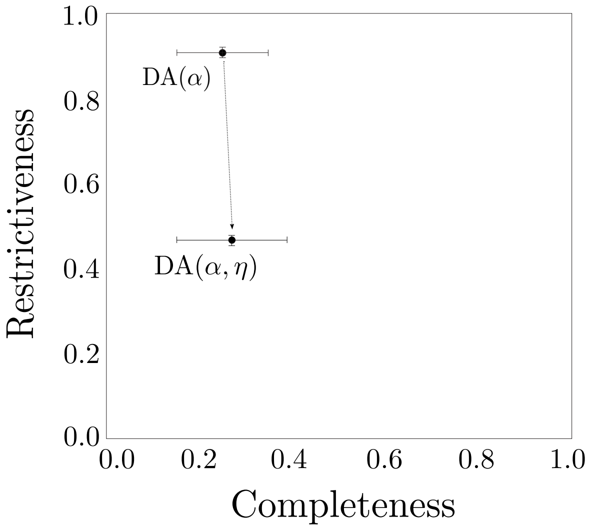
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Restrictiveness
Completeness
CPT(α)/DA(α)/EU(α)
DA(η)
CPT(δ)
CPT(γ)
CPT(α, δ)
DA(α, η)
CPT(α, γ)
CPT(δ, γ)
CPT(α, δ, γ)

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

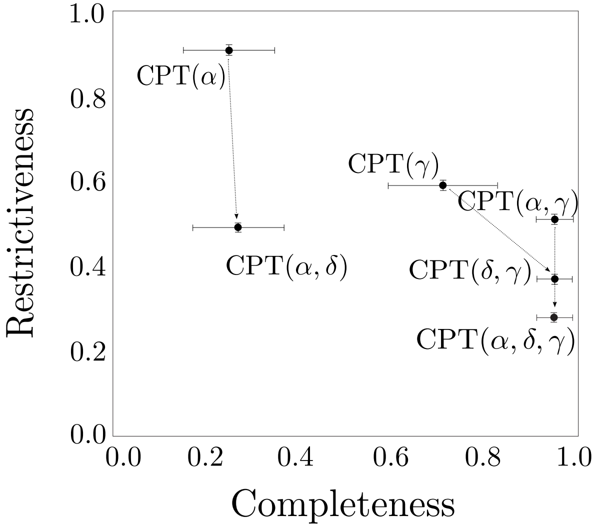


Role of Probability Weighting Parameter δ in CPT

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

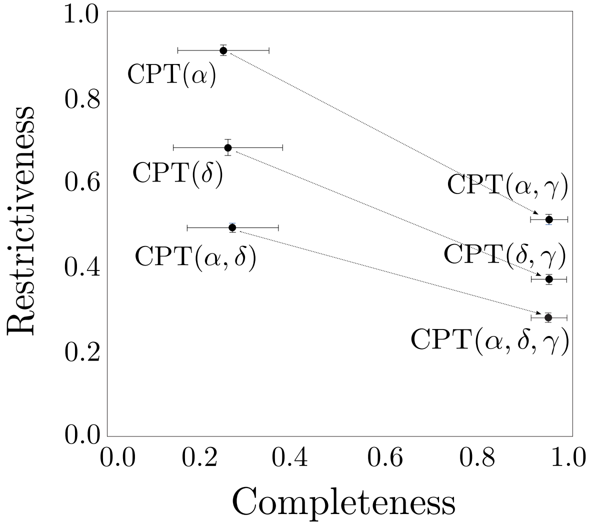


Role of Probability Weighting Parameter γ in CPT

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

- On the other hand, the curve slope parameter γ plays an important role in capturing real risk preferences.

Summary of

Applications 2 and 3

Application 2: Initial Play in Games

- Predict distribution of initial play in 466 3×3 matrix games from Fudenberg and Liang (2019)

Application 2: Initial Play in Games

- Predict distribution of initial play in 466 3×3 matrix games from Fudenberg and Liang (2019)
- Evaluate three models:
  - Poisson Cognitive Hierarchy Model (PCHM)
  - Logit PCHM (i.e., PCHM with logistic best replies)
  - Logit Level-1 (logit choice model where the "utility" of each action is its expected payoff against uniform play)

Application 2: Initial Play in Games

- Predict distribution of initial play in 466 3×3 matrix games from Fudenberg and Liang (2019)

- Evaluate three models:
  - Poisson Cognitive Hierarchy Model (PCHM)
  - Logit PCHM (i.e., PCHM with logistic best replies)
  - Logit Level-1 (logit choice model where the "utility" of each action is its expected payoff against uniform play)

- To evaluate restrictiveness, we generate synthetic distributions of play for each of the 466 games, satisfying:
  - Any strictly dominated action is played with frequency $\leq 1/3$
  - Any strictly dominant action is played with frequency $\geq 1/3$

Restrictiveness and Completeness of Each Model

	# Param	Restrictiveness	Completeness
PCHM	1	0.992	0.436
		(<0.001)	(0.017)
logit level-1	1	0.970	0.727
		(<0.001)	(0.015)
logit PCHM	2	0.971	0.729
		(0.003)	(0.014)

Restrictiveness and Completeness of Each Model

	# Param	Restrictiveness	Completeness
PCHM	1	0.992	0.436
		(<0.001)	(0.017)
logit level-1	1	0.970	0.727
		(<0.001)	(0.015)
logit PCHM	2	0.971	0.729
		(0.003)	(0.014)

- All three models are very restrictive.

Restrictiveness and Completeness of Each Model

	# Param	Restrictiveness	Completeness
PCHM	1	0.992	0.436
		(<0.001)	(0.017)
logit level-1	1	0.970	0.727
		(<0.001)	(0.015)
logit PCHM	2	0.971	0.729
		(0.003)	(0.014)

- All three models are very restrictive.
- Logit Level-1 and Logit PCHM are nearly twice as complete as PCHM, but only slightly less restrictive.

Restrictiveness and Completeness of Each Model

	# Param	Restrictiveness	Completeness
PCHM	1	0.992	0.436
		(<0.001)	(0.017)
logit level-1	1	0.970	0.727
		(<0.001)	(0.015)
logit PCHM	2	0.971	0.729
		(0.003)	(0.014)

- All three models are very restrictive.
- Logit Level-1 and Logit PCHM are nearly twice as complete as PCHM, but only slightly less restrictive.
- Despite having different parameter counts, Logit Level-1 and Logit PCHM are nearly identically restrictive and complete.

Application 3: Network Diffusion

- Setting: Prediction of microfinance takeup rates following diffusion of information in social networks.

Application 3: Network Diffusion

- Setting: Prediction of microfinance takeup rates following diffusion of information in social networks.
- We use data from Banerjee et al (2019):
  - Certain "leaders" in 43 villages in Karnatka, India were given information about a microfinance program
  - Outcome variable is average takeup of the program after a period of time

Application 3: Network Diffusion

- Setting: Prediction of microfinance takeup rates following diffusion of information in social networks.
- We use data from Banerjee et al (2019):
  - Certain "leaders" in 43 villages in Karnatka, India were given information about a microfinance program
  - Outcome variable is average takeup of the program after a period of time
- We evaluate two kinds of models:

  OLS regressions based on different sets of network statistics.

  Structural model of information diffusion based on Banerjee et al (2019).

  - Individuals pass info onto their neighbors, and are more likely to take up the more they hear about it.

Restrictiveness and Completeness of Each Model

	# Param	Restrictiveness	Completeness
Linear Models			
Eigenvector Centrality of Leaders	1	0.9762	0.2577
		(0.0003)	(0.1101)

Restrictiveness and Completeness of Each Model

	# Param	Restrictiveness	Completeness
Linear Models			
Eigenvector Centrality of Leaders	1	0.9762	0.2577
		(0.0003)	(0.1101)
+ Degree Centrality of Leaders	2	0.9526	0.3385
		(0.0004)	(0.1193)

Restrictiveness and Completeness of Each Model

	# Param	Restrictiveness	Completeness
Linear Models			
Eigenvector Centrality of Leaders	1	0.9762	0.2577
		(0.0003)	(0.1101)
+ Degree Centrality of Leaders	2	0.9526	0.3385
		(0.0004)	(0.1193)
+ Degree Centrality of All Villagers	3	0.9288	0.3471
		(0.0005)	(0.1151)

Restrictiveness and Completeness of Each Model

	# Param	Restrictiveness	Completeness
Linear Models			
Eigenvector Centrality of Leaders	1	0.9762	0.2577
		(0.0003)	(0.1101)
+ Degree Centrality of Leaders	2	0.9526	0.3385
		(0.0004)	(0.1193)
+ Degree Centrality of All Villagers	3	0.9288	0.3471
		(0.0005)	(0.1151)
+ Betweenness Centrality of Leaders	4	0.9053	0.3475
		(0.0006)	(0.1158)
+ Clustering Coefficient	5	0.8816	0.3516
		(0.0007)	(0.1191)
+ Average Path Length	6	0.8579	0.3516
		(0.0007)	(0.1191)
+ Proportion of Connected Villagers	7	0.8342	0.3575
		(0.0008)	(0.1229)
+ Proportion of Leaders	8	0.8101	0.3604
		(0.0008)	(0.1237)

Restrictiveness and Completeness of Each Model

	# Param	Restrictiveness	Completeness
Linear Models			
Eigenvector Centrality of Leaders	1	0.9762	0.2577
		(0.0003)	(0.1101)
+ Degree Centrality of Leaders	2	0.9526	0.3385
		(0.0004)	(0.1193)
+ Degree Centrality of All Villagers	3	0.9288	0.3471
		(0.0005)	(0.1151)
+ Betweenness Centrality of Leaders	4	0.9053	0.3475
		(0.0006)	(0.1158)
+ Clustering Coefficient	5	0.8816	0.3516
		(0.0007)	(0.1191)
+ Average Path Length	6	0.8579	0.3516
		(0.0007)	(0.1191)
+ Proportion of Connected Villagers	7	0.8342	0.3575
		(0.0008)	(0.1229)
+ Proportion of Leaders	8	0.8101	0.3604
		(0.0008)	(0.1237)
Structural Model	4	0.9408	0.0674
		(0.0036)	(0.0452)

Restrictiveness and Completeness of Each Model

	# Param	Restrictiveness	Completeness
Linear Models			
Eigenvector Centrality of Leaders	1	0.9762	0.2577
		(0.0003)	(0.1101)
+ Degree Centrality of Leaders	2	0.9526	0.3385
		(0.0004)	(0.1193)
+ Degree Centrality of All Villagers	3	0.9288	0.3471
		(0.0005)	(0.1151)
+ Betweenness Centrality of Leaders	4	0.9053	0.3475
		(0.0006)	(0.1158)
+ Clustering Coefficient	5	0.8816	0.3516
		(0.0007)	(0.1191)
+ Average Path Length	6	0.8579	0.3516
		(0.0007)	(0.1191)
+ Proportion of Connected Villagers	7	0.8342	0.3575
		(0.0008)	(0.1229)
+ Proportion of Leaders	8	0.8101	0.3604
		(0.0008)	(0.1237)
Structural Model	4	0.9408	0.0674
		(0.0036)	(0.0452)

Takeaways

- OLS models:
  - More regressors $\Longrightarrow$ model becomes weakly more complete and less restrictive.
  - But completeness eventually plateaus while restrictiveness continues to decrease at roughly the same rate.

Takeaways

- OLS models:
  - More regressors $\implies$ model becomes weakly more complete and less restrictive.
  - But completeness eventually plateaus while restrictiveness continues to decrease at roughly the same rate.
- Structural model:
  - Restrictiveness is very high, 0.94, suggesting that the model imposes substantial restrictions across village configurations.
  - But the model's completeness is only 0.07, so it does not capture much of the variation in the village takeup rate.
  - The structural model is both less complete and less restrictive than the simple OLS model built on eigenvector centrality.

Conclusion

- When a theory fits the data well, it matters whether this is
  - because the theory captures important regularities in the data
  - because the theory is so flexible that the only constraints it imposes are basic background constraints that we already know
- We provide a practical, algorithmic approach for evaluating the restrictiveness of a theory, which can also be used to investigate the role of specific parameters.
- A final note:
  - Relative to highly flexible ML/statistical methods, economic theory is distinguished in part by the structure it imposes
  - A question then emerges of exactly how restrictive our models actually are compared to the nearly nonparametric approaches used in high-dimensional statistical modeling
  - The proposed measure offers a way to quantify this

Thank You

Estimating Restrictiveness

Assumption 1: $\mathbb{E}_{f \sim \text{Unif}(\mathscr{F})}[d(f_{\text{base}}, f)] > 0$.

Proposition

Under Assumption 1,

$$\frac{\sqrt{M}\left(\hat{r}_M - r\right)}{\hat{\sigma}_{\hat{r}}} \xrightarrow{d} \mathcal{N}\left(0, 1\right).$$

where the asymptotic variance estimator $\hat{\sigma}_{\hat{r}}^2$ is defined by

$$\hat{\sigma}_{\hat{r}}^2 := \frac{\hat{\sigma}_{\mathcal{G}}^2 - 2 \cdot \hat{r} \cdot \hat{\sigma}_{\mathcal{G}, f_{base}} + \hat{r}^2 \cdot \hat{\sigma}_{f_{base}}^2}{\left(\frac{1}{M}\sum_{m=1}^{M} d(f_{base}, f_m)\right)^2}, \tag{1}$$

with $\hat{\sigma}_{\mathcal{G}}^2$ being the sample variance of $d(\mathcal{G}, f_m)$, $\hat{\sigma}_{f_{base}}^2$ the sample variance of $d(f_{base}, f_m)$, and $\hat{\sigma}_{\mathcal{G}, f_{base}}^2$ the sample covariance of $d(\mathcal{G}, f_m)$ and $d(f_{base}, f_m)$, across $m = 1, ..., M$.

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.

Proof Sketch

$$\text{A1-A5} \iff e(\mathcal{G}, \mathscr{F}, d) = \mathbb{E}_{f \sim \mathsf{Unif}(\mathscr{F})}\left[c \cdot d(\mathcal{G}, f)\right]$$

Clearly the axioms are satisfied by the representation. Other direction:

- Fix an arbitrary model $\mathcal{G}$ and distance d, and define

$$\nu(\mathscr{F}) = e(\mathcal{G}, \mathscr{F}, d) \cdot \mu(\mathscr{F}) \quad \forall \, \text{measurable } \mathscr{F}$$

Proof Sketch

$$\text{A1-A5} \iff e(\mathcal{G}, \mathscr{F}, d) = \mathbb{E}_{f \sim \text{Unif}(\mathscr{F})}\left[c \cdot d(\mathcal{G}, f)\right]$$

Clearly the axioms are satisfied by the representation. Other direction:

- Fix an arbitrary model $\mathcal{G}$ and distance d, and define

$$\nu(\mathscr{F}) = e(\mathcal{G}, \mathscr{F}, d) \cdot \mu(\mathscr{F}) \quad \forall \text{ measurable } \mathscr{F}$$

- A1 (Nonnegativity) and A4 (Linearity) $\Rightarrow \nu$ is a measure.

Proof Sketch

$$\text{A1-A5} \iff e(\mathcal{G}, \mathscr{F}, d) = \mathbb{E}_{f \sim \text{Unif}(\mathscr{F})}\left[c \cdot d(\mathcal{G}, f)\right]$$

Clearly the axioms are satisfied by the representation. Other direction:

- Fix an arbitrary model $\mathcal{G}$ and distance d, and define

$$\nu(\mathscr{F}) = e(\mathcal{G}, \mathscr{F}, d) \cdot \mu(\mathscr{F}) \quad \forall \text{ measurable } \mathscr{F}$$

- A1 (Nonnegativity) and A4 (Linearity) $\Rightarrow \nu$ is a measure.

- The Radon-Nikodym theorem implies existence of a function $h_{\mathcal{G},d} : \mathscr{F}^* \to \mathbb{R}$ such that $\nu(\mathscr{F}) = \int_{\mathscr{F}} h_{\mathcal{G},d}(f) d\mu$.

Proof Sketch

$$\text{A1-A5} \iff e(\mathcal{G}, \mathscr{F}, d) = \mathbb{E}_{f \sim \mathsf{Unif}(\mathscr{F})}\left[c \cdot d(\mathcal{G}, f)\right]$$

Clearly the axioms are satisfied by the representation. Other direction:

- Fix an arbitrary model $\mathcal{G}$ and distance d, and define

$$\nu(\mathscr{F}) = e(\mathcal{G}, \mathscr{F}, d) \cdot \mu(\mathscr{F}) \quad \forall \text{ measurable } \mathscr{F}$$

- A1 (Nonnegativity) and A4 (Linearity) $\Rightarrow \nu$ is a measure.
- The Radon-Nikodym theorem implies existence of a function $h_{\mathcal{G},d} : \mathscr{F}^* \to \mathbb{R}$ such that $\nu(\mathscr{F}) = \int_{\mathscr{F}} h_{\mathcal{G},d}(f) d\mu$.
- Use this to show that: $e(\mathcal{G}, \mathscr{F}, d) = \int h_{\mathcal{G},d}(f) d\mu_{\mathscr{F}}$

Proof Sketch

$$\text{A1-A5} \iff e(\mathcal{G}, \mathscr{F}, d) = \mathbb{E}_{f \sim \text{Unif}(\mathscr{F})} \left[c \cdot d(\mathcal{G}, f) \right]$$

Clearly the axioms are satisfied by the representation. Other direction:

- Fix an arbitrary model $\mathcal{G}$ and distance d, and define

  $$\nu(\mathscr{F}) = e(\mathcal{G}, \mathscr{F}, d) \cdot \mu(\mathscr{F}) \quad \forall \text{ measurable } \mathscr{F}$$

- A1 (Nonnegativity) and A4 (Linearity) $\Rightarrow$ ν is a measure.
- The Radon-Nikodym theorem implies existence of a function $h_{\mathcal{G},d} : \mathscr{F}^* \to \mathbb{R}$ such that $\nu(\mathscr{F}) = \int_{\mathscr{F}} h_{\mathcal{G},d}(f) d\mu$.
- Use this to show that: $e(\mathcal{G}, \mathscr{F}, d) = \int h_{\mathcal{G},d}(f) d\mu_{\mathscr{F}}$
- A2 (Monotonicity) and A3 (Rescaling) $\Rightarrow$ $h_{\mathcal{G},d}(f)$ is linear in the distance between the model $\mathcal{G}$ and the mapping f.

Proof Sketch

$$\text{A1-A5} \iff e(\mathcal{G}, \mathscr{F}, d) = \mathbb{E}_{f \sim \text{Unif}(\mathscr{F})}\left[c \cdot d(\mathcal{G}, f)\right]$$

Clearly the axioms are satisfied by the representation. Other direction:

- Fix an arbitrary model $\mathcal{G}$ and distance d, and define

  $$\nu(\mathscr{F}) = e(\mathcal{G}, \mathscr{F}, d) \cdot \mu(\mathscr{F}) \quad \forall \text{ measurable } \mathscr{F}$$

- A1 (Nonnegativity) and A4 (Linearity) $\Rightarrow \nu$ is a measure.
- The Radon-Nikodym theorem implies existence of a function $h_{\mathcal{G},d} : \mathscr{F}^* \to \mathbb{R}$ such that $\nu(\mathscr{F}) = \int_{\mathscr{F}} h_{\mathcal{G},d}(f) d\mu$.
- Use this to show that: $e(\mathcal{G}, \mathscr{F}, d) = \int h_{\mathcal{G},d}(f) d\mu_{\mathscr{F}}$
- A2 (Monotonicity) and A3 (Rescaling) $\Rightarrow h_{\mathcal{G},d}(f)$ is linear in the distance between the model $\mathcal{G}$ and the mapping f.
- A5 (Symmetry) pins down the uniform distribution.