Quantifying the Restrictiveness of Theories

Drew Fudenberg[*] Wayne Gao [†] Annie Liang[‡]

April 19, 2020

Abstract

We propose an algorithm for quantifying the restrictiveness of economic models. Our restrictiveness measure is evaluated on simulated, hypothetical data sets that are drawn at random from a distribution that satisfies some application-dependent content restrictions, such as that people should prefer more money to less. For each such data set, we measure the extent to which the best version of the model (i.e the parameters that give the lowest cross-validated prediction error) improves on a naive prediction rule such as guessing at random, compared to the best achievable improvement. Models that can fit almost all data well are not restrictive. We illustrate the proposed approach with two applications: using Cumulative Prospect Theory to predict certainty equivalents for lotteries, and using the Poisson Cognitive Hierarchy Model to predict the distribution of initial play in games.

[*] Department of Economics, MIT

[†] Department of Economics, U. Pennsylvania

[‡] Department of Economics, U. Pennsylvania

1 Introduction

When a model does a good job of fitting the available data, is it because the model is so flexible that it would fit most possible data, or does it achieve this by capturing structure present in the outcomes of interest?

One approach to determine what a model allows and what it rules out is to formally characterize the empirical content of a model through representation theorems. But representation theorems don't exist for most economic models, and even when there are such theorems for a very general version of the model, there generally aren't any for the functional forms commonly used in applied work to fit data. Moreover, the predictive success of a model is usually based on whether its predictions are approximately rather than exactly correct, and there are not representation theorems for the approximate predictions of most economic models.

Our goal in this paper is to provide a quantitative measure of model restrictiveness that can be practically computed across a variety of applications. Specifically, we propose an algorithm that determines the ability of a theory to approximate a wide range of data. Our approach is to first stipulate some basic application-dependent content restrictions, such as that people should prefer more money to less. Then we generate random data sets that obey these properties, and determine the restrictiveness of a model based on its performance on this hypothetical data. A more restrictive model is one that performs less well on random data sets: it encodes structure beyond what is present in the basic restrictions.

To measure a model's performance, we use the *completeness* measure of Fudenberg et al. (2019), which is the extent to which the best model in the class (i.e the parameters that give the lowest cross-validated prediction error) improve on a naive prediction rule such as guessing at random. A theory that is very complete captures most of the important regularities in the observed behavior. But if a theory can approximate almost all conceivable data, its ability to fit the data doesn't speak to its relevance. An ideal model would have very high completeness so that it does a great job of predicting real outcomes, but also very high restrictiveness, so that it is not consistent with various sorts of counterfactual data.

We illustrate our method with two classic prediction problems from experimental economics—predicting certainty equivalents for binary lotteries and predicting initial play in matrix games—and evaluate models in these domains from the dual perspectives of completeness and restrictiveness. In our first application, we evaluate the performance of a four-parameter version of Cumulative Prospect Theory (CPT) for prediction of data from Bruhin et al. (2010), consisting of certainty equivalents reported across a population of subjects for a set of 50 binary lotteries. We find that while CPT is almost fully complete (achieving a completeness of 0.93), it is also very flexible: When we generate hypothetical data sets (restricted to satisfy first-order stochastic dominance), the average completeness of the model is 0.63. This suggests that the four parameter version of CPT is rich enough to provide a reasonably good fit for any plausible data set.

Here restrictiveness provides a new perspective on the problem of how richly to parameterize a model. Minimizing cross-validated prediction error already leads to one bound on this: overparameterized models can overfit to training data and perform poorly on test data. But since the test and training sets are drawn from actual observations, determining which version of a model has the lowest cross-validated prediction error doesn't tell us how much of a model's success is due to its unrestrictiveness, and how much is due to the fact that it tracks regularities that are present in the data.

Free parameters improve a model's ability to fit to data, but decrease its restrictiveness. One way of evaluating the value of additional parameters is to compare how much they increase completeness, compared to how much they decrease restrictiveness. We next compare our initial four-parameter specification of CPT with various alternative specifications from the literature that have fewer parameters. Our results point to the importance of the nonlinear probability weighting parameters in CPT. Specifically, we find that including only the nonlinear probability weighting parameters achieves most of the completeness of the four-parameter specification while being substantially more restrictive.

Our second application is to the prediction of initial play in matrix games. We use data from Fudenberg and Liang (2019), including play in 466 3×3 normal-form games. We find that the Poisson Cognitive Hierarchy Model (PCHM) (Camerer et al., 2004) achieves a completeness of 0.44, which might suggest that it is less revealing or insightful than CPT. But the restrictiveness of the PCHM is 0.92, meaning it rules out most possible behaviors. This tells us that the PCHM captures a systematic regularity in the actual data that is not imposed by our background assumptions on the hypothetical data. Note that this is a joint commentary on the content of the models and the restrictiveness of the background data constraints: In settings where prior knowledge or intuition more sharply restrict the conceivable data, we expect a model's restrictiveness to be higher.

We then compare the PCHM with *logit level-1*, which assumes the distribution is a logistic best reply to the uniform distribution, and to *logit PCHM*, which allows for logistic best replies in the PCHM (Wright and Leyton-Brown, 2014). We find that logit level-1 is simultaneously more complete and more restrictive than the PCHM, which suggests that it is a better model of initial play (although we suspect not a better model of play given repetition and feedback). Moreover, it is almost as complete as the logit PCHM and substantially more restrictive (0.93 as opposed to 0.82).

2 Related Work

Koopmans and Reiersol (1950) pointed out that unless a theory is *observationally restrictive*, it cannot be refuted from data, and provided definitions for whether or not a theory has any restrictiveness at all.[1] Selten (1991) proposed measuring the restrictiveness by the fraction of possible data sets that could be exactly explained by the theory. Like Selten (1991), we propose a measure for how restrictive a model is, but we relax exact consistency to a quantitative measure of accuracy; that is, we ask how well an arbitrary data set can be approximated by the theory. (See further discussion in Section 3.3). A second, more fundamental, difference is that we focus on the question of how to (algorithmically) compute completeness and restrictiveness,

[1] Relatedly, a long line of econometric literature on *overidentification* provide theoretical conditions and statistical tests for the overidentification of econometric models: for example, Sargan (1958), Hausman (1978), Hansen (1982), and Chen and Santos (2018).

while the Selten (1991) measures must be analytically determined. This may only be feasible in very special cases, such as in the Harless and Camerer (1994) study of a data set where subjects made three choices from pairs of binary lotteries. For our applications, and many others, the number of possible observations is much larger, which makes it difficult if not impossible to analytically determine which observations are consistent with the theory. We demonstrate how our algorithmic approach can nevertheless provide insight into the restrictiveness of the theory.

Our work is also related to representation theorems in decision theory, which describe the empirical content of different models. Our work complements this by providing an approach for when representation theorems either do not exist or do not apply to the specific functional form that is used by the analyst. For example, although there are representation theorems that characterize which data are consistent with general Cumulative Prospect Theory specification (Quiggin, 1982; Yaari, 1987), there are no representation theorems for the popular functional form we use here, and the same is true for the Poisson Cognitive Hierarchy Model. Moreover, even in settings where there are representation theorems that characterize the behavior consistent with a theory, it can be computationally challenging to bring them to the data: For example, the Harless and Camerer (1994) exercise would be much harder on larger menus of binary lotteries, on 3-outcome lotteries, or if subjects had been asked to report certainty equivalents.

This paper is related to the vast statistic and econometric literature on model selection, which dates back to Cox (1961, 1962). Our restrictiveness measure may be useful as part of model selection, but it has a different goal. Different from classic measures like AIC and BIC, it is not based on observed data, nor is it designed to guard against overfitting. Instead, it proposes a practical procedure for evaluating the restrictiveness of a parametric modeling class within a class of permissible models.[2] Similarly, although VC dimension—which provides another measure for the "span" of a model—is related to our restrictiveness measure at a high level, it is generally

[2] This paper also has a different goal than the extensive econometric literature that studies how the "restrictiveness" of an econometric model may affect the identification of parameters and the efficiency of estimators.

nontrivial to determine the VC dimension of any given model.[3] In contrast, our metric is (by design) easy to compute.

Finally, the paper utilizes recent development in the statistics literature, specifically Austern and Zhou (2020) on the asymptotic theory of cross-validation risk estimator, which we use to evaluate the error of a given model class in a given data set.

3 Key Definitions

3.1 Preliminaries

Let X be an observable (random) *feature vector* taking values in a finite set $\mathcal{X}$, and Y be a (random) *outcome* of interest taking values in a finite-dimensional set $\mathcal{Y}$. Each observation is a pair $Z_i = (X_i, Y_i)$, which we assume are i.i.d with distribution P^*. The marginal distribution on features P^*_X is known (or chosen by the analyst) while the conditional distribution $P^*_{Y|X}$ is not.

The analyst is interested in predicting a statistic of the conditional distribution

$$s(x) \equiv \varphi(P^*_{Y|X}(x)).$$

The statistic take values in a space S. Two important special cases include:

(a) $s(x) = \mathbb{E}_{P^*_{Y|X}}[Y \mid X = x]$, so that the analyst seeks to predict the conditional expectation of Y at each x.

(b) $s(x) = P^*_{Y|X}(x)$ and $S = \Delta(\mathcal{Y})$, so that the analyst seeks to predict the conditional distribution of Y at each x.

Any function $f : \mathcal{X} \to S$ is said to be a *predictive mapping* or simply *mapping.* Write $\mathcal{F}$ for the set of all models. We suppose that the problem comes equipped with a loss function $l : \mathcal{F} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, where $l(f, (x, y))$ is the error to predicting $f(x)$ when the realized outcome is y. For example, a common loss function for real-valued

[3] The VC dimension is known for very few economic models. A recent exception is the work of Basu and Echenique (2020) for various models of decision-making under uncertainty.

predictions $f(x)$ is mean-squared error, $l(f,(x,y)) = (y - f(x))^2$, and a common loss function for prediction of distributions $f(\cdot \mid x) \in \Delta(\mathcal{Y})$ is negative log-likelihood, $l(f,(x,y)) = -\log f(y \mid x)$.

The *(expected) error* of mapping f for predicting a new test case is

$$e\left(f\right) := \mathbb{E}_{P^*}\left[l(Y_i, f(X_i))\right] \tag{1}$$

and the mapping that minimizes error is

$$f^*(x) := \min_{f \in \mathcal{F}} e(f). \tag{2}$$

We call f^* the *best mapping.*

We will focus on evaluating parametric economic models $\mathcal{F}_\Theta = \{f_\theta\}_{\theta \in \Theta}$, where Θ is a finite-dimensional, closed, and compact set. We assume that the loss function and the model class satisfy:

Assumption 1. *(a) $l(f_\theta,(x,y))$ is continuous with respect to θ on Θ.*

(b) $l(f,(x,y)) = 0$ if $f(x) = y$.

3.2 Completeness

We review here the definition of *completeness* from Fudenberg et al. (2019), which is the amount that a model improves predictions over a naive rule, compared to the best achievable improvement given the available features.

The mapping in $\mathcal{F}_\Theta$ that minimizes error is

$$f_{\theta^*} = \arg\min_{f \in \mathcal{F}_\Theta} e(f).$$

Following Fudenberg et al. (2019), we normalize this error relative to two baselines: the error achieved by a naive mapping f_{naive} suited to the problem, and the error achieved by the best mapping f^*. We assume throughout that the naive mapping is worse than the best model in $\mathcal{F}_\Theta$:

Assumption 2. $e(f_{naive}) \geq e(f_{\theta^*})$

A sufficient condition for this is $f_{\text{naive}} \in \mathcal{F}_\Theta$, as it is in our subsequent applications.

The *completeness* of model $\mathcal{F}_\Theta$ is defined as the ratio of the error reduction achieved by the model compared to the achievable reduction:

Definition 1. *The completeness of model $\mathcal{F}_\Theta$ is*

$$\kappa^* := \frac{e\left(f_{naive}\right) - e\left(f_{\theta^*}\right)}{e\left(f_{naive}\right) - e\left(f^*\right)}. \tag{3}$$

Assumption 2 implies that the error of the naive mapping is an upper bound on error of the best mapping in $\mathcal{F}_\Theta$, so completeness κ^* is upper bounded by 1. Since also $e(f^*)$ is the lowest achievable error by definition, κ^* is nonnegative. Thus, completeness ranges between 0 and 1, where a model with $\kappa^* = 1$ predicts as well as the best mapping f^*, while a model with $\kappa^* = 0$ predicts no better than the naive mapping.

3.3 Restrictiveness

One explanation for a very high observed completeness measure is simply that the model is flexible enough to accommodate any pattern of behavior. We would thus like to distinguish high completeness because a model includes most mappings from $\mathcal{X}$ to $\mathcal{Y}$, versus because the model includes the "right" regularities, namely those that are observed in actual data. We now propose an algorithmic method for quantifying the restrictiveness of a model, which allows us to separate these cases.

Our strategy is to generate random mappings f from a set $\mathcal{F}_\mathcal{M}$ of "permissible mappings," and evaluate how well these mappings can be approximated using the model $\mathcal{F}_\Theta$. The more mappings from $\mathcal{F}_\mathcal{M}$ that can be approximated, the less restrictive that model is.

The set $\mathcal{F}_\mathcal{M}$ is chosen to encode prior knowledge about the setting. For example, when predicting certainty equivalents for lotteries, we may assume that people prefer more money to less. We further specify a distribution μ on $\mathcal{F}_\mathcal{M}$ chosen by the analyst, where μ is interpreted to be the analyst's prior over the space of mappings. In our applications below, we take μ to be uniform on $\mathcal{F}_\mathcal{M}$.[4]

[4] It can also be instructive to compute restrictiveness with respect to different choices of μ—

Formally, for any two mappings f and f', define the *discrepancy*

$$d(f, f') = \mathbb{E}_{P_X}\big(l(f(X), f'(X)\big)$$

to be the average loss in predicting according to f when the data is generated by f'. Further define

$$d(\mathcal{F}_\Theta, f) = \inf_{f' \in \mathcal{F}_\Theta} d(f', f)$$

to be the discrepancy between f and the closest mapping in $\mathcal{F}_\Theta$, so that $d(\mathcal{F}_\Theta, f)/d(f_{\text{naive}}, f)$ is a normalized measure of the average discrepancy between $\mathcal{F}_\Theta$ and f, relative to the naive prediction rule introduced in the previous section. We call this ratio the *normalized discrepancy*; it is bounded between 0 and 1 from Assumption 2.[5]

The restrictiveness of model $\mathcal{F}_\Theta$ is then defined to be the average normalized discrepancy between $\mathcal{F}_\Theta$ and random mappings f with distribution μ.

Definition 2. *The* restrictiveness *of model* $\mathcal{F}_\Theta$ *is* $r := \mathbb{E}_\mu \left[\dfrac{d(\mathcal{F}_\Theta, f)}{d(f_{naive}, f)} \right]$.

Smaller values of r correspond to less restrictive models: If $\mathcal{F}_\Theta = \mathcal{F}_M$ (so that the model is completely unrestrictive), then $r = 0$ for every choice of μ. Shrinking $\mathcal{F}_\mathcal{M}$ reduces the restrictiveness measure: A low restrictiveness score means that the model imposes very few restrictions, or makes very few predictions, beyond the properties that are already imposed by $\mathcal{F}_\mathcal{M}$.

Relationship to completeness. For some loss functions, the expected error of mapping f can be decomposed into the expected error of the best mapping f^*, and the discrepancy between f and f^*:

$$e(f) = e(f^*) + d(f, f^*). \tag{4}$$

including those that have support on different permissible sets $\mathcal{F}_\mathcal{M}$—as we do in Appendix B.1.

[5] Normalizing in this way allows us to make meaningful comparisons across different problem domains, such as we do here across initial play and certainty equivalents. It also makes our measure less sensitive to rescaling: for example, in our application to certainty equivalents, scaling up the payoffs results in large changes to mean-squared error, and hence potentially to the discrepancy between the model and the best mapping.

As we show in Appendix A, this holds for mean-squared error and negative log-likelihood, the loss functions that we use in our applications. Appendix A also proves the following claim:

Claim 1. *When the decomposition in (4) is valid, then* $\kappa^* = 1 - \frac{d(\mathcal{F}_\Theta, f^*)}{d(f_{naive}, f^*)}$

That is, completeness is simply 1 minus the discrepancy between $\mathcal{F}_\Theta$ and the best mapping f^*. Thus when (4) holds we can define

$$\kappa(f) = 1 - \frac{d(\mathcal{F}_\Theta, f)}{d(f_{\text{naive}}, f)}$$

to be the completeness of the model for predicting data best fit by an arbitrary mapping f. Restrictiveness is then

$$r = 1 - \mathbb{E}_\mu(\kappa(f)) \tag{5}$$

or the "inverse" of average completeness on arbitrary data. Thus, large κ^* and small r means that the model has high completeness for predicting the actual data, but low completeness for predicting hypothetical behaviors.

An alternative "area" measure. An alternative measure of restrictiveness is $1 - \mu(\mathcal{F}_\Theta)$; that is, the fraction of possible mappings that are consistent with the model. (This is very similar to Selten (1991)'s proposed "area" measure.) Our measure of restrictiveness is substantively different from this, as it measures how well the model $\mathcal{F}_\Theta$ *approximates* a randomly drawn mapping f in $\mathcal{F}_\mathcal{M}$. We define restrictiveness in this way to allow for quantification of the degree of error. A model that doesn't include most mappings from $\mathcal{F}_\mathcal{M}$ can nevertheless have low restrictiveness by our measure if it approximates most mappings very well. In particular, while the area measure $\mu(\mathcal{F}_\Theta)$ concludes that all finite models $\mathcal{F}_\Theta$ are completely restrictive, our measure of restrictiveness does not, and can be instructive for ordering models within this class.

Sensitivity to μ. We might prefer that the restrictiveness measure doesn't respond too sensitively to small changes in μ. We demonstrate now that it does not.

For any two measures $\mu, \mu' \in \Delta(\mathcal{F})$,

$$\mathbb{E}_{\mu}\left[\frac{d(\mathcal{F}_{\Theta}, f)}{d(f_{\text{naive}})}\right] - \mathbb{E}_{\mu'}\left[\frac{d(\mathcal{F}_{\Theta}, f)}{d(f_{\text{naive}})}\right] \leq \int \frac{d(\mathcal{F}_{\Theta}, f)}{d(f_{\text{naive}})} \cdot |d\mu - d\mu'| \leq 2 \cdot \delta_{TV}(\mu, \mu')$$

where δ_{TV} is the total variation distance. Thus for any two measures that are close in total variation distance, the corresponding restrictiveness measures must also be close.

Combining restrictiveness and completeness into a single measure. We take the view that it is preferable for a model to be simultaneously more complete and more restrictive, which implies a partial ordering over models. There are many ways to complete this ordering. One possibility is to use a lexicographic ordering, where models are first ordered by completeness, and then ordered by restrictiveness. Another is to impose a functional form for combining completeness κ^* and restrictiveness r, such as $\kappa^* + r$. When the representation of restrictiveness in (5) is valid, $\kappa^* + r = \kappa^* - \mathbb{E}_{\mu}(\kappa)$, so this metric can be interpreted as the difference between completeness on actual data versus hypothetical data. Another possible measure is the CDF of the distribution of completeness measures $\kappa(f)$ (where f is drawn according to μ) evaluated at the completeness κ^* for the real data. This would tell us the probability that the model explains the actual data better than it explains a randomly generated data set. In the present paper, we report κ^* and r separately, and leave it to the analyst's discretion whether or how to combine these two metrics.

4 Estimates and Test Statistics

We now discuss how to implement our approach in practice. We suppose that the analyst has access to a finite sample of data $\{Z_i := (X_i, Y_i)\}_{i=1}^{N}$ drawn from the unknown true distribution P^*, which is used to estimate completeness.

4.1 Estimating Completeness κ^*

For a given data set $\mathbf{Z}_N = \{(X_i, Y_i)\}_{i=1}^N$ and a given model $\widetilde{\mathcal{F}}$, we use K-fold cross validation to estimate the out-of-sample prediction error of the model. (In our applications, we take the standard choice of $K = 10$.) Specifically, we randomly divide $\mathbf{Z}_N$ into K (approximately) equal-sized groups. For notational simplicity assume that $J_N = \frac{N}{K}$ is an integer. Let $k(i)$ denote the group number of observation Z_i, and for each group $k = 1, ..., K$, define

$$\hat{f}^{-k} := \arg\min_{f \in \mathcal{F}} \frac{1}{N - J_N} \sum_{k(i) \neq k} (Y_i - f(X_i))^2$$

to be the mapping from $\widetilde{\mathcal{F}}$ that minimizes error for prediction of observations outside of group k. This estimated mapping is used for prediction of the k-th test set, and

$$\hat{e}_k := \frac{1}{J_N} \sum_{k(i)=k} \left(Y_i - \hat{f}^{-k}(X_i)\right)^2$$

is its out-of-sample error on the k-th test set. Then,

$$CV(\mathcal{F}) := \frac{1}{K} \sum_{k=1}^{K} \hat{e}_k$$

is the average test error across the K folds. This is an estimator for the unobservable expected error of the best mapping from class $\mathcal{F}$.

Setting $\widetilde{F}$ to be respectively $\mathcal{F}_\Theta$, $\mathcal{F}$, or $\mathcal{F}_{\text{naive}} = \{f_{\text{naive}}\}$, we can compute $CV(\mathcal{F}_\Theta)$, $CV(\mathcal{F})$ and $CV(\mathcal{F}_{\text{naive}})$ from the data, leading to the following estimator for κ^*:

$$\hat{\kappa}^* = \frac{CV(\mathcal{F}_{naive}) - CV(\mathcal{F}_\Theta)}{CV(\mathcal{F}_{naive}) - CV(\mathcal{F})}$$

It is crucial that the denominator in $\hat{\kappa}^*$ does not vanish asymptotically, so we impose the following assumption:

Assumption 3 (Naive Rule is Imperfect)**.** $e(f_{naive}) - e(f^*) > 0$.

This assumption is quite weak, as it simply says that the naive mapping performs strictly worse in expectation than the best mapping. Under additional technical conditions, we show, by applying and adapting Proposition 5 in Austern and Zhou (2020), that $\hat{\kappa}^*$ is asymptotically normal. See Appendix C for details. To obtain the standard error, we use a variance estimator adapted from Proposition 1 in Austern and Zhou (2020). Specifically, define

$$l(Z_i, f) = (Y_i - f(X_i))^2$$

to be the test error of mapping f for prediction of observation $Z_i = (X_i, Y_i)$. Then, for the k-th test set, let $f_{\hat{\theta}^{-k}}$ and $\hat{f}^{-k}$ be the estimated mappings from model $\mathcal{F}_\Theta$ and $\mathcal{F}$, respectively. The difference in their test errors on observation Z_i is $\Delta(Z_i) = l\left(Z_i, f_{\hat{\theta}^{-k}}\right) - l\left(Z_i, \hat{f}^{-k}\right)$ and the average difference across all observations in test fold k is

$$\overline{\Delta}_k = \frac{1}{J_N} \sum_{k(i)=k} \Delta(Z_i).$$

The sample variance of the difference in test errors is correspondingly

$$\hat{\sigma}^2_{\Delta,k} = \frac{1}{J_N - 1} \sum_{k(i)=k} \left(\Delta(Z_i) - \overline{\Delta}_k\right)^2.$$

Based on this, we define the following variance estimator for $\hat{\kappa^*}$:

$$\hat{\sigma}^2_{\hat{\kappa}^*} := \frac{\frac{1}{K}\sum_{k=1}^{K} \hat{\sigma}^2_{\Delta,k}}{\left[CV\left(f_{\text{naive}}\right) - CV\left(\mathcal{F}\right)\right]^2} \tag{6}$$

We establish the asymptotic distribution of our proposed estimators via the following theorem.

Theorem 1. *Under Assumption 3 and some regularity conditions*[6]*:*

$$\frac{\sqrt{N}\left(\hat{\kappa}^* - \kappa^*\right)}{\hat{\sigma}_{\hat{\kappa}^*}} \xrightarrow{d} \mathcal{N}\left(0, 1\right).$$

[6] See Appendix C for details of these assumptions.

Consequently, the $(1-\alpha)$ *two-sided confidence interval for* κ^* *is given by*

$$\left[\hat{\kappa}^* - q_{1-\alpha/2}\cdot\frac{1}{\sqrt{N}}\hat{\sigma}_{\hat{\kappa}},\ \hat{\kappa}^* - q_{\alpha/2}\cdot\frac{1}{\sqrt{N}}\hat{\sigma}_{\hat{\kappa}}\right]$$

where $\hat{\sigma}_{\hat{\kappa}}$ *is given in (6).*

4.2 Computing Restrictiveness r

We provide an algorithm for computing r: Sample M times from the distribution μ on $\mathcal{F}_{\mathcal{M}}$, and for each sampled $f_m \in \mathcal{F}_{\mathcal{M}}$, compute $r_m := \frac{d(\mathcal{F}_\Theta, f_m)}{d(f_{\text{naive}}, f_m)}$. The sample mean $\overline{r}_M := \frac{1}{M}\sum_{m=1}^{M} r_m$ is an estimator for restrictiveness. In principle, the number of simulations we run, M, can be taken as large as we want, so $\overline{r}$ can be made arbitrarily close to r by the Law of Large Numbers. Nevertheless, the approximation error under a given finite M can be quantified using standard statistical inference methods. We focus on the case where the distribution of r_m is nondegenerate:

Assumption 4. *The distribution of* r_m *is non-degenerate.*

Assumption 4 is a very mild condition that can be easily verified, as it is sufficient for any two r_m and r'_m to be distinct.

The sample variance is

$$\hat{\sigma}^2_{\overline{\kappa}} := \frac{1}{M}\sum_{m=1}^{M}\left(r_m - \overline{r}_M\right)^2, \tag{7}$$

and by the standard Central Limit Theorem:

Proposition 1. *Under Assumption 4,*

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

The $(1-\alpha)$*-th confidence interval for* κ^* *is given by*

$$\left[\overline{r}_M - q_{1-\alpha/2}\cdot\frac{1}{\sqrt{M}}\hat{\sigma}_{\overline{\kappa}},\ \overline{r}_M - q_{\alpha/2}\cdot\frac{1}{\sqrt{M}}\hat{\sigma}_{\overline{\kappa}}\right]$$

where $\hat{\sigma}_{\overline{\kappa}}$ *is given in (7).*

One-sided hypothesis tests on r—e.g. for the null hypothesis that the model is unrestrictive, $r = 0$—can be also carried out in standard ways. We again note that the confidence intervals here simply serve to measure the approximation error of r based on a finite number of simulations, and do not reflect randomness in experimental data.

5 Application 1: Certainty Equivalents

5.1 Setting

We consider the problem of predicting certainty equivalents for a set of binary lotteries from Bruhin et al. (2010). Each lottery is described as a tuple $x = (\overline{z}, \underline{z}, p)$ where $\overline{z} \geq \underline{z}$, and the feature space $\mathcal{X}$ consists of the 50 tuples associated with lotteries in the Bruhin et al. (2010) data. The outcome space is $\mathcal{Y} = \mathbb{R}$, where each outcome is a reported certainty equivalent. An observation (X_i, Y_i) then consists of a lottery and a reported certainty equivalent. Note that the variation in Y for fixed X reflects the fact that different subjects report different certainty equivalents for the same lottery. In Section 7, we consider subject-level heterogeneity.

A predictive mapping for this problem is any function $f : \mathcal{X} \to \mathbb{R}$ mapping the 50 lotteries into predicted certainty equivalents. To evaluate the accuracy of predictions, we use mean-squared error: $l(f, (x, y)) = -(f(x) - y)^2$.

The economic model that we consider is a four-parameter version of *Cumulative Prospect Theory* indexed by $\theta = (\alpha, \beta, \gamma, \delta) \in \mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R}_+$, which predicts

$$f_\theta(\overline{z}, \underline{z}, p) = w(p)v(\overline{z}) + (1 - w(p))v(\underline{z})$$

where

$$v(z) = \begin{cases} z^\alpha & \forall z \geq 0 \\ -(-z)^\beta & \forall z < 0 \end{cases} \tag{8}$$

is a value function for money, and

$$w(p) = \frac{\delta p^\gamma}{\delta p^\gamma + (1 - p)^\gamma} \tag{9}$$

is a probability weighting function.[7] We specify $\mathcal{F}_\Theta$ as the set of all such functions f_θ, and refer to this model as CPT-$(\alpha, \beta, \gamma, \delta)$.

5.2 Completeness

We first report the completeness of the model on the Bruhin et al. (2010) data. The naive benchmark is set to be the *expected value* of the lottery, which predicts $f_{\text{naive}}(\overline{z}, \underline{z}, p) = p\overline{z} + (1-p)\underline{z}$. The out-of-sample error of this naive rule is 103.31, while the out-of-sample error of CPT-$(\alpha, \beta, \gamma, \delta)$ is 67.84. The best achievable error in the problem is estimated to be 65.51. Thus, completeness for CPT-$(\alpha, \beta, \gamma, \delta)$ is estimated to be

$$\frac{103.81 - 67.84}{103.81 - 65.51} = 0.93$$

This high level of completeness suggests that CPT-$(\alpha, \beta, \gamma, \delta)$ is a very good description of behavior, but it is also possible that the high completeness is because the model is flexible enough to mimic most functions from binary lotteries to certainty equivalents. To determine whether this is the case, we next compute the restrictiveness of the model on this domain.

5.3 Restrictiveness

We define the permissible set $\mathcal{F}_M$ to be all mappings satisfying the following criteria:

1. $\underline{z} \leq f(\overline{z}, \underline{z}, p) \leq \overline{z}$

2. if $\overline{z} \geq \overline{z}'$, $\underline{z} \geq \underline{z}'$, and $p \geq p'$ then $f(\overline{z}, \underline{z}, p) \geq f(\overline{z}', \underline{z}', p')$

3. if $\overline{z} \geq \underline{z}$, $p \geq p'$, then $f(\overline{z}, \underline{z}, p) \geq f(\overline{z}, \underline{z}, p')$

The restriction in (1) requires that the certainty equivalent is within the range of the possible payoffs. The restrictions in (2) and (3) require the function f to respect first-order stochastic dominance. Note that in the Bruhin et al. (2010) lottery data, there are many pairs of lotteries that can be compared via (2) and (3), so these conditions are not vacuous.

[7] This parametric form for $w(p)$ was first suggested by Goldstein and Einhorn (1987) and Lattimore et al. (1992).

Our primitive distribution μ is uniform over the set of permissible mappings $\mathcal{F}_{\mathcal{M}}$. Below generate 100 random mappings from $\mathcal{F}_{\mathcal{M}}$ and plot the distribution of normalized discrepancies with respect to these mappings.

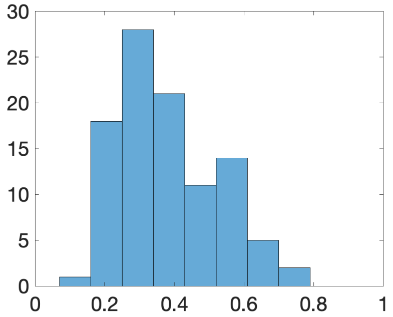


Figure 1: Distribution of normalized discrepancies

The restrictiveness of the model (i.e. the average normalized discrepancy) is 0.37, with a standard error of 0.02. This means that the average completeness of CPT-$(\alpha, \beta, \gamma, \delta)$ with respect to simulated mappings is 0.63, which is less than the completeness of 0.93 that we find on the actual data, but still fairly high.

Thus, while the high completeness of CPT-$(\alpha, \beta, \gamma, \delta)$ on the Bruhin et al. (2010) data says that it does a very good job of explaining the systematic variation in the that data, its relatively low restrictiveness means that the model would have been reasonably complete for almost any data that respects first-order stochastic dominance. This suggests that the model does not encode substantial structure about perception of risk beyond what is implied by FOSD, at least for binary lotteries.

In Appendix B.1, we consider alternative permissible sets $\mathcal{F}_M$. First, we drop the FOSD restrictions in (2) and (3), keeping only the range restriction in (1). The restrictiveness of CPT-$(\alpha, \beta, \delta, \gamma)$ on this larger set is 0.39, only slightly higher than the restrictiveness of 0.37 that we find for the main specification of $\mathcal{F}_{\mathcal{M}}$. We also consider restrictiveness relative to a permissible set $\mathcal{F}_{\mathcal{M}}$ that contains *only* mappings that violate FOSD. The restrictiveness of CPT-$(\alpha, \beta, \gamma, \delta)$ is 0.31 for this choice of permissible mappings, meaning that the model improves substantially on the naive mapping even for approximating behaviors in which people prefer less money to more. These observations all suggest that CPT-$(\alpha, \beta, \gamma, \delta)$ is not a particularly restrictive model. Of course, our analysis so far leaves open the possibility that the unrestrictiveness of the 4-parameter CPT model is specific to binary lotteries.[8] In Appendix B.3 we consider a set of 3-outcome lotteries from Bernheim and Sprenger (2020); here we find that the restrictiveness of CPT-$(\alpha, \beta, \gamma, \delta)$ is 0.57, so the model is more restrictive than on binary lotteries, although much less restrictive than the models of initial play that we study in the next section.

5.4 Comparison of Models

Adding free parameters to a model decreases its restrictiveness, but also weakly increases completeness, provided that there is enough data that overfitting is not a concern. All else equal, we prefer parameters that increase completeness without substantially decreasing restrictiveness. So one way of evaluating the value of additional parameters is to compare the increase in completeness that they permit, relative to the decrease in the restrictivenes of the model.

We next compare CPT-$(\alpha, \beta, \gamma, \delta)$ with more restrictive special cases that have been studied in the literature: $\delta = 1$, as Tversky and Kahneman (1992), $\alpha = \beta = 1$, which corresponds to a risk-neutral CPT agent whose utility function over money is $u(z) = z$ but exhibits nonlinear probability weighting, and $\delta = \gamma = 1$, which corresponds to an Expected Utility decision-maker whose utility function is as given in (8). The distribution of normalized discrepancies under these more restrictive models are shown in Figure 2 below.

[8] Although most experiments on choice under risk use binary lotteries, CPT and its simpler predecessor Prospect Theory coincide unless the lottery has 3 or more possible outcomes.

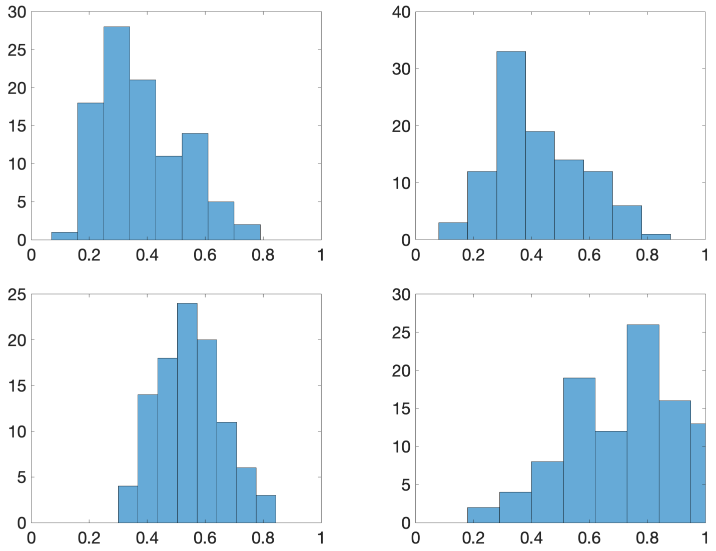


Figure 2: *Top-Left:* CPT-$(\alpha, \beta, \delta, \gamma)$ *Top-Right:* CPT-(α, β, γ), *Bottom-Left:* CPT-(δ, γ), *Bottom-Right:* CPT-(α, β).

Less general specifications are always at least weakly more restrictive, but the restrictiveness of a model must be considered jointly with its completeness. Table 5.4 reports completeness and restrictiveness measures for all four specifications of CPT.

	Completeness	N	Restrictiveness	M
CPT-$(\alpha, \beta, \delta, \gamma)$	0.93	8906	0.37	100
	(0.01)		(0.02)	
CPT-(α, β, γ)	0.76	8906	0.43	100
	(0.04)		(0.01)	
CPT-(δ, γ)	0.89	8906	0.55	100
	(0.02)		(0.01)	
CPT-(α, β)	0.13	8906	0.71	100
	(0.06)		(0.02)	

Table 1: Completeness and restrictiveness measures for each model in the certainty equivalent setting. N is the number of observations in the data used to estimate completeness. M is the number of generated mappings from $\mathcal{F}_{\mathcal{M}}$ for computation of restrictiveness.

We find that the model CPT-(δ, γ) is both more restrictive and also more complete than CPT-(α, β, γ), although it cannot be directly ranked relative CPT-(α, β) and the original CPT-$(\alpha, \beta, \delta, \gamma)$. Moreover, adding the parameters α and β over δ and

γ improves completeness only from 0.89 to 0.93, but drops restrictiveness from 0.55 to 0.37. This suggests that the probability weighting parameters δ and γ are more important. Figure 3 plots these measures and their 95% confidence intervals.

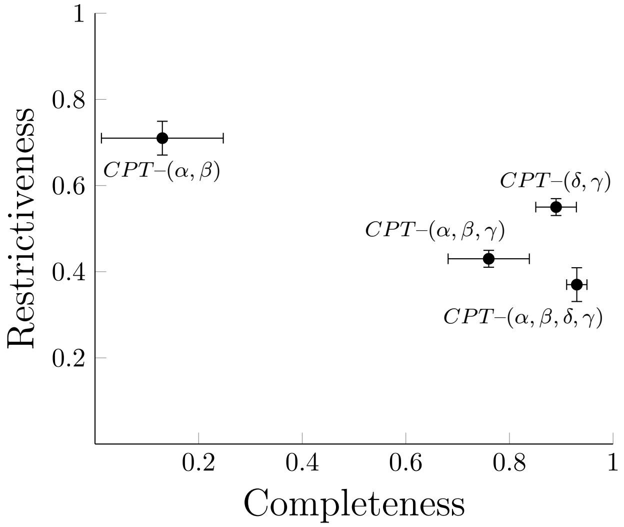


Figure 3: Comparison of CPT specifications

6 Application 2: The Distribution of Initial Play

6.1 Setting

Our second application is to predicting the distribution of initial play in games. Here the feature space $\mathcal{X}$ consists of the 466 unique 3×3 matrix games from Fudenberg and Liang (2019), each described as a vector in $\mathbb{R}^{18}$. The outcome space is $\mathcal{Y} = \{a_1, a_2, a_3\}$, the set of row player actions, and the analyst seeks to predict the conditional distribution over $\mathcal{Y}$ for each game (interpreted as choices made by a population of subjects for the same game). Thus, $S = \Delta(\mathcal{Y})$, the set of all distributions over row player actions. A predictive mapping is any function $f : \mathcal{X} \to S$ taking the 466 games into predicted distributions of play. We use negative log-likelihood as the loss function: $l(f, (x, y)) = -\log f(y \mid x)$.

We define the naive mapping to predict the uniform distribution for every game: $f_{\text{naive}}(x) = (1/3, 1/3, 1/3)$ for every x. Additionally, we consider three economic models for this prediction task. The *Poisson Cognitive Hierarchy Model* (PCHM) supposes that there is a distribution over players of differing levels of sophistication: The *level-0* player randomizes uniformly over his available actions, while the *level-1* player best responds to level-0 play (Stahl and Wilson, 1994, 1995; Nagel, 1995). Camerer et al. (2004) defines the play of level-k players, $k \geq 2$, to be the best response to a perceived distribution

$$p_k(h, \tau) = \frac{\pi_\tau(h)}{\sum_{l=0}^{k-1} \pi_\tau(l)} \qquad \forall \ h \in \mathbb{N}_{<k} \tag{10}$$

over (lower) opponent levels, where π_τ is the Poisson distribution with rate parameter τ. The parameter τ is the only free parameter of the model, and the naive mapping is nested as $\tau = 0$.

We also evaluate a model that we call *logit level-1*, which has a single free parameter $\lambda \geq 0$. For each action a_i, the predicted frequency with which a_i is played is

$$\frac{\exp\left(\lambda \cdot u(a_i)\right)}{\sum_{i=1}^{3} \exp\left(\lambda \cdot u(a_i)\right)}.$$

The model nests prediction of uniform play (our naive rule) as $\lambda = 0$, and predicts a degenerate distribution on the level-1 action when λ is sufficiently large.

Finally, we consider a model that we call *logit PCHM* (see e.g. Wright and Leyton-Brown (2014)), which replaces the assumption of exact maximization in the PCHM with a logit best response. This model has two free parameters: $\lambda, \tau \in \mathbb{R}_+$. The level-0 player chooses $g_0 = (1/3, 1/3, 1/3)$, as in the PCHM. Recursively define for each $k \geq 1$

$$v_k(a_i) = \sum_{h=0}^{k-1} p_k(h, \tau) \left(\sum_{j=1}^{3} g_h(j) u(a_i, a_j) \right)$$

to be the expected payoff of action a_i against a player whose type is distribution according to $p_k(\cdot, \tau)$, where $p_k(h, \tau)$ is as defined in (10), and define

$$g_k(a_i) = \frac{\exp(\lambda \cdot v_k(a_i))}{\sum_{j=1}^{3} \exp(\lambda \cdot v_k(a_j))}$$

to be the distribution of level-k play. We aggregate across levels using a Poisson distribution with rate parameter τ.

6.2 Completeness

The error of the naive uniform prediction rule is $-log(1/3) \approx 1.10$ and our estimate for the best achievable error is 0.831, while the errors of the PCHM, level-1(α), and logit PCHM are respectively 0.981, 0.904, and 0.901. Thus the completenesses of these models, as reported in Table 6.3, are respectively 0.436, 0.727, and 0.729.

As observed in Wright and Leyton-Brown (2014), logit PCHM substantially improves upon the completeness of the PCHM. Perhaps surprisingly, we find that almost all of this improvement is obtained by simply adding the logit parameter to the level-1 model; that is, the further improvement from allowing for multiple levels of sophistication is negligible.

The strong performance of logit level-1 for predicting initial play is consistent with an earlier result in Fudenberg and Liang (2019), where we looked at prediction of the modal action and found that the level-1 model performed quite well. Our prediction task in the present paper is more demanding, since the goal is to predict the full distribution of play. It is perhaps striking that level-1 play with a logit noise parameter achieves 72% of the achievable improvement over a naive rule. These results suggest that initial play is rather unstrategic, but systematically so.[9]

6.3 Restrictiveness

We turn now to evaluating the restrictiveness for these models. Compared to the case of preferences over binary lotteries, economic theory provides very little in the way of a prior restrictions on initial play.[10] We define the permissible set $\mathcal{F}_M$ to include all mappings satisfying the following very weak conditions:

[9] In Fudenberg and Liang (2019), we found that modal play was sometimes better described by equilibrium notions than level-1. Since such regularities cannot be accommodated by the logit level-1 model, these may explain the gap between the completeness of logit level-1 and full completeness.

[10] Classic game theory alone would suggest that dominant strategies have probability 1 and dominated strategies have probability 0, but this is inconsistent with our data (and most experimental data of play in games).

1. If an action is strictly dominated, then the frequency with which it is chosen does not exceed 1/3.[11]

2. If an action is strictly dominant, then the frequency with which it is chosen is at least 1/3.[12]

For each of the PCHM, level-1(α), and logit PCHM, we generate 100 mappings from a uniform distribution μ over the set of permissible mappings $\mathcal{F}_{\mathcal{M}}$, and evaluate the normalized discrepancy between the model and the generated mappings.[13] The distributions of normalized discrepancies are shown in the figure below.

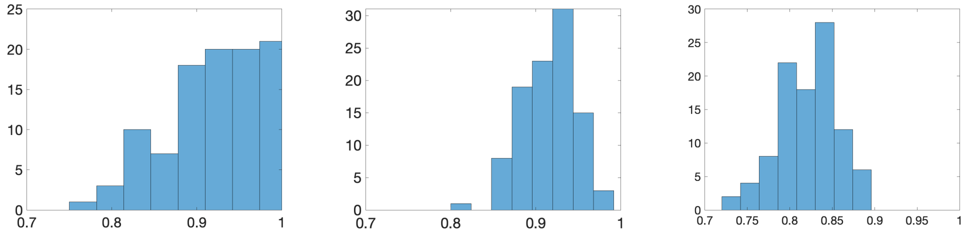


Figure 4: *Left to right*: logit level-1, PCHM, logit PCHM

We find that logit level-1's restrictiveness is 0.930, PCHM's restrictiveness is 0.915, and logit-PCHM's restrictiveness is 0.822. Indeed, across all of these mappings and models, the discrepancy is always at least 0.72. What this implies is that *almost all* distributions of initial play are inconsistent with the PCHM, level-1(α), and logit PCHM, in that no parameter values substantially improve upon predicting $(1/3, 1/3, 1/3)$. Equivalently, the completenesses of these models across the simulated mappings is bounded above by 0.28, while the completeness of these models on the actual data ranged from 0.436 to 0.729.

Simply comparing the completeness of the PCHM, 0.436, against the completeness of CPT-($\alpha, \beta, \gamma, \delta$), 0.93, suggests that the PCHM is a "worse" model of initial play

[11] In the actual data, the median strictly dominated action receives a frequency of 0.03 and the max frequency is 0.35.

[12] In the actual data, the median strictly dominant action receives a frequency of 0.86 and the min frequency is 0.69.

[13] The discrepancy between mappings is given by the average Kullback-Leibler divergence between the predicted distributions, see Appendix A for details.

than CPT is of certainty equivalents for lotteries. The contrast in their restrictivenesses (0.915 vs. 0.31) tells us that while PCHM does not capture all of the observed behaviors, it more successfully rules out behaviors that we do not observe.

Table 6.3 summarizes completeness and restrictiveness measures for all three models.

	Completeness	N	Restrictiveness	M
PCHM	0.436 (0.017)	21,393	0.915 (0.003)	100
logit level-1	0.727 (0.015)	21,393	0.930 (0.005)	100
logit PCHM	0.729 (0.014)	21,393	0.822 (0.003)	100

Table 2: Completeness and restrictiveness measures for each model in the initial play setting. N is the number of observations in the data used to estimate completeness. M is the number of generated mappings from $\mathcal{F}_{\mathcal{M}}$ for computation of restrictiveness.

We find that both logit level-1 and logit PCHM are substantially more complete than the baseline PCHM. Moreover, logit level-1 is simultaneously more complete and more restrictive than the PCHM, and it is substantially more restrictive than logit PCHM at the cost of only a slight and not statistically significant decrease in completeness.

Thus logit level-1 may be preferable to the PCHM and logit PCHM for prediction of initial play. Since logit level-1 is a completely unstrategic model—in particular, the model's predictions do not rely on the payoffs to the column player—we expect that its completeness would change if we looked at data from subjects who played the game several times and learned from feedback, though as long as we didn't change the set of feasible mappings its restrictiveness would not.

7 Heterogeneous Risk Preferences

Our analysis above considered representative agent models. In some cases, the analyst may have auxiliary data on the subjects that can be used to improve predictions. We show now how completeness and restrictiveness can be evaluated in this case.

Specifically, we return to our first application and group subjects into three clusters identified by Bruhin et al. (2010). We fit CPT-$(\alpha, \beta, \gamma, \delta)$ (henceforth CPT) for each cluster, allowing parameter values to vary across groups. Table 7 reports completeness measures cluster by cluster.

	Cluster 1	Cluster 2	Cluster 3
Naive	39.90	150.10	99.94
	(4.98)	(7.24)	(7.97)
CPT	30.08	62.67	66.38
	(3.38)	(5.07)	(3.97)
Best Achievable Error	29.39	52.08	64.81
	(3.22)	(4.35)	(3.88)
Completeness	0.93	0.89	0.96
	(0.01)	(0.03)	(0.03)
N	1341	2292	5273

Both the performance of the naive expected value rule, as well as the best achievable performance, vary substantially across clusters. For example, the behavior of subjects in cluster 1 is roughly consistent with expected value (the error of the naive rule is 39.90), while the behavior of subjects in cluster 2 departs substantially from this benchmark (the error of the naive rule is 150.10). The best achievable prediction for these groups of subjects is also very different (ranging from 29.39 to 64.81). The completeness of CPT, however, is roughly stable across the clusters.

The average completeness, weighted by proportion of observations in each cluster, is 0.935, which is very close to what we found for the representative agent model. This may seem surprising at first, since allowing for parameters to vary across subjects improves the accuracy of predictions. But the best mapping from the extended feature space $\mathcal{X}' = \mathcal{X} \times \{1, 2, 3\}$ to $\mathcal{Y}$ is more predictive than the best mapping considered previously. Thus what we find is that the completeness of CPT with three clusters, *relative to the best three-cluster mapping*, is comparable to the completeness of the representative-agent version of CPT, *relative to the best representative-agent mapping.*

Similarly, when measuring restrictiveness, we extend the set of permissible mappings to the domain $\mathcal{X}'$. Each generated pattern of behavior is thus a triple (f_1, f_2, f_3) of mappings from the original $\mathcal{F}_{\mathcal{M}}$. We ask how well these tuples can be approximated using mappings (g_1, g_2, g_3) from CPT-$(\alpha, \beta, \gamma, \delta)$. It is straightforward to see that the restrictiveness of the three-cluster CPT is identical to the restrictiveness of the representative-agent model.[14]

8 Conclusion

When a theory fits the data well, it matters whether this is because the theory captures important regularities in the data, or whether the theory is so flexible that it can explain any behavior at all. We provide a practical, algorithmic approach for evaluating how the restrictiveness of a theory, and demonstrate that it reveals new insights into models from two economic domains. The method is easily applied to other models from different domains.

[14] Note that this is true for any number of exogenously specified clusters.

References

Austern, M. and W. Zhou (2020): "Asymptotics of Cross-Validation," *arXiv preprint arXiv:2001.11111.*

Basu, P. and F. Echenique (2020): "On the falsifiability and learnability of decision theories," *Theoretical Economics*, forthcoming.

Bernheim, D. and C. Sprenger (2020): "Direct Tests of Cumulative Prospect Theory," Working Paper.

Bruhin, A., H. Fehr-Duda, and T. Epper (2010): "Risk and Rationality: Uncovering Heterogeneity in Probability Distortion," *Econometrica.*

Camerer, C. F., T.-H. Ho, and J.-K. Chong (2004): "A cognitive hierarchy model of games," *The Quarterly Journal of Economics*, 119, 861–898.

Chen, X. and A. Santos (2018): "Overidentification in regular models," *Econometrica*, 86, 1771–1817.

Cox, D. R. (1961): "Tests of separate families of hypotheses," in *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, vol. 1, 23.

——— (1962): "Further results on tests of separate families of hypotheses," *Journal of the Royal Statistical Society: Series B (Methodological)*, 24, 406–424.

Fudenberg, D., J. Kleinberg, A. Liang, and S. Mullainathan (2019): "Measuring the Completeness of Theories," Working Paper.

Fudenberg, D. and A. Liang (2019): "Predicting and Understanding Initial Play," *American Economic Review.*

Goldstein, W. M. and H. J. Einhorn (1987): "Expression theory and the preference reversal phenomena," *Psychological review*, 94, 236.

Hansen, L. P. (1982): "Large sample properties of generalized method of moments estimators," *Econometrica: Journal of the Econometric Society*, 1029–1054.

Harless, D. and C. Camerer (1994): "The Predictive Utility of Generalized Expected Utility Theories," *Econometrica*, 62, 1251–1289.

Hausman, J. A. (1978): "Specification tests in econometrics," *Econometrica: Journal of the econometric society*, 1251–1271.

Koopmans, T. and O. Reiersol (1950): "The Identification of Structural Characteristics," *The Annals of Mathematical Statistics.*

Lattimore, P. K., J. R. Baker, and A. D. Witte (1992): "The influence of probability on risky choice: A parametric examination," *Journal of Economic Behavior & Organization*, 17, 315–436.

Nagel, R. (1995): "Unraveling in Guessing Games: An Experimental Study," *American Economic Review*, 85, 1313–1326.

Quiggin, J. (1982): "A Theory of Anticipated Utility," *Journal of Economic Behavior and Organization*, 3, 323–343.

Sargan, J. D. (1958): "The estimation of economic relationships using instrumental variables," *Econometrica: Journal of the Econometric Society*, 393–415.

Selten, R. (1991): "Properties for a Measure of Predictive Success," *Mathematical Social Sciences*, 21, 153–167.

Stahl, D. O. and P. W. Wilson (1994): "Experimental evidence on players' models of other players," *Journal of Economic Behavior and Organization*, 25, 309–327.

——— (1995): "On players' models of other players: Theory and experimental evidence," *Games and Economic Behavior*, 10, 218–254.

Tversky, A. and D. Kahneman (1992): "Advances in Prospect Theory: Cumulative Representation of Uncertainty," *Journal of Risk and Uncertainty*, 5, 297–323.

Wright, J. R. and K. Leyton-Brown (2014): "Level-0 meta-models for predicting human behavior in games," *Proceedings of the fifteenth ACM conference on Economics and computation*, 857–874.

Yaari, M. (1987): "The Dual Theory of Choice under Risk," *Econometric*, 55, 95–115.

A Supplementary Material to Section 3.3

A.1 Proof of Claim 1

Suppose the decomposition in (4) is valid. Then,

$$\begin{aligned}\kappa^* = \frac{e\left(f_{\text{naive}}\right) - e\left(f_{\theta^*}\right)}{e\left(f_{\text{naive}}\right) - e\left(f^*\right)} &= \frac{(e(f^*) + d(f_{\text{naive}}, f^*)) - (e(f^*) + d(f_{\theta^*}, f^*))}{(e(f^*) + d(f_{\text{naive}}, f^*)) - (e(f^*) + d(f^*, f^*))} \\ &= \frac{d(f_{\text{naive}}, f^*) - d(f_{\theta^*}, f^*)}{d(f_{\text{naive}}, f^*)} \qquad \text{(A.1)}\end{aligned}$$

using in the final step that $d(f^*, f^*) = 0$. It remains to show $d(f_{\theta^*}, f^*) = \inf_{\theta\in\Theta} d(f_\theta, f^*)$, which follows since

$$\theta^* = \arg\min_{\theta\in\Theta} e(f_\theta) = \arg\min_{\theta\in\Theta} \left(e(f^*) + d(f_\theta, f^*)\right) = \arg\min_{\theta\in\Theta} d(f_\theta, f^*).$$

A.2 The decomposition in (4) is valid in our two applications

We show now that the decomposition in (4) is satisfied for the two loss functions used in our applications:

Mean-Squared Error. Suppose $S = \mathcal{Y} = \mathbb{R}$ and the loss function is $l(f, (x, y)) = (y - f(x))^2$. The following decomposition is standard:

$$\begin{aligned} e\left(f\right) &:= \mathbb{E}_{P^*}\left[\left(Y - f\left(X\right)\right)^2\right] \\ &= \mathbb{E}_{P^*}\left[\left(Y - f^*\left(X\right)\right)^2\right] + \mathbb{E}_{P^*}\left[\left(f\left(X\right) - f^*\left(X\right)\right)^2\right] = e\left(f^*\right) + d\left(f, f^*\right)\end{aligned}$$

Negative Log-Likelihood. Suppose $S = \Delta(\mathcal{Y})$ where $\mathcal{Y}$ is a finite set, and the loss function is $l(p, (x, y)) = -\log p(y \mid x)$ for any mapping $p : \mathcal{X} \to S$. Then,

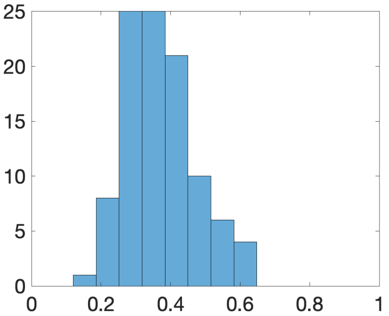
$$\begin{aligned} d(p, p^*) &= \sum_x p^*\left(x\right) \sum p^*\left(y|\, x\right) \log\left(\frac{p^*\left(y|\, x\right)}{p\left(y|\, x\right)}\right) \\ &= \mathbb{E}_{P^*}\left[\log p^*\left(y|\, x\right)\right] - \mathbb{E}_{P^*}\left[\log p\left(y|\, x\right)\right] = -e(p^*) + e(p).\end{aligned}$$

So $e(p) = e(p^*) + d(p, p^*)$ as desired.

B Supplementary Material for Application 1

B.1 Different Specifications for the Permissible Set

Larger Permissible Set $\mathcal{F}_M$. Define an alternative permissible set of mappings to include all functions $f : \mathcal{X} \to \mathbb{R}$ satisfying $f(\overline{z}, \underline{z}, p) \in [\underline{z}, \overline{z}]$. We sample 100 functions from a uniform distribution over this set and report the distribution of associated measures in the figure below:



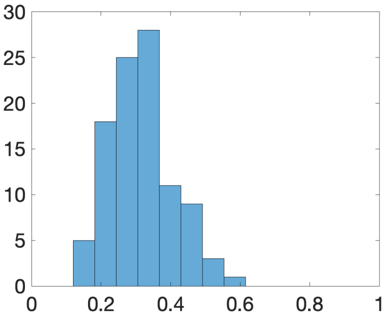
The restrictiveness here is 0.39 (with a standard error of 0.01). This is barely higher than the restrictiveness of 0.37 under our original specification of $\mathcal{F}_M$ from the main text, where we also require the permissible mappings to respect FOSD. The mean naive error is 329.24 (compared to 178.73 under the original $\mathcal{F}_M$), while the mean CPT error is 124.80 (compared to 58.21 under the original $\mathcal{F}_M$).

"Reverse" Permissible Set $\mathcal{F}_M$. We consider now a permissible set of mappings designed to violate FOSD. The set consists of all functions $f : \mathcal{X} \to \mathbb{R}$ satisfying $f(\overline{z}, \underline{z}, p) \in [\underline{z}, \overline{z}]$ and also

$$\overline{z} \geq \overline{z}', \ \underline{z} \geq \underline{z}', \ p \geq p' \quad \Longrightarrow \quad f(\overline{z}, \underline{z}, p) < f(\overline{z}', \underline{z}', p')$$

thus ensuring a violation of FOSD (since there exist multiple lotteries satisfying the LHS condition).

We sample 100 functions from a uniform distribution over this set and report the distribution of associated measures in the figure below:



The restrictiveness here is 0.31 (with a standard error of 0.01), which is lower than the restrictiveness of 0.37 under our main specification of $\mathcal{F}_M$, but still different from zero. What this tells us is that even if behavior were to violate FOSD, the CPT model would still improve substantially upon the Expected Value benchmark. The mean naive error is 427.28 (compared to 178.73 under the original $\mathcal{F}_M$), while the mean CPT error is 132.77 (compared to 58.21 under the original $\mathcal{F}_M$).

B.2 Parameter Estimates

We report below the parameter estimates for each of the models that we consider. In the first column, we report the estimated parameters on the actual data. In the second, we report the average parameter estimates for approximation of the simulated mappings.

	Real Data	Generated Mappings
CPT-$(\alpha, \beta, \delta, \gamma)$	(1.03,0.98,0.53,0.5)	(1.05,0.98,1.24,0.40)
CPT-(α, β, γ)	(0.98,1.01,0.50)	(1.06,0.95,0.38)
CPT-(δ, γ)	(0.70,0.50)	(1.12,0.24)
CPT-(α, β)	(0.98,0.99)	(1.02,0.95)

	Real Data	Generated Mappings
PCHM	$\tau = 0.5$	$\tau = 0.1$
logit level-1	$\lambda = 0.02$	$\lambda = 0.0018$
logit PCHM	$(\tau, \lambda) = (1.4, 0.11)$	$(\tau, \lambda) = (1.05, 0.02)$

B.3 Three-Outcome Lotteries

We use a set of 18 three-outcome lotteries from Bernheim and Sprenger (2020) (listed below) and evaluate the restrictiveness of Cumulative Prospect Theory for predicting certainty equivalents for these lotteries.

z_1	z_2	z_3	p_1	p_2	p_3
34	24	18	0.1	0.3	0.6
34	24	18	0.4	0.3	0.3
34	24	18	0.6	0.3	0.1
32	24	18	0.1	0.3	0.6
32	24	18	0.4	0.3	0.3
32	24	18	0.6	0.3	0.1
30	24	18	0.1	0.3	0.6
30	24	18	0.4	0.3	0.3
30	24	18	0.6	0.3	0.1
24	23	18	0.3	0.1	0.6
24	23	18	0.3	0.4	0.3
24	23	18	0.3	0.6	0.1
24	21	18	0.3	0.1	0.6
24	21	18	0.3	0.4	0.3
24	21	18	0.3	0.6	0.1
24	19	18	0.3	0.1	0.6
24	19	18	0.3	0.4	0.3
24	19	18	0.3	0.6	0.1

By convention, $z_1 \geq z_2 \geq z_3$. We estimate CPT-$(\alpha, \beta, \delta, \gamma)$ which predicts

$$w(p_1)v(z_1) + w(p_2)v(z_2) + (1 - w(p_1) - w(p_2))v(z_3)$$

for each lottery, where v and w are as defined in the main text. (Note that because these lotteries are only over gains, the parameter β is not used.)

A predictive mapping takes these 18 lotteries into certainty equivalents. The permissible mappings are defined by an adapted version of the restrictions used for two-outcome lotteries: (1) each certainty equivalent has to be in the range of the lottery outcomes, and (2) if a lottery first-order stochastically dominates another, then its certainty equivalent must be higher. We generate 100 random mappings from a uniform distribution over mappings satisfying these properties.

Below, we compare the distribution of normalized discrepancies for our binary lottery setting from Figure 5 with the distribution for these three-outcome lotteries.

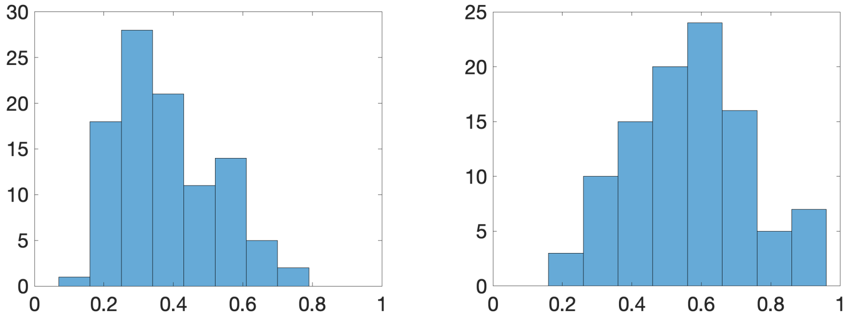


Figure 5: *Left:* Binary lotteries; *Right:* Three-outcome lotteries

The restrictiveness of CPT-$(\alpha, \beta, \delta, \gamma)$ on this set of three-outcome lotteries is 0.566, with a standard error of 0.017. Thus CPT-$(\alpha, \beta, \gamma, \delta)$ is about 1.5 times as restrictive as a model of certainty equivalents for three-outcome lotteries than as a model of certainty equivalents for binary lotteries. Even this higher restrictiveness is substantially less than what we find for models of initial play.

C Proof of Theorem 1: Asymptotics of the Completeness Estimator

C.1 Preliminary Definitions

We now introduce some definitions and notation that will be useful in the derivation of the asymptotic distribution of the CV-based completeness estimator.

C.1.1 Finite-Sample Out-of-Sample Error

Let $\mathbf{Z}_N := (Z_i)_{i=1}^N$ be a random sample of observations in a given data set, and let $Z_{N+1} \sim P^*$ denote a random variable with the same distribution P^* that is independent of $\mathbf{Z}_N$. For a given data set $\mathbf{Z}_N$ and a given model $\mathcal{F}$, we define the conditional out-of-sample error (given data set $\mathbf{Z}_N$) as

$$e_{\mathcal{F}}\left(\mathbf{Z}_N\right) := \mathbb{E}\left[l\left(Z_{N+1}, \hat{f}_{\mathbf{Z}_N}\right) \middle| \mathbf{Z}_N\right],$$

where $\hat{f}_{\mathbf{Z}_N} \in \mathcal{F}$ is an estimator, or an algorithm, that selects a mapping $\hat{f}_{\mathbf{Z}_N}$ within the model $\mathcal{F}$ based on data $\mathbf{Z}_N$. We also define the out-of-sample error, with expectation taken over different possible data sets $\mathbf{Z}_N$, as

$$e_{\mathcal{F},N} := \mathbb{E}\left[e_{\mathcal{F}}\left(\mathbf{Z}_N\right)\right].$$

From the definition of the K-fold cross-validation estimator, it can be easily shown that $\mathbb{E}\left[CV\left(\mathcal{F}\right)\right] = e_{\mathcal{F},\frac{K-1}{K}N}$. As a result, the asymptotic distribution of $CV\left(\mathcal{F}\right) - e_{\mathcal{F},\frac{K-1}{K}N}$ has been studied in the statistics and machine learning literature. Our analysis below will be based on the results in Austern and Zhou (2020) on the asymptotic distribution of $CV\left(\mathcal{F}\right) - e_{\mathcal{F},\frac{K-1}{K}N}$.

C.1.2 Joint Parametrization of $\mathcal{F}_\Theta$ and $\mathcal{F}_\mathcal{M}$

Recall that the model $\mathcal{F}_\Theta$ is parametrized by $\theta \in \Theta$, and f_θ denotes a generic function in $\mathcal{F}_\Theta$. Motivated by the applications in this paper, we assume that $\mathcal{F}_\mathcal{M}$ can be smoothly parameterized by a finite-dimensional parameter $\beta \in \mathcal{B}_\mathcal{M} \subseteq \mathbb{R}^{d_\mathcal{M}}$ and use the notation $f_{[\beta]} \in \mathcal{F}_\mathcal{M}$ to denote a generic function in $\mathcal{F}_\mathcal{M}$. Since by assumption $f^* \in \mathcal{F}_\mathcal{M}$, we can define a parameter β^* to represent it, i.e. $f_{[\beta^*]} = f^*$.

For arbitrary parameters θ and β, write

$$l_\Theta\left(Z_i, \theta\right) := l\left(Z_i, f_\theta\right), \quad l_\mathcal{B}\left(Z_i, \beta\right) := l\left(Z_i, f_{[\beta]}\right).$$

We define the estimation mappings in $\mathcal{F}_\Theta$ and $\mathcal{F}_\mathcal{M}$ by

$$\hat{\theta}\left(\mathbf{Z}_N\right) := \arg\min_{\theta\in\Theta} \frac{1}{N}\sum l_\Theta\left(Z_i, \theta\right),$$
$$\hat{\beta}\left(\mathbf{Z}_N\right) := \arg\min_{\beta\in\mathcal{B}_\mathcal{M}} \frac{1}{N}\sum l_\mathcal{B}\left(Z_i, \beta\right).$$

Let $\alpha := \left(\theta', \beta'\right)'$ denote the concatenation of the parameters $\theta \in \mathcal{F}_\Theta$ and $\beta \in \mathcal{B}_\mathcal{M}$, $\alpha^* := \left(\theta^{*'}, \beta^{*'}\right)'$ to be the parameters associated with the best mappings in $\mathcal{F}_\Theta$ and $\mathcal{F}_\mathcal{M}$, and also define

$$\hat{\alpha}\left(\mathbf{Z}_N\right) := \left(\hat{\theta}'\left(\mathbf{Z}_N\right), \hat{\beta}'\left(\mathbf{Z}_N\right)\right)'$$

$$= \arg \min_{\theta\in\Theta,\beta\in\mathcal{B}_{\mathcal{M}}} \frac{1}{N}\sum_{i=1}^{N}\left[l_{\Theta}\left(Z_i,\theta\right)+l_{\mathcal{B}}\left(Z_i,\beta\right)\right],$$

to be an estimator for α^*. Finally, define

$$\Delta l\left(Z_i;\theta,\beta\right):=l\left(Z_i,f_\theta\right)-l\left(Z_i,f_{[\beta]}\right)=l_{\Theta}\left(Z_i,\theta\right)-l_{\mathcal{B}}\left(Z_i,\beta\right).$$

C.2 Assumptions and Lemmas Based on Austern and Zhou (2020)

Assumption 5 (Conditions for Asymptotics of CV Estimator)**.**

1. *$l_{\Theta}\left(z,\theta\right)$ and $l_{\mathcal{B}}\left(z,\beta\right)$ are twice differentiable and strictly convex in θ and β.*
2. *$\mathbb{E}\left[\sup_{\theta\in\Theta} l_{\Theta}^{4}\left(Z_i,\theta\right)\right]<\infty$ and $\mathbb{E}\left[\sup_{\beta\in\mathcal{B}} l_{\mathcal{B}}^{4}\left(Z_i,\beta\right)\right]<\infty$.*
3. *There exist open neighborhoods $\mathcal{O}_{\theta^*}$ and $\mathcal{O}_{f^*}$ of $\theta^*$and β^* in Θ and $\mathcal{B}$ such that*
   - *(a) $\mathbb{E}\left[\sup_{\theta\in\mathcal{O}_{\theta^*}}\left\|\nabla_\theta l_{\Theta}\left(Z_i,\theta\right)\right\|^{16}\right]<\infty$ and $\mathbb{E}\left[\sup_{\beta\in\mathcal{O}_{\beta^*}}\left\|\nabla_\beta l_{\mathcal{B}}\left(Z_i,\beta\right)\right\|^{16}\right]<\infty$.*
   - *(b) $\mathbb{E}\left[\sup_{\theta\in\mathcal{O}_{\theta^*}}\left\|\nabla_\theta^2 l_{\Theta}\left(Z_i,\theta\right)\right\|^{16}\right]<\infty$ and $\mathbb{E}\left[\sup_{\beta\in\mathcal{O}_{\beta^*}}\left\|\nabla_\beta l_{\mathcal{B}}\left(Z_i,\beta\right)\right\|^{16}\right]<\infty$.*
   - *(c) there exist some $\delta>0$ such that $\nabla_\theta^2 l_{\Theta}\left(Z_i,\theta\right)\geq c$ a.s. and $\nabla_\beta^2 l_{\mathcal{B}}\left(Z_i,\beta\right)\geq c$ a.s. uniformly on $\mathcal{O}_{\theta^*}$ and $\mathcal{O}_{f^*}$.*

Lemma C.1 (Application of Proposition 5 of Austern and Zhou, 2020)**.** *Under Assumption 5:*

$$\sqrt{N}\left[CV\left(\mathcal{F}_\Theta\right)-CV\left(\mathcal{F}_{\mathcal{M}}\right)-\left(e_{\mathcal{F}_\Theta,\frac{K-1}{K}N}-e_{\mathcal{F}_{\mathcal{M}},\frac{K-1}{K}N}\right)\right]\xrightarrow{d}\mathcal{N}\left(0,Var\left(\Delta l\left(Z_i,f_{\theta^*},f^*\right)\right)\right).$$

Proof. Proposition 5 of Austern and Zhou (2020) establishes the asymptotic normality of cross-validation risk estimator and its asymptotic variance under parametric settings where the loss function used for training is the same as the loss function used for evaluation. Applying Proposition 5 of Austern and Zhou (2020) under Assumption 5 to θ,β and $\alpha=(\theta,\beta)$, we obtain:

$$\sqrt{N}\left(CV\left(\mathcal{F}_\Theta\right)-e_{\mathcal{F}_\Theta,\frac{K-1}{K}N}\right)\xrightarrow{d}\mathcal{N}\left(0,Var\left(l\left(Z_i,f_{\theta^*}\right)\right)\right),$$

$$\sqrt{N}\left(CV\left(\mathcal{F}_{\mathcal{M}}\right)-e_{\mathcal{F}_{\mathcal{M}},\frac{K-1}{K}N}\right)\xrightarrow{d}\mathcal{N}\left(0,Var\left(l\left(Z_i,f^*\right)\right)\right),$$

$$\sqrt{N}\left(CV\left(\mathcal{F}_\Theta\right)+CV\left(\mathcal{F}_{\mathcal{M}}\right)-e_{\mathcal{F}_\Theta,\frac{K-1}{K}N}-e_{\mathcal{F}_{\mathcal{M}},\frac{K-1}{K}N}\right)\xrightarrow{d}\mathcal{N}\left(0,Var\left(l\left(Z_i,f_{\theta^*}\right)+l\left(Z_i,f^*\right)\right)\right).$$

Using the equality $Var\left(X+Y\right)+Var\left(X-Y\right)=2Var\left(X\right)+2Var\left(Y\right)$, we then deduce that

$$\sqrt{N}\left[CV\left(\mathcal{F}_{\Theta}\right)-CV\left(\mathcal{F}_{\mathcal{M}}\right)-\left(e_{\mathcal{F}_{\Theta},\frac{K-1}{K}N}-e_{\mathcal{F}_{\mathcal{M}},\frac{K-1}{K}N}\right)\right]\xrightarrow{d}\mathcal{N}\left(0,Var\left(\Delta l\left(Z_i,f_{\theta^*},f^*\right)\right)\right).$$

□

Lemma C.2 (Application of Proposition 1 of Austern and Zhou, 2020)**.** *Under Assumption 5,*

$$\hat{\sigma}^2_{\Delta}\xrightarrow{p}Var\left(\Delta l\left(Z_i,f_{\theta^*},f^*\right)\right).$$

Proof. Applying Proposition 1 of Austern and Zhou (2020) under Assumption 5 to θ,β and $\alpha=(\theta,\beta)$:

$$\begin{aligned}\hat{\sigma}^2_{CV(\mathcal{F}_{\Theta})}:=&\frac{1}{K}\sum_{k=1}^{K}\frac{1}{J_N-1}\sum_{k(i)=k}\left(l\left(Z_i,f_{\hat{\theta}^{-k}}\right)-\frac{1}{J_N}\sum_{k(j)=k}l\left(Z_j,f_{\hat{\theta}^{-k}}\right)\right)^2\\&\xrightarrow{p}Var\left(l\left(Z_i,f_{\theta^*}\right)\right).\end{aligned}$$

and

$$\begin{aligned}\hat{\sigma}^2_{CV(\mathcal{F}_{\mathcal{M}})}:=&\frac{1}{K}\sum_{k=1}^{K}\frac{1}{J_N-1}\sum_{k(i)=k}\left(l\left(Z_i,f_{\left[\hat{\beta}^{-k}\right]}\right)-\frac{1}{J_N}\sum_{k(j)=k}l\left(Z_j,f_{\left[\hat{\beta}^{-k}\right]}\right)\right)^2\\&\xrightarrow{p}Var\left(l\left(Z_i,f^*\right)\right).\end{aligned}$$

and

$$\begin{aligned}&\hat{\sigma}^2_{CV(\mathcal{F}_{\Theta})+CV(\mathcal{F}_{\mathcal{M}})}\\:=&\frac{1}{K}\sum_{k=1}^{K}\frac{1}{J_N-1}\cdot\sum_{k(i)=k}\\&\left(l\left(Z_i,f_{\hat{\theta}^{-k}}\right)+l\left(Z_i,f_{\left[\hat{\beta}^{-k}\right]}\right)-\frac{1}{J_N}\sum_{k(j)=k}\left[l\left(Z_j,f_{\left[\hat{\beta}^{-k}\right]}\right)+l\left(Z_i,f_{\hat{\theta}^{-k}}\right)\right]\right)^2\\\xrightarrow{p}&Var\left(l\left(Z_i,f_{\theta^*}\right)+l\left(Z_i,f^*\right)\right),\end{aligned}$$

Hence:

$$\begin{aligned}\hat{\sigma}^2_{\Delta}&=2\hat{\sigma}^2_{CV(\mathcal{F}_{\Theta})}+2\hat{\sigma}^2_{CV(\mathcal{F}_{\mathcal{M}})}-\hat{\sigma}^2_{CV(\mathcal{F}_{\Theta})+CV(\mathcal{F}_{\mathcal{M}})}\\&\xrightarrow{p}2Var\left(l\left(Z_i,f_{\theta^*}\right)\right)+2Var\left(l\left(Z_i,f^*\right)\right)-2Var\left(l\left(Z_i,f_{\theta^*}\right)+l\left(Z_i,f^*\right)\right)\\&=Var\left(\Delta l\left(Z_i,f_{\theta^*},f^*\right)\right)\end{aligned}$$

□

C.3 Proof of Asymptotic Normality of $\hat{\kappa}^*$

Lemma C.1 characterizes the limit distribution of

$$\sqrt{N}\left[CV\left(\mathcal{F}_\Theta\right) - CV\left(\mathcal{F}_\mathcal{M}\right) - \left(e_{\mathcal{F}_\Theta,\frac{K-1}{K}N} - e_{\mathcal{F}_\mathcal{M},\frac{K-1}{K}N}\right)\right]$$

which we now show is also the limit distribution of

$$\sqrt{N}\left[CV\left(\mathcal{F}_\Theta\right) - CV\left(\mathcal{F}_\mathcal{M}\right) - \left(e_{\mathcal{F}_\Theta} - e_{\mathcal{F}_\mathcal{M}}\right)\right].$$

To see this, notice that

$$\begin{aligned}
e_{\Theta,\frac{K-1}{K}N} - e_{\mathcal{F}_\Theta} &= \mathbb{E}\left[l_\Theta\left(Z_i,\hat{\theta}^{-k(i)}\right) - l\left(Z_i,\theta^*\right)\right] \\
&= \mathbb{E}\left[\nabla l_\Theta\left(Z_i,\theta^*\right)\cdot\left(\hat{\theta}^{-k(i)}-\theta^*\right) + \left(\hat{\theta}^{-k(i)}-\theta^*\right)'\nabla^2 l_\Theta\left(Z_i,\tilde{\theta}\right)\cdot\left(\hat{\theta}^{-k(i)}-\theta^*\right)\right] \\
&= 0 + \mathbb{E}\left[\left(\hat{\theta}^{-k(i)}-\theta^*\right)'\nabla^2 l_\Theta\left(Z_i,\tilde{\theta}\right)\cdot\left(\hat{\theta}^{-k(i)}-\theta^*\right)\right] \\
&= \frac{1}{N-J_N}\mathbb{E}\left[\sqrt{N-J_N}\left(\hat{\theta}^{-k(i)}-\theta^*\right)'\nabla^2 l_\Theta\left(Z_i,\tilde{\theta}\right)\cdot\sqrt{N-J_N}\left(\hat{\theta}^{-k(i)}-\theta^*\right)\right] \\
&= c\frac{1}{N-J_N} + o\left(\frac{1}{N-J_N}\right) \\
&= c\frac{K}{K-1}\cdot\frac{1}{N} + o\left(\frac{1}{N}\right)
\end{aligned}$$

and hence

$$\sqrt{N}\left(e_{\Theta,\frac{K-1}{K}N} - e_\Theta\right) = o_p\left(1\right).$$

Similarly, $\sqrt{N}\left(e_{\mathcal{F}_\mathcal{M},\frac{K-1}{K}N} - e_{\mathcal{F}_\mathcal{M}}\right) = o_p\left(1\right)$.

Hence:

$$\sqrt{N}\left[CV\left(\mathcal{F}_\Theta\right) - CV\left(\mathcal{F}_\mathcal{M}\right) - \left(e_{\mathcal{F}_\Theta} - e_{\mathcal{F}_\mathcal{M}}\right)\right] \xrightarrow{d} \mathcal{N}\left(0, Var\left(\Delta l\left(Z_i, f_{\theta^*}, f^*\right)\right)\right).$$

Then, by Lemma C.2, Assumption 3 and the continuous mapping theorem, we have

$$\frac{\sqrt{N}\left(\hat{\kappa}^* - \kappa^*\right)}{\hat{\sigma}_{\hat{\kappa}^*}} \xrightarrow{d} \mathcal{N}\left(0,1\right).$$