Inference of Preference Heterogeneity from Choice Data

Annie Liang[*]

First Version: October 4, 2016
This Version: August 15, 2018

Abstract

Suppose that an analyst observes inconsistent choices from either a single decision-maker, or a population of agents. Can the analyst determine whether this inconsistency arises from choice error (imperfect maximization of a single preference) or from preference heterogeneity (deliberate maximization of multiple preferences)? I model choice data as generated from imperfect maximization of a small number of preferences. The main results show that (a) simultaneously minimizing the number of inferred preferences and the number of unexplained observations can exactly recover the number of underlying preferences with high probability; (b) simultaneously minimizing the *richness* of the set of preferences and the number of unexplained observations can exactly recover the choice implications of the decision maker's underlying preferences with high probability.

1 Introduction

Let X be a finite set of choice alternatives, and consider an analyst who observes choices (by either a single decision-maker, or a population of subjects) from various subsets of X. Empirical choice data of this nature is often inconsistent, and cannot be explained as perfect maximization of a single preference.

There are two different perspectives for how to interpret such inconsistency. One view is that inconsistency emerges from *preference heterogeneity.* There is abundant

[*] University of Pennsylvania. I am especially grateful to Jerry Green. I would also like to thank David Ahn, Emily Berger, Gabriel Carroll, Ian Crawford, Aluma Dembo, Drew Fudenberg, Ben Golub, David Laibson, Jay Lu, Erik Madsen, Eric Maskin, Jose Montiel Olea, Krishna Pendakur, Gleb Romanyuk, Andrei Shleifer, Ran Shorrer, Ran Spiegler, Tomasz Strzalecki, and Yufei Zhao for useful comments and discussions.

evidence that choices depend on details about the choice context—for example, Einav et al. (2012) find that just over 30% of their subject pool makes decisions across six different financial domains that can be rationalized using a common risk preference. Additionally, choice data aggregated over a population of decision-makers often exhibits cross-sectional heterogeneity in preferences—for example, Crawford and Pendakur (2012) study household consumption decisions over different kinds of milk, and find that no more than two-thirds of observations in their data set can be rationalized using a single utility function. In both of these cases, choice inconsistencies are understood to reflect intentional maximization which is welfare-relevant.

Another view is that inconsistencies reflect *errors*. For example, the decision-maker may be inattentive, or the analyst may make mistakes while recording observations. In these cases, inconsistency reflects choices that are welfare-reducing and not indicative of genuine preference.

Preference heterogeneity and error are distinct sources of inconsistency, with different implications for welfare-assessment and for prediction. To take a stark example, compare two hypothetical choice data sets: one generated by *perfect maximization of two different preferences*, and one generated by *maximization of a single preference with trembling error*. Application of classical approaches such as Houtman and Maks (1985) can fail to distinguish between these data sets, especially if the same fraction of both data sets is rationalizable using a single preference. Nevertheless, the underlying choice mechanics are quite different: the inconsistency represented in the first data set can be expected to be stable across future observations, while the inconsistency represented in the second is idiosyncratic.

A basic question then is *how many* choice domains or subpopulations are present in the data, where a special case of interest is whether there is evidence of multiple preferences or a single preference with error. (The question of "how many preferences" is pursued, for example, in Crawford and Pendakur (2012) in the case of household consumption decisions and Dean and Martin (2010) for individual choices over lotteries.)

At two ends for interpreting the data are the classical approaches of Houtman and Maks (1985) and Kalai et al. (2002), both of which rule out one of the two sources of inconsistency described above. Specifically, we can find a "best-fit" single preference (rationalizing the largest fraction of observations) and interpret the remaining observations as choice errors (Houtman and Maks, 1985), or find the smallest number of preferences that perfectly rationalizes the choice data (Kalai et al., 2002). When preference multiplicity and choice errors are simultaneously present in the data, the Houtman and Maks (1985) solution underestimates the number of preferences (since by design it assumes a single preference), while the Kalai et al. (2002) solution overestimates (since choice errors are attributed preferences). These misinterpretations have potentially large consequences both for welfare evaluation and also for out-of-sample predictions.[1]

The purpose of this paper is to develop a method to determine the "best" intermediate solution. I consider data generated according to a (generalized) random utility model. The decision-maker (DM) chooses from choice set $A \subseteq X$ by sampling a preference according to a distribution μ_A, and maximizing the sampled preference. I suppose that each μ_A is in fact a perturbation of a "sparse" μ_A^*, whose support is a small number of preferences K that are constant across choice sets.[2] The goal is to recover from the choice data the underlying number of preferences K.

The proposed approach, presented in Section 4, minimizes a weighted sum of the number of preferences attributed to the decision-maker, and the number of unexplained observations (choices that cannot be rationalized by any of the recovered preferences). Intuitively, the approach imposes a cost on each recovered preference, so that a preference is recovered if and only if it explains sufficiently many observations that would otherwise be considered error. The classic Houtman and Maks (1985) and Kalai et al. (2002) solutions are returned for special choices of weights—the former is returned when the cost of preferences relative to unexplained observations is sufficiently high, and the latter is returned when the cost of preferences relative to unexplained observations is sufficiently low.

The main result in Section 5 provides a set of weights (which depend on primitives of the choice model) given which the proposed approach exactly recovers the "true" number of preferences with sufficiently many observations.[3] Informally, these conditions require that the K preferences are sufficiently differentiated in the sampled data, so that choice inconsistencies that arise from genuine preference heterogeneity resemble other inconsistencies in the choice data, whereas choice inconsistencies that arise from error appear idiosyncratic. The Kalai et al. (2002) approach is shown to recover the number of underlying preferences when the probability of choice error is zero; to the best of my knowledge, this is the first statistical justification for the Kalai et al. (2002) approach. Additionally, the special case of discerning between choice data that is generated by imperfect maximization of a single preference, versus choice data that reflects multiplicity of preference, is considered in Section 5.4.

The set of weights which allow for recovery depend on primitives of the choice model. Next, I explain a way in which we can "test" particular assumptions about

[1] See Appendix A for examples in which use of these approaches to predict choice behaviors leads to suboptimally prediction accuracy, and the magnitude of the potential gains is substantial.

[2] This paper takes a nonparametric, or "model-free" approach—see prior work in Varian (1982), Famulari (1995), Houtman and Maks (1985), and Kalai et al. (2002)).

[3] I assume that the DM may be presented with the same choice problem multiple times.

these unobservables based on the data. Intuitively, the main theorem provides an *interval* of weights that recover the same solution. Thus, if our inferred solution is the true number of underlying preferences, then it should be robust to nearby choices of weights. We can determine from the data the actual range of values of weights over which our inferred solution remains stable, and Corollary 1 shows how we can use this range to bound the key primitives (the extent of differentiation of the underlying preferences, and the probability of error).

Section 7 revisits an analysis conducted in Crawford and Pendakur (2012), in which the Kalai et al. (2002) approach is used to discover the number of preference types among 500 subjects. Crawford and Pendakur (2012) find that five preferences are needed to perfectly rationalize their data set. I show how the proposed approach can be used to identify some of these preferences as noise.

Section 8 turns to the question of recovering the preferences themselves. Inference of multiple preferences from choice data is an ill-posed problem, and Section 5.1 presents several negative results that help to clarify the reasons for this. In Proposition 1, I show that most sets of orderings are indistinguishable based on their choice implications, so that even in the absence of choice error, most sets of multiple preferences cannot be recovered. This result is very much in the spirit of Ambrus and Rozen (2013), which studies a broad (but different) class of multi-self models and shows that these models have no testable implications without prior restrictions on the number of selves involved in a decision.[4]

In view of these results, I suggest that a more appropriate object of recovery is the set of *choice implications* of the decision maker's preferences—that is, the choice observations that are consistent with maximization of one of these preferences. I define equivalence classes for sets of preferences, where two sets belong to the same equivalence class if they have the same choice implications, and ask whether we can recover the equivalence class to which the true set of preferences belongs. Section 8.2 shows that this is indeed possible, but that penalizing the number of inferred preferences is not the appropriate criterion for this goal. Intuitively, penalizing only the number of preferences results in inference of sets of preferences whose choice implications are as diverse as possible. I propose an alternative criterion, minimizing a weighted sum of the number of unexplained observations and the *richness* of the set of preferences, as measured through the number of unique choice implications. Proposition 2 shows that under certain conditions on the choice model described above, this approach will exactly recover the equivalence class of choice implications

[4] Ambrus and Rozen (2013) study several different choice-set independent aggregation rules over preferences, whereas I consider a specific aggregation rule (in which one preference is assigned "dictator") that varies across choice problems.

containing those of the true model.

Finally, Section 8.3 considers a richer kind of data set, which includes auxiliary information on the choice contexts active during different observations. I show that with this additional information, we can (under certain conditions) recover the exact set of preferences.

Taken together, these results suggest that appropriately penalizing the complexity of the inferred choice model—for example, via the number of preferences used or the number of choice implications—can be leveraged towards recovering stable features of preference from inconsistent choice data.

2 Example

In an adaptation on the Luce and Raiffa dinner (Luce and Raiffa, 1957), suppose that a large number of consumers are observed to choose entrées from different restaurant menus. Each menu includes at least two entrées from the set $\{x_1, x_2, \ldots, x_N\}$; additionally, a special of frog legs (denoted x_{N+1}) is sometimes included.

As in Sen (1993), the presence of frog legs signals a high quality chef and encourages consumers to choose entrées that are harder to prepare. Suppose that entrées are ordered $x_1, x_2, \ldots, x_N$ from least to most difficult to prepare. When frog legs are present on the menu A, consumers choose each entrée $x \in A$ with probability

$$c_1(x|A) = \frac{e^{-\gamma u_1(x)}}{\sum_{x' \in A} e^{-\gamma u_1(x')}} \qquad \text{when } x \neq x_{N+1}$$

where $\gamma \in \mathbb{R}_+$ is a logit parameter, and the utility function $u_1(x_k) = k$ assigns a higher payoff to entrées that are more difficult to prepare. Fix $c_1(x|A) = 0$ for every $x \notin A$ and also $c_1(x_{N+1}|A) = 0$, so that frog legs are never themselves chosen.

When frog legs are *not* present on the menu, consumers' choices follow the logit choice rule

$$c_2(x|A) = \frac{e^{-\gamma u_2(x)}}{\sum_{x' \in A} e^{-\gamma u_2(x')}} \qquad \forall\, x \in A$$

where the utility function $u_2(x_k) = N - k + 1$ assigns higher payoffs to entrées that are easier to prepare. Fix $c_2(x|A) = 0$ for every $x \notin A$.

For concreteness, fix the logit parameter to be $\gamma = 2$, so that consumers choose the most preferred alternatives (under respectively u_1 or u_2) with high probability. For example, given the choice set $\{x_1, x_2, x_3\}$, the most preferred alternative x_1 is chosen with probability 0.86, alternative x_2 with probability 0.12, and alternative x_3 with probability 0.02.

Choices from n menus are observed, where menus are sampled uniformly at random (with repetitions permitted) from the set of all menus that contain at least two entrées from $\{x_1, \ldots, x_N\}$. I make minimal assumptions about the analyst's knowledge about the consumers' choice rule; in particular, he does not know c_1 and c_2, and does not know that there are two choice rules. He simply observes, and seeks to rationalize, the generated choices $\{(x_i, A_i)\}_{i=1}^n$.

The key feature of this example is that there are two distinct reasons why observed choices are unlikely to be consistent with maximization of any single ordering: first, consumers use different choice rules in different observations; second, choice is stochastic (i.e. maximization is imperfect). At extremes, we can interpret the data in a way that rules out either of the two sources of inconsistency. For example, we can insist on a single preference and find the "best-fit" preference, interpreting the remaining choice observations as error. Another alternative is to ascribe to consumers as many preferences as are needed to perfectly rationalize the data. Neither of these choices matches our narrative above, and both lead to misinterpretations of the data with consequences for welfare evaluation and prediction (discussed further below). The key question for this paper is how to determine that there are two primary preferences.

Consider $N = 10$, so that there are ten main entrées. The choice data can be represented using an *error-preference tradeoff graph*: For each number of preferences k, we find the percentage of the choice data that *cannot* be rationalized using k preferences.[5]

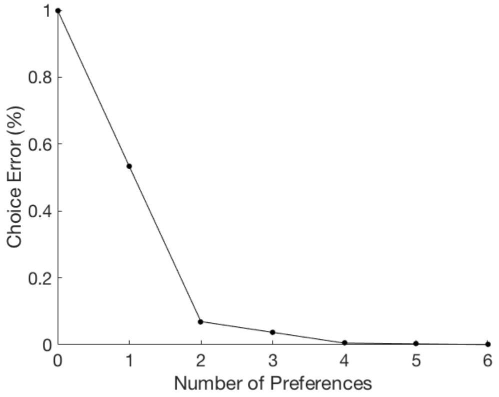


Figure 1: *Error-Preference Tradeoff Graph.* The (expected) fraction of the choice data that cannot be rationalized using any set of k preferences.

[5] In practice, we would use the fraction of the actual choice data that is unexplained (interpret these as *choice errors*); for the purpose of this illustration, Figure 1 reports the expected fractions.

From this figure, we see that it is possible to rationalize (in expectation) 47% of the choice data using a single preference, while two preferences can rationalize almost all of the data (93%). In contrast, the addition of a third ordering increases the completeness of explanation by only 3%, and the addition of the fourth preference contributes even less.[6] Thus, each preference *up to* the second preference helps to rationalize a large fraction of the data, while each preference *after* the second preference contributes only a marginal improvement in explanation. The substantial drop in marginal explanation after the second ordering suggests the presence of two structural preferences, with additional trembling noise.

This is not the interpretation of either the Houtman and Maks (1985) or the Kalai et al. (2002) approaches. Direct application of Houtman and Maks (1985) produces an (expected) inconsistency measure of 53%. This implies a quite irrational DM, when in fact the DM perfectly maximizes in most choice observations. Direct application of Kalai et al. (2002) errs in the other direction, overestimating the number of active preferences as six. This can lead to errors in policy interventions: with high probability, some (x, A) will be observed where x is neither $\succ_1$- nor $\succ_2$-maximal in A. Prescription of alternative x for the DM is guaranteed not to be the optimal solution, but the Kalai et al. (2002) approach does not reveal this.

We can also see, informally, that prediction of future choices based on these two approaches can be misguided.[7] The best *single* preference recovered under Houtman and Maks (1985) predicts incorrectly in a majority of new choice problems, while some of the six preferences recovered under the Kalai et al. (2002) approach may perform worse than random guessing. A more rigorous treatment of the topic of out-of-sample prediction accuracy is deferred to Appendix A.

This discussion highlights some cases in which the outputs of the Houtman and Maks (1985) and Kalai et al. (2002) approaches are not well-suited to the analyst's goals. The key question is then how to identify an optimal solution intermediate to the two approaches (in this case, identifying two preferences). The sections below propose a method for doing this.

[6] The expected fraction of error given k preferences is: 0.5335 for $k = 1$, 0.0694 for $k = 2$, 0.0372 for $k = 3$, 0.0050 for $k = 4$, 0.0027 for $k = 5$, and 0.0004 for $k = 6$.

[7] This discussion is informal, since we have not fixed a model for choosing which preference to use in predicting new choice observations, which is required for evaluating multiple-preference models. The additional analysis pursued in Appendix A uses an extension described in Section 8.3 that makes these out-of-sample comparisons possible.

3 Conceptual Framework

3.1 Choice Model

Let X be a finite set of alternatives. A *preference* P is a strict linear ordering over X, and the set of all preferences is denoted $\mathscr{P}$. A *choice set* is a subset $A \subseteq X$, and 2^X denotes the set of all possible choice sets. I will primarily use the interpretation in which choices are generated by a single individual, although the framework applies also to the setting in which choices are generated by a population of subjects.

The decision-maker (DM) chooses from choice set A by sampling a preference according to a distribution $\mu_A \in \Delta(\mathscr{P})$, and maximizing the sampled preference. This corresponds to a standard generalization of the *random utility model (RUM)*, where the decision-maker's distribution over preferences is permitted to vary across choice sets. I refer to $\mu = (\mu_A)_{A \in 2^X}$ as the decision-maker's RUM. Then, the probability that alternative x is selected from choice set A is

$$c(x|A) = \mu_A(\{P : x \text{ is } P\text{-maximal in } A\}). \tag{1}$$

An analyst observes the decision-maker's choices from n choice sets, sampled (with replacement) from a distribution $\pi \in \Delta(2^X)$. A *choice observation* is a pair (x, A), corresponding to choice of alternative x from set A, and the observed data is a multiset of choice observations $D = \{(x_i, A_i)\}_{i=1}^n$. For simplicity, I will refer to D as simply a *set* of choice observations, although it should be understood that the same observation may appear multiple times. Finally, the ex-ante probability of observing any (x, A) (taking into account both the randomness in which choice sets are presented to the DM, and also the randomness in his choice) is

$$\nu(x, A) = \pi(A)c(x|A). \tag{2}$$

Notice that by explicitly modeling the sampling of choice sets, I depart from a standard assumption that the analyst knows the stochastic choice rule c, and thus has available to him an "idealized" data set where choices are made infinitely often from each choice set. In this paper, the analyst observes only a finite number of choices. If π assigns positive probability to every choice set, then the idealized data is returned as a limiting case when we take the number of observations to infinity.

3.2 Separation of Preference from Error

Consider the general choice framework described in the previous section. When the distributions μ_A are not degenerate, then violations of the Independence of Irrelevant

Alternatives axiom are expected.[8]

There are two different perspectives for how to interpret these choice inconsistencies. One view is that the inconsistencies emerge from *preference heterogeneity*. For example, it may be that preference depends on unobserved features of the environment beyond the choice set, so that different preferences are cued in different choice observations. A related explanation is that if choices are generated by a population of decision-makers, then inconsistencies may reflect cross-sectional heterogeneity in preferences across the population. According to both of these interpretations, the randomness over outcomes is reflective of intentional maximization which is welfare-relevant. A second view is that these inconsistencies describe *measurement errors* or *choice errors*, which are welfare-reducing and not indicative of genuine preference. The choice model described previously permits both kinds of inconsistency to be simultaneously present.

In general, it will not always be possible to separate preference heterogeneity from error; even conceptually. This paper studies a setting in which there is a *small* set of preferences that are maximized *most* of the time, and suggests that preference heterogeneity and error can be meaningfully distinguished in this case. In particular, we may think of the small stable set of preferences as the "true" preferences—reflecting, for example, different sub-populations or different choice contexts—and the choices that are inconsistent with these preferences as error.

3.3 Underlying "Sparse" Choice Model

Formally, I consider the setting in which the RUM μ is well-approximated by an underlying μ^*, supported on a "sparse" set of preferences.

Specifically, suppose that the DM has a set $\mathcal{P}$ consisting of K preferences. In the absence of choice error, his choice from set A corresponds to maximization of a preference sampled from a distribution μ^*_A, where $\mu^*_A(\mathcal{P}) = 1$. (Interpret non-degenerate distributions μ^*_A as reflecting variation in the activation of preferences. For example, if preferences correspond to different choice contexts, then μ^*_A is the empirical distribution of contexts for the choice problem A.) Thus, when K is small, the RUM $\mu^* = (\mu^*_A)_{A \in 2^X}$ is supported on only a small number of preferences, relative to the complete set of preference orderings $\mathscr{P}$. In analogy to (1) and (2), define $c^*(x|A) = \mu^*_A(\{P : x \text{ is } P\text{-maximal in } A\})$ for the stochastic choice rule associated with μ^*, and $\nu^*(x, A) = \pi(A)c^*(x|A)$ for the frequency of observation of (x, A) under RUM μ^* and sampling distribution π.

[8] Here, and throughout the paper, I refer to the classic (deterministic) version of IIA. Naturally, violations of the stochastic version of IIA may also be present.

We do not observe choices generated under RUM μ^*, but rather choices generated under its perturbation μ. The relationship between these RUMs is given as follows: For each choice set A, there is a map $g_A : \mathscr{P} \to \Delta(\mathscr{P})$ such that

$$\mu_A = \mu^*_A G_A$$

where μ_A and μ^*_A are $1 \times |\mathscr{P}|$ vectors (choose an arbitrary indexing of preferences) and G_A is the $|\mathscr{P}| \times |\mathscr{P}|$ Markov matrix associated with g_A.[9] It is important that choice errors do not occur frequently; formally, suppose there is a (uniform) bound on probability of error p such that the diagonal entries in G_A are at least $1-p$ (for every choice set A). Informally, this guarantees that the "right" preference is maximized most of the time. I will refer to p throughout as the *probability of error*.

When K and p are both small, as is the primary case of interest, then most choices under the RUM μ correspond to maximization of a small number of preferences. Section 5 provides conditions on the sampling distribution π, the probability of error p, and the underlying RUM μ^* under which recovery of the number of preferences K is possible.[10]

Note the following special cases of the model:

Example 1. If $K = 1$, so that there is only one underlying preference, then each μ^*_A is degenerate on that preference and the observed choice data corresponds to imperfect maximization of a single preference ordering.

Example 2. If $K > 1$ but each μ^*_A is degenerate, then we return the multiple preference model introduced in Kalai et al. (2002), where different preferences are cued in different choice problems.

Throughout, I will take the perspective that μ^* is the DM's true RUM, and K describes the cardinality of the DM's true set of preferences. An alternative interpretation is that the DM's "true" RUM is μ, but we prefer a more parsimonious description—specifically, an approximate representation by an RUM which assigns positive probability only to a small set of preferences. From this perspective, our problem is that of how many preferences are needed to explain most of the choice data generated under μ.

Another interpretation of the proposed framework, building on Rubinstein and Salant (2008) and Bernheim and Rangel (2009), is the following. Let $\mathscr{C}$ be a set of contexts that are relevant to the DM's preference but unobserved by the analyst.[11]

[9] Index the preferences $P_1, \dots P_{N!}$. The i-th column of G_A is $g_A(P_i)$.

[10] See Section 5.4 for application of the approach in an example in which the probability of p is large.

[11] These are called *frames* in Rubinstein and Salant (2008) and *ancillary conditions* in Bernheim and Rangel (2009).

Each context is associated with a preference; in a slight abuse of notation, let P_C be the preference associated with context $C \in \mathscr{C}$. Choice sets A are associated with a distribution μ_A over contexts, and the probability of observation of choice of x from A is given by

$$c(x|A) = \mu_A(\{C : x \text{ is } P_C\text{-maximal in } A\}).$$

The main question of this paper is whether it is possible to recover the number of contexts, given observation of pairs (x, A), when the contexts i are hidden. Section 8.3 pursues this interpretation further, considering the case in which contexts are *observed*.

4 Analysis of the Choice Data

Fix a data set $D = \{(x_i, A_i)\}_{i=1}^n$. A *multiple preference rationalization* of this data is any set of preference orderings $\mathcal{P} \subseteq \mathscr{P}$. The number of observations in D that are inconsistent with all preferences in $\mathcal{P}$ is

$$\varepsilon(D, \mathcal{P}) := \#\{(x, A) \in D : x \text{ is not } P\text{-maximal in } A \text{ for any } P \in \mathcal{P}\}.$$

I call this the number of implied choice errors.[12] Say that $\mathcal{P}$ constitutes a *perfect rationalization* of D if $\varepsilon(D, \mathcal{P}) = 0$. If we restrict to rationalizations that consist of k orderings, the minimal number of implied choice errors in D is

$$\varepsilon_k(D) := \min_{|\mathcal{P}|=k} \varepsilon(D, \mathcal{P}).$$

Say that the data set D is *k-rationalizable* (Kalai et al., 2002) if there is some set of k orderings that perfectly rationalizes D, so that $\varepsilon_k(D) = 0$.

It is useful to represent D as the linear interpolation of points in $\{(k, \varepsilon_k(D))\}_{k\in\mathbb{Z}_+}$. Henceforth, I will refer to this as an *Error-Preference Tradeoff Graph*. If we consider the convex hull of this graph, then each point represents a particular weighted minimization of the number of orderings k (ascribed to the DM), and the number of implied choice errors $\varepsilon_k(D)$.

Formally, for every value λ (which defines a weighting over these objectives), there is a corresponding solution:

Definition 4.1. *For every $\lambda \in \mathbb{R}_+$ and data set D, define*

$$K_\lambda^*(D) = \operatorname*{argmin}_{k\in\mathbb{Z}_+} \left[k + \lambda\varepsilon_k(D)\right]. \tag{3}$$

[12] Naturally, this is only one of many possible definitions for choice error. In particular, other notions of error may be preferred assuming different models of preference aggregation (see e.g. the multiple-ordering models of Rubinstein and Salant (2006), Fudenberg and Levine (2006), Green and Hojman (2007), Manzini and Mariotti (2007, 2009)).

This solution is depicted in the figure below. When there are multiple solutions to the problem above, I will take $K^*_\lambda(D)$ to mean the smallest value of k in the minimizing set.

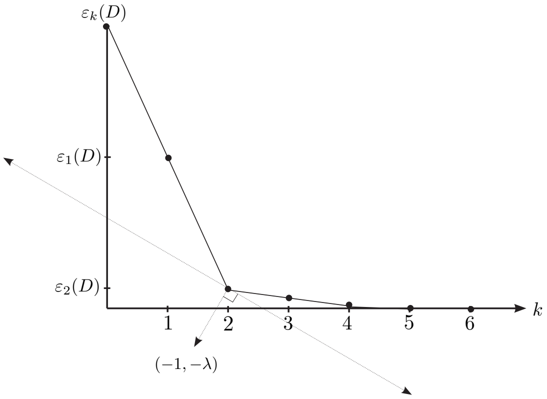


Figure 2: Define E to be the set of points lying above the linear interpolation of $\{(k, \varepsilon_k(D))\}_{k\in\mathbb{Z}_+}$. Then, the problem in (3) returns a solution with k orderings if and only if the line with normal vector $(-1, -\lambda)$ properly supports E at $(k, \varepsilon_k(D))$.

Intuitively, $1/\lambda$ is the "cost" of each ordering, so that an ordering is attributed to the DM if and only if it explains at least $1/\lambda$ observations that would otherwise be interpreted as choice error. As $\lambda \to 0$, the cost of errors becomes increasingly small relative to the cost of orderings, so the analyst prefers to attribute a single ordering to the DM and interpret the unexplained observations as choice errors. As $\lambda \to \infty$, the cost of choice errors becomes increasingly large relative to the cost of orderings, so the analyst prefers to use as many orderings as necessary to perfectly rationalize the data.

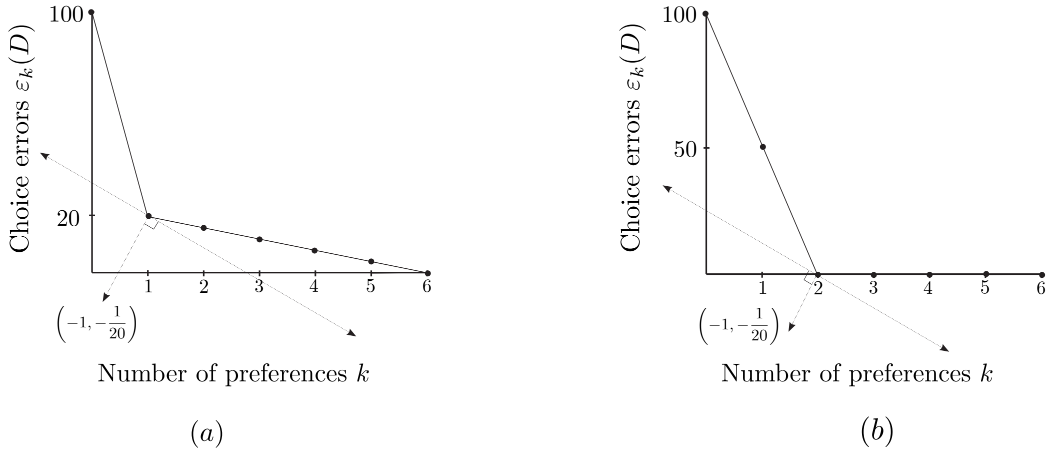
In particular, if $\lambda < \frac{1}{\varepsilon_1(D)}$ (where, recall, $\varepsilon_1(D)$ is the necessary number of unexplained observations if the DM is modeled with a single preference), then the approach returns the Houtman and Maks (1985) solution, and if $\lambda > 1$, then the approach returns the Kalai et al. (2002) solution.

Observation 1. *For every data set D,*

(a) $K^*_\lambda(D) = 1$ for every $\lambda < \frac{1}{\varepsilon_1(D)}$, and

(b) $K^*_\lambda(D) = L$ for every $\lambda > 1$, where L is the smallest integer such that D is L-rationalizable.

The intervals provided in Observation 1 are sufficient but not necessary for recovery of the Houtman and Maks (1985) and Kalai et al. (2002) solutions. In particular, it misleadingly suggests that *different* sets of choices for λ are persistently associated with each of these approaches. Indeed, the same choice of λ can recover either solution, depending on the realized data. For example, choice of $\lambda = 1/20$ selects $K_\lambda^* = 1$ (the Houtman and Maks (1985) solution) in panel (a) of the figure below, and selects $K_\lambda^* = 2$ (the Kalai et al. (2002) solution) in panel (b).



(a)

(b)

These selections align with a heuristic intuition that the data set in (a) corresponds to imperfect maximization of a single ordering, while the data set in (b) corresponds to perfect maximization of two orderings. The main results below formalize these intuitions, relating the "optimal" choice(s) of λ to the primitives of the choice model described in Section 3.

5 Recovering the Number of Preferences

5.1 No Error Baseline: $\mu = \mu^*$

Consider first the recovery problem for an idealized baseline in which the decision-maker's RUM is exactly μ^*, so that there are no choice errors. In fact, recovery of the number of preferences K is not guaranteed even in this setting. The key condition needed for recovery of K is that preferences are sufficiently differentiated in the data. This differentiation depends *jointly* on the sampling procedure π and also on the underlying RUM μ^*.

Basic challenges to recovery are illustrated in Examples 3-5 below. In Example 3, recovery is not feasible because the DM's preferences are not sufficiently differentiated by their choice implications; in Example 4, recovery is not feasible because a preference is insufficiently sampled; and in Example 5, recovery is not feasible because preferences agree on the sampled choice sets. The exercise of recovery is not obviously meaningful in these cases, and such settings will be subsequently ruled out.

Example 3. The set of choice alternatives is $X = \{x_1, x_2, x_3\}$ and the DM's preferences are $\mathcal{P} = \{P_1, P_2, P_3\}$, where

$$\begin{aligned} &x_1 \, P_1 \, x_2 \, P_1 \, x_3 \\ &x_1 \, P_2 \, x_3 \, P_2 \, x_2 \\ &x_3 \, P_3 \, x_2 \, P_3 \, x_1 \end{aligned}$$

Notice that every choice observation consistent with maximization of some ordering in $\mathcal{P}$ is also consistent with maximization of an ordering in $\mathcal{P}' = \{P_1, P_3\}$. Fix any RUM μ^* supported on $\mathcal{P}$, and any sampling distribution π over choice sets. Then, for all data sets D generated under this model, there is no value of λ given which $K^*_\lambda(D) = 3$.

Example 4. The DM's preferences are $\mathcal{P} = \{P_1, P_2\}$, but only the first preference is sampled; that is, $\mu^*_A(P_1) = 1$ for every choice set A. For all data sets D generated under this model, there is no value of λ given which $K^*_\lambda(D) = 2$.[13]

Example 5. The DM's preferences are $\mathcal{P} = \{P_1, P_2\}$ where

$$\begin{aligned} &x_1 \, P_1 \, x_2 \, P_1 \, x_3 \\ &x_1 \, P_2 \, x_3 \, P_2 \, x_2 \end{aligned}$$

The sampling procedure π puts probability 1 on the choice set $\{x_1, x_2, x_3\}$. Since P_1 and P_2 agree on this choice set, for every RUM μ^* (supported on $\mathcal{P}$), every data set D generated under this model, and every choice of λ, the proposed approach yields $K^*_\lambda(D) = 1$.

These examples highlight that sufficient differentiation of preferences in the data requires first that preferences have different choice implications, and second that there is opportunity for these choice implications to be observed (namely, that the corresponding choice problems and preferences are sampled in the data). The previous examples have the property that the number of underlying preferences cannot be recovered from *any* number of choice observations. Each of these obstacles to recovery may also appear in a more moderate degree: for example, some preference may be rarely sampled, causing its choice implications to appear rarely in the data.

Following, I define a measure for how *differentiated* the K underlying preferences are in the choice data. This measure is defined as a property of the primitives π and

[13] I am grateful to an anonymous referee for suggesting this example.

μ^*. Sufficient differentiation in preferences serves a dual role: it simultaneously makes recovery of the number of preferences possible, and it also justifies consideration of this number as an object of interest. Note that in each of the previous examples, recovery of the number of "true" preferences is an arguably misguided exercise—for example, the set of three preferences $\mathcal{P}$ in Example 3 is a needlessly complex way to rationalize choice data that can also be explained using $\mathcal{P}' \subset \mathcal{P}$.

As a preliminary step, I first define a generalization of IIA:[14]

Definition 5.1. *For any integer k, say that choice observations $\{(x_i, A_i)\}_{i=1}^k$ are in* k-violation of IIA *if*

(1) $x_i \neq x_j$ for every $i \neq j$,

(2) $x_i \in \bigcap_{j=1}^k A_j$ for every $i = 1, \ldots, k$.

The first condition requires that every chosen alternative is different, and the second condition requires that each of the chosen alternatives is available in all of the observed choice sets. An immediate implication is that the set of choice observations $\{(x_i, A_i)\}_{i=1}^k$ cannot be rationalized using fewer than k orderings (without introducing choice error). Notice also that every pair of choice observations from $\{(x_i, A_i)\}_{i=1}^k$ constitutes a (standard) violation of IIA. These observations suggest that the decision maker possesses at least k different preferences.

Of special interest are the sets of choice observations that are in K-violation of IIA, where K is the true number of underlying preferences. Below, I define *differentiation* to be the (limiting) fraction of choice observations that can be partitioned into disjoint K-violations of IIA.

Definition 5.2 (Differentiation)**.** *For each data set D, define $g(D)$ to be the largest number of disjoint subsets of choice observations in D that are in K-violation of IIA. The* differentiation parameter *for primitives (π, μ^*) is*

$$d(\pi, \mu^*) = \liminf_{n \to \infty} \mathbb{E}_{(\nu^*)^n} \left[\frac{1}{n} g(D_n) \right] \tag{4}$$

where $(\nu^*)^n$ is the product measure corresponding to n i.i.d. draws from ν, and $D_n \sim (\nu^*)^n$ is a random data set of size n.

To interpret the differentiation parameter $d(\pi, \mu^*)$, suppose that n choice observations are generated from the choice model described by primitives π and μ^*. Then, in expectation, there is a partitioning of the realized choice data such that at least

[14] Observations (x, A) and (x', B) are in violation of the Independence of Irrelevant Alternatives (IIA) axiom if $x, x' \in B$ and $x, x' \in A$.

$n \cdot d(\pi, \mu^*)$ partition elements are in K-violation of IIA. Recalling that choice observations in K-violation of IIA require K preferences for perfect rationalization, large values of $d(\pi, \mu^*)$ imply that use of fewer than K orderings cannot rationalize most of the data, and thus encourages recovery of K preferences.

There are some intrinsic restrictions on the size of the differentiation parameter. A basic bound is:

$$0 \leq d(\pi, \mu^*) \leq 1/K \tag{5}$$

for every π and μ^*. The lower bound was attained in Examples 3-5:

Observation 2. *Fix any π and μ^* obeying the restrictions described in Example 3, 4, or 5. Then, $d(\pi, \mu^*) = 0$.*

Generalizing from Example 3 in particular:

Definition 5.3. *Let*

$$\mathbb{C}(\mathcal{P}) = \left\{(x, A) \,:\, x \textit{ is } P\textit{-maximal in } A \textit{ for some } P \in \mathcal{P},\;\; A \in 2^X\right\}$$

be the set of unique choice implications of preferences in $\mathcal{P}$.

Observation 3. *If $\mathbb{C}(\mathcal{P})$ does not contain any K-violations of IIA, then $d(\pi, \mu^*) = 0$ for every sampling distribution π and every RUM μ^* supported on $\mathcal{P}$.*

The upper bound $d(\pi, \mu^*) = 1/K$ is attained if the data can eventually be (approximately) completely partitioned into disjoint K-violations of IIA. For example:[15]

Example 6. The DM's preferences are $\mathcal{P} = \{P_1, \ldots, P_K\}$. Define x_i^* to be the P_i-maximal element from X, and suppose that all x_i^* are unique. Given any choice set A, the DM samples uniformly over $\mathcal{P}$. Only choice sets containing $\{x_i^*\}_{i=1}^K$ are sampled with positive probability. Then, $d(\pi, \mu^*) = 1/K$.

Example 7. Consider $X = \{x_1, \ldots, x_N\}$ and define $\mathcal{P} = \{P_1, P_2\}$ where

$$\begin{gathered} x_1 \, P_1 \, x_2 \, P_1 \;\ldots\; P_1 \, x_N \\ x_N \, P_2 \, x_{N-1} \, P_2 \;\ldots\; P_2 \, x_1 \end{gathered}$$

Let each μ_A assign equal probability to both preferences, and let π be an arbitrary distribution over non-singleton choice sets. Then, $d(\pi, \mu^*) = 1/2$.

[15] In general, the value of $d(\pi, \mu^*)$ can vary significantly depending on the sampling distribution π. For example, as long as there is a single choice set on which all of the preferences disagree, then exclusive sampling of this choice set will result in $d(\pi, \mu^*)$ attaining the upper bound $1/K$. Similarly, as long as there is a single choice set on which all of the preferences agree, then exclusive sampling of this choice set will result in $d(\pi, \mu^*)$ attaining the lower bound 0.

The claim below says that so long as the differentiation parameter is strictly positive, then we can recover the number of orderings using any $\lambda > 1$ (which implements the Kalai et al. (2002) solution).

Claim 1. *Suppose $d(\pi, \mu^*) > 0$ and $\mu = \mu^*$. Then,*

$$\Pr(K^*_\lambda(D_n) = K) \to 1 \quad \textit{as} \quad n \to \infty$$

for every $\lambda > 1$.

The proof is clear and omitted. If the DM imperfectly maximizes, however, then $\lambda > 1$ will not generally be the best choice for recovery of K, and the size of the differentiation parameter $d(\pi, \mu^*)$ will be important. I turn to this case now.

5.2 Main Case: $\mu \neq \mu^*$

When there are no choice errors, all choice "inconsistencies" are (by assumption) representative of preference heterogeneity. Thus, the only obstacle to recovery is identification—will all of the K preferences be represented in the data?

This identification problem is also present when there are choice errors. But in addition, the possibility of choice error complicates our problem by introducing a new source of inconsistency. Choice errors may artificially inflate the inferred number of preferences (if we mistakenly interpret errors as preference), and they may also artificially reduce the inferred number of preferences (since an imperfectly maximized data set can in some cases be rationalized using fewer orderings than its perfectly maximized counterpart).

The proposed approach separates error from preference by looking for structure in the inconsistencies. A large set of choice inconsistencies that is "internally consistent"—i.e. rationalizable using the same preference—are interpreted as preference. Inconsistencies which are "internally *inconsistent*" are taken to represent error. This distinction requires a commitment to a notion of a large set, which is governed by choice of the parameter λ.

The main theorem provides values of λ under which the proposed approach recovers the true number of orderings as the number of observations grows large:

Theorem 1. *Define*

$$\overline{p} = d(\pi, \mu^*)(1 - p)^K. \tag{6}$$

Choose any $\tilde{p} \in [p, \overline{p}]$ and set $\lambda = 1/(\tilde{p}n)$. Then,

$$\Pr(K^*_\lambda(D_n) = K) \to 1 \qquad \textit{as } n \to \infty.$$

The condition in (6) defines a value $\overline{p}$ that is increasing in the differentiation parameter $d(\pi, \mu^*)$, decreasing in the probability of error p, and decreasing in the number of preferences K. Theorem 1 says that if each ordering ascribed to the DM is required to uniquely explain at least pn observations, but not more than $\overline{p}n$ observations, then the proposed approach will recover the number of underlying orderings given sufficiently many choice observations. Choosing $\lambda < 1/(\overline{p}n)$ may result in an underestimate of the number of preferences, and choosing $\lambda > 1/(pn)$ may result in an overestimate.

If either the differentiation parameter $d(\pi, \mu^*)$ is too small, or the probability of error p and number of preferences K too large, the condition in (6) will yield $\overline{p} < p$, in which case $K^*_\lambda(D)$ may not recover K for any value of λ. Specifically, it follows from the bounds on the differentiation parameter described in (5) that Condition (6) requires

$$\frac{1}{K} > \frac{p}{(1-p)^K}.$$

For example, if the DM has more than ten underlying preferences, then the probability of error cannot exceed $p = 0.05$, and if the DM has 5 underlying preferences, then the probability of error cannot exceed $p = 0.1$.[16]

The proof of Theorem 1 is deferred to Appendix B.2, but a sketch of the main ideas follows. The key idea is to identify every data set with an undirected hypergraph[17] (henceforth *graph*) in the following way: every node corresponds to a choice observation, and there is an edge between a set of nodes if and only if the corresponding observations are not consistent with maximization of any single ordering. The proof notes that a data set is k-rationalizable if and only if the corresponding graph is k-colorable.[18,19] Thus, the problem in (3) can be re-cast as finding the smallest number of colors k such that a large subset of nodes are k-colorable.

Let us consider a data set generated by repeated sampling from ν^* (which, recall, corresponds to choice without errors) instead of the actual distribution ν. Since by construction, this data set corresponds to *perfect* maximization of K orderings, the corresponding graph must admit a K-coloring. Moreover, since every set of observations in K-violation of IIA creates a complete K-partite subgraph, and at least one such set exists,[20] the corresponding graph cannot be colored by fewer than K colors. The challenge is to show that even when the graph is perturbed by choice

[16] I have not made efforts to optimize this bound, and it can be improved in future work.

[17] A *hypergraph* is a generalization of a graph in which edges may connect more than two vertices.

[18] A *k-coloring* of a graph is a partition of its vertex set V into k color classes such that no edge in E connects two nodes of the same color. A graph is *k-colorable* if it admits an k-coloring.

[19] This equivalence is shown by taking each color class to represent consistency with a distinct ordering.

[20] This is implied by $d(\pi, \mu^*) > 0$. If $d(\pi, \mu^*) = 0$, then the interval $[1/(\overline{p}n), 1/(pn)]$ is empty, and

error, with high probability it will remain the case that a large subset of the nodes can be colored by K colors, but no fewer.

To show that K colors are sufficient to color most nodes, I use Hoeffding's inequality to upper bound the number of imperfectly maximized choice observations by $1/\lambda$ (with high probability) as n gets large. To show that K colors are needed, I use the assumption in (6) and repeated applications of Hoeffding's inequality to lower bound the number of disjoint complete K-partite subgraphs by $1/\lambda$ (with high probability) as n gets large. This relies crucially on the differentiation parameter $d(\pi, \mu^*)$ being sufficiently large. Since each complete K-partite subgraph cannot be colored by fewer than K colors, the number of such subgraphs provides an approximate[21] lower bound on the number of nodes that are uniquely colored by each of the first K colors. Thus, each of the first K orderings uniquely explains at least $1/\lambda$ observations, and the marginal $(K+1)$-st ordering explains strictly fewer than $1/\lambda$ additional observations, so the proposed approach correctly returns K orderings.

5.3 Evaluating Assumptions

Since the number of orderings K, the probability of error p, and the differentiation parameter $d(\pi, \mu^*)$ are not known, the expression in (6) cannot be directly computed from the data. Nevertheless, we can infer properties of these unknowns. Theorem 1 provides an *interval* of values of λ that recover the same solution. If the inferred $K^*_\lambda(D)$ is the "correct" number of underlying preferences, then we can use the range of values of λ that induce this solution to bound $d(\pi, \mu^*)$ and p. To ease notation, d is used throughout this section in place of $d(\pi, \mu^*)$.

Formally, for each number of preferences k, define

$$\overline{\lambda}_k(D) = \max\{\lambda' \, : \, K^*_{\lambda'}(D) = k\} \qquad \underline{\lambda}_k(D) = \min\{\lambda' \, : \, K^*_{\lambda'}(D) = k\} \tag{7}$$

to be the largest and smallest values of λ' that return the solution k. An implication of Theorem 1 is that different choices of λ' in the interval $[1/(\overline{p}n), 1/(pn)]$ eventually yield the same solution.[22] Thus:

Corollary 1. *Choose any λ satisfying the condition in Theorem 1, and define $\underline{\lambda}(D_n)$*

the theorem is vacuously true.

[21] The proof considers the number of such subgraphs that are additionally *perfectly maximized*; this is a lower bound on the number of nodes that are uniquely colored by each of the first K colors.

[22] In fact, the proof of Theorem 1 in the appendix establishes a slightly stronger claim than stated: not only does each $K^*_\lambda(D_n)$ converge to K for $\lambda \in [1/(\overline{p}n), 1/(pn)]$, but convergence is uniform in this interval.

and $\overline{\lambda}(D_n)$ *as in (7). Then,*

$$\Pr\left(\left[\frac{1}{d(1-p)^K n}, \frac{1}{pn}\right] \subseteq \left[\underline{\lambda}_K(D_n), \overline{\lambda}_K(D_n)\right]\right) \to 1 \qquad \text{as } n \to \infty \tag{8}$$

The set of values in (8) is loosely an "inversion" of our main result: Theorem 1 says that if preferences are sufficiently differentiated and error is sufficiently small, then we can recover the number of preferences with an appropriate choice of λ. Corollary 1 asks, if λ is the appropriate choice, how differentiated could the preferences have been, and how high must the probability of error have been?

Given conjecture of any solution k, we can use (8) to back out implied properties about the (unobservable) primitives $d(\pi, \mu^*)$ and p. Each of $\overline{\lambda}_k := \overline{\lambda}_k(D_n)$, $\underline{\lambda}_k := \underline{\lambda}_k(D_n)$, and n can be computed directly from the data; thus, we can use

$$\left\{(d', p') \ : \ \left[\frac{1}{d'(1-p')^k n}, \frac{1}{pn}\right] \subseteq \left[\underline{\lambda}_k, \overline{\lambda}_k\right]\right\} \tag{9}$$

to (eventually) bound the possible values of d and p. Notice also that every d' and p' in the set (9) satisfy

$$\frac{1}{\underline{\lambda}_k(1 - 1/(\underline{\lambda}_k n))^k n} \leq d' \leq \frac{1}{\overline{\lambda}_k(1 - 1/(\overline{\lambda}_k n))^k n} \tag{10}$$

$$\frac{1}{\overline{\lambda}_k n} \leq p' \leq \frac{1}{\underline{\lambda}_k n} \tag{11}$$

Intuitively, when the differentiation parameter d is large, and the probability of error p is small, then we expect the interval $\overline{\lambda}_K - \underline{\lambda}_K$ (over which K is recovered) to be large. Hence, if $\overline{\lambda}_k - \underline{\lambda}_k$ is small, this implies either that k is not the underlying number of preferences, or that in fact d is small and p is large.

These bounds are applied in Section 7 to arbitrate between different solutions k.

5.4 Testing Rationality $(K = 1)$

The main part of this paper seeks to recover the "true" number of underlying preferences, but an important special case (with long precedence in the literature) concerns whether the choice data should be rationalized using a single preference.

Suppose that $K = 1$ in the proposed framework; then, the differentiation parameter $d(\pi, \mu^*)$ is trivially 1, since every choice observation constitutes a 1-violation of IIA. An immediate corollary of Theorem 1 is:

Corollary 2. *Set any* λ *satisfying* $0 \leq \lambda \leq 1/(pn)$*. Then,* $\Pr(K^*_\lambda(D_n) = 1) \to 1$ *as* $n \to \infty$.

It is interesting also to consider choice models that *don't* naturally correspond to maximization of a single preference, and see what the proposed approach recovers in those cases. Ideally, the proposed approach should behave differently for choice models that are far from this description, versus those that are close. As one comparison, I apply the proposed approach to a class of logit choice rules, where the probability of error is allowed to vary.

Example 8 (Single Preference with Logit Error)**.** Let $X = \{x_1, x_2, x_3\}$. The DM imperfectly maximizes the preference ordering $x_3 P x_2 P x_1$. Specifically, his probability of choosing alternative x from choice set A is given by

$$c(x|A) = \frac{e^{-\gamma u(x)}}{\sum_{x' \in A} e^{-\gamma u(x')}} \qquad \forall x \in A.$$

where $\gamma > 0$ and $u(x_k) = k$ assigns a higher payoff to alternatives that are higher ranked by P.

Within this class of choice rules, the probability of choosing the most preferred outcome is governed by the logit parameter γ. Lower choices of γ return higher probabilities of error, with the extreme case $\gamma = 0$ corresponding to uniform selection over the available alternatives. Intuitively, the choice rule is better described as "imperfect maximization of a single preference" when γ is large. Claims 2 and 3 below formalize two senses in which the proposed solution aligns with this intuition.

Claim 2 says that the size of the set of values of λ that recover a single ordering is monotonically increasing in γ, so that the more concentrated choice behaviors are, the more "slack" there is in recovery of $K = 1$.

Claim 2. *Define* $\overline{\lambda}(D) := \max\{\lambda : K^*_\lambda(D) = 1\}$ *and* $\underline{\lambda}(D) := \min\{\lambda : K^*_\lambda(D) = 1\}$. *Then, for every* n*, the expected size* $\overline{\lambda}(D_n) - \underline{\lambda}(D_n)$ *is increasing in* γ.[23]

This claim is illustrated below in Figure 3 for various choices of γ.

Thus, when γ is small, the proposed approach recovers $K = 1$ from a large set of values of λ. Claim 3 provides a complementary result: it says for every fixed value of λ, the recovered number of preferences is (weakly) decreasing with γ.

Claim 3. *For each choice of* λ *and quantity of data* n*, the expected value of* $K^*_\lambda(D_n)$ *(where the expectation is taken over data sets* D_n*) is weakly decreasing in* γ*.*

[23] Note that $\underline{\lambda}(D_n) = 0$ for every data set D_n, so a simpler statement of this result says that the expected size of $\overline{\lambda}(D_n)$ is increasing in γ.

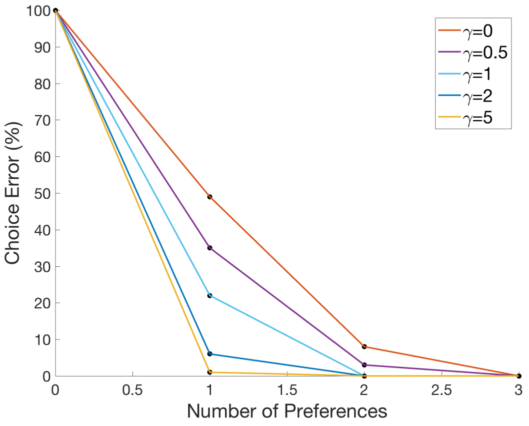


Figure 3: Error-Preference Tradeoff Graph for various choices of γ in Example 8 (assuming a uniform sampling rule over non-singleton choice sets, and simulating 100 choice observations). The set of choices of λ that recover one preference is larger for the curves corresponding to higher γ.

Thus, within the class of logit choice rules described above, the conditions for recovery of a single preference are least likely to be satisfied for $\gamma = 0$.

Another case that is conceptually distinct from "imperfect maximization of a single ordering" is when the DM exhibits small trembles around multiple preferences. As one such example, return to our setting from Section 2:

Example 9 (Two Preferences with Logit Error)**.** Depending on the choice set, the DM applies either of two logit choice rules (described in Section 2), each of which imperfectly maximizes a single preference. The logit parameter for both choice rules is $\gamma = 2$, so the probability of error is small for all choice sets. Given the idealized choice data set (corresponding to an infinite number of observations), one can show that the solution $K^*_\lambda(D_n) = 1$ is recovered for all values of $\lambda \leq 1/(0.46n)$. This interval is in fact smaller than the interval of values of λ that recover $K^*_\lambda(D_n) = 1$ for uniformly random choices ($\gamma = 0$ in Example 8), which is $\lambda \leq 1/(0.22n)$. Thus, imperfect maximization of two preferences is less suggestive of a single underlying preference (under the proposed approach), relative to if the choice data were completely random.

6 Extensions: Continuous Utility

So far, we have considered a decision maker whose preferences are orderings over a discrete set X. I now show that the main results extend to the case in which $(\underline{X}, \tau)$

is a topological space.

Formally, suppose that choice sets are compact subsets $\underline{A} \subseteq \underline{X}$, repeatedly sampled according to a distribution π.[24] The set $\mathscr{U} = \{u_\theta\}_{\theta \in \Theta}$ is a parametric family of continuous utility functions $u_\theta : \underline{X} \to \mathbb{R}$. Conditional on observation of choice set $\underline{A}$, the DM maximizes a utility function u_θ, where θ is sampled from a (Borel-measurable) distribution $\mu_{\underline{A}} \in \Delta(\Theta)$. Write $\underline{D}$ for a typical outcome of the choice data.

As before, I assume that the DM possesses an underlying "sparse" set of utility functions. Formally, each $\mu_{\underline{A}}$ can be rewritten as

$$\mu_{\underline{A}} = \mu_A G_A$$

where G_A is a Markov kernel on $(\Theta, \mathcal{B})$, and μ_A is supported on a finite set of utility functions $\mathcal{U} \subset \mathscr{U}$. The goal is to determine the number of utility functions $K := |\mathcal{U}|$.[25]

The proposed approach minimizes a weighted average of the number of inferred utility functions and the number of unexplained observations. For every set of utility functions $\mathcal{U}$, let

$$\underline{\varepsilon}(\underline{D}, \mathcal{U}) = \# \left\{ (x, \underline{A}) \in \underline{D} \, : \, x \neq \max_{x' \in \underline{A}} u(x') \text{ for any } u \in \mathcal{U} \right\}$$

be the number of choice observations in $\underline{D}$ that are not consistent with maximization of any utility function in $\mathcal{U}$. Then,

$$\underline{\varepsilon}_k(\underline{D}) = \min_{|\mathcal{U}|=k} \underline{\varepsilon}(\underline{D}, \mathcal{U})$$

is the minimal number of observations in $\underline{D}$ that are unexplained if we rationalize the DM's choices using k utility functions.

The solution below simultaneously minimizes the number of utility functions k and the implied choice error $\underline{\varepsilon}_k(\underline{D})$:

Definition 6.1. *For every $\lambda \in \mathbb{R}_+$, define*

$$\underline{K}^*_\lambda(\underline{D}) = \operatorname*{argmin}_{k \in \mathbb{N}} \left[k + \lambda \underline{\varepsilon}_k(\underline{D}) \right]. \tag{12}$$

As before, when there are multiple solutions, take $\underline{K}^*_\lambda(\underline{D})$ to mean the minimal value.

Corollary 3 below shows that this solution recovers the "correct" number of utility functions K under conditions that directly parallel the previous section. The statement below follows as a corollary to Theorem 1 (where the differentiation parameter $d(\pi, \mu^*)$ is defined as in Section 3.2):

[24] Some care is required in specifying the correct σ-algebra over choice sets; for example, one can take the Borel σ-algebra associated with the product topology of τ.

[25] Observe that as before, no parametric assumptions are made regarding the distribution of error; future work may include such assumptions to strengthen the recovery results.

Corollary 3. *Define*

$$\overline{p} = d(\pi, \mu^*)(1-p)^K. \tag{13}$$

Choose any $\tilde{p} \in [p, \overline{p}]$ *and set* $\lambda = 1/(\tilde{p}n)$*. Then,*

$$\Pr(K^*_\lambda(\underline{D}_n) = K) \to 1 \qquad \text{as } n \to \infty.$$

Thus, the proposed approach recovers the number of underlying utility functions as the number of observed choices gets large.

Why do the conditions of Theorem 1 extend to this more general setting? The key observation is that choice data generated in this way can be mapped into discrete choice data, where we reduce $\underline{X}$ to the finite set

$$X = \{x \in \underline{X} \,:\, (x, \underline{A}) \in \underline{D} \text{ for some } \underline{A} \subseteq \underline{X}\}.$$

This set consists of all choice alternatives that are observed to be chosen. For example, take $\underline{X} = \mathbb{R}$, and suppose we observe

$$\underline{D} := \{(3, [0,4]),\ (2, [1,4]),\ (8, [0,10])\}.$$

Then, labelling '3' as x_1, '2' as x_2, and '8' as x_3, we can redefine the set of choice alternatives as $X = \{x_1, x_2, x_3\}$, and the choice data as

$$D := \{(x_1, \{x_1, x_2\}),\ (x_2, \{x_1, x_2\}),\ (x_3, \{x_1, x_2, x_3\})\}.$$

This is a standard mapping in the literature, and yields a data set of the form introduced in Section 2.1.

A lemma in Appendix B.4 shows that the new problem posed in Definition 6.1 is equivalent to the original problem posed in Definition 4.1, in the sense that the solution to

$$\operatorname*{argmin}_{k \in \mathbb{Z}_+} \left[k + \lambda \underline{\varepsilon}_k(\underline{D})\right],$$

is the same as the solution to

$$\operatorname*{argmin}_{k \in \mathbb{Z}_+} \left[k + \lambda \varepsilon_k(D)\right].$$

It immediately follows that the conditions for recovery stated in the previous section are also the conditions needed in the present setting.

7 Application

This section describes an example application of the proposed approach, which builds on an analysis from Crawford and Pendakur (2012) (henceforth CP). CP study the consumption decisions of Danish households over six different kinds of milk, where the purchases are aggregated over a month. The relevant choice information is the quantity of each kind of milk purchased during this time (written as a quantity vector $\mathbf{q} \in \mathbb{R}^6$), and also the price index at which these purchases were made (written as a price vector $\mathbf{p} \in \mathbb{R}^6$). The main sample in CP consists of 500 households, so the choice data is $\{(\mathbf{p}_1, \mathbf{q}_1), \ldots, (\mathbf{p}_{500}, \mathbf{q}_{500})\}$.[26]

Let us map these choice observations into the present framework, using the relabelling described in Section 6. Index the observations by $i = 1, \ldots, 500$, and define $\mathbf{x}_i = (\mathbf{p}_i, \mathbf{q}_i)$. Take $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_{500}\}$ to be the set of choice alternatives. For each observation i, the set

$$A_i = \{\mathbf{x}_j \,:\, \mathbf{p}_i \cdot \mathbf{q}_i \geq \mathbf{p}_j \cdot \mathbf{q}_j\}$$

consists of every alternative $\mathbf{x}_j$ that is less costly than the selected alternative $\mathbf{x}_i$. These are the alternatives in X that could have been chosen when the alternative $\mathbf{x}_i$ was chosen.[27] The observed data from CP is now rewritten

$$D = \{(\mathbf{x}_1, A_1), \ldots, (\mathbf{x}_{500}, A_{500})\}.$$

This data set is equivalent to the original data in the sense described in the previous section.[28]

There are many ways to rationalize the choice data D. For example, following the proposal of Houtman and Maks (1985), we can find the single ordering that explains the largest fraction of the data, and interpret the remaining observations as choice error. CP find that no single preference explains more than two-thirds of the observations.

Alternatively—and this is the main approach taken in CP—we can seek the minimal set of preferences that explains every observation (thus following the proposal of

[26] Their data also includes a household indicator and covariates describing each household, but these are outside of the proposed framework.

[27] That is, if $\mathbf{p}_i \cdot \mathbf{q}_i \geq \mathbf{p}_j \cdot \mathbf{q}_j$, then $\mathbf{x}_j \in A_i$.

[28] Suppose there exists a set of k utility functions such that m observations in the original data are consistent with maximization of a utility function from this set; that is,

$$u(\mathbf{q}_i) > u(\mathbf{q})$$

for all quantity vectors $\mathbf{q}$ satisfying $\mathbf{p}_i^T \mathbf{q}_i \geq \mathbf{p}_i^T \mathbf{q}_i$. Then we can find a set of k preference orderings on X such that m observations in the relabelled data are consistent with maximization of a preference from this set, and vice versa.

Kalai et al. (2002)). CP find that no more than five orderings are needed to perfectly rationalize the data.

The present paper interprets the two solutions above as edge cases among a set of rationalizations of the data, each of which entails a different tradeoff between maximization of fit to the data and minimization of the number of orderings used. Figure 4 provides approximations from CP for the fraction of unexplained observations $\varepsilon_k(D)$ (for the purpose of illustration of the approach, I will treat these approximations as the exact values of $\varepsilon_k(D)$).[29] For example, with a single ordering, we must leave 179 (of the 500) observations unexplained; using two orderings, we must leave 79 observations unexplained; and with five orderings, we can explain all of the observations. The intuition that this paper formalizes is that there is significant variation in the degree of evidence for each of the five orderings, and this variation can be used to evaluate them.

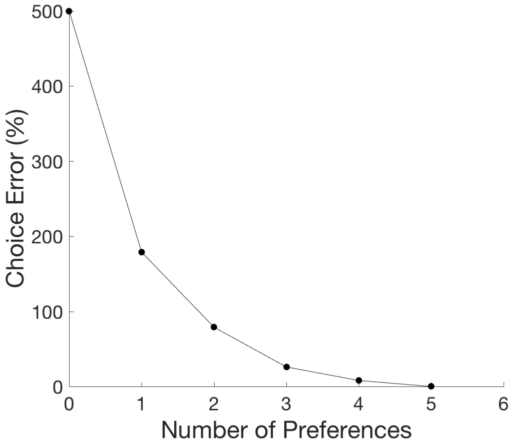


Figure 4: Error-Preference Tradeoff Graph for the Crawford and Pendakur (2012) data set.

There are five possible solutions to (3) given this data set, each of which holds for a range of choices of λ, as shown below:

[29] These values correspond to an algorithm that computes an upper bound on the number of types needed to explain a given number of observations, so the true values of $\varepsilon_k(D)$ are weakly smaller than those reported below. Crawford and Pendakur (2012) provide also lower bounds that show that the errors in this approximation are not large.

1	2	3	4	5
$\lambda \in [0, 0.01]$	$[0.01, 0.019]$	$[0.019, 0.056]$	$[0.056, 0.125]$	$[0.125, \infty)$

Table 1: Different numbers of preferences k (first row) are induced by different intervals of λ.

One rule-of-thumb is to set $\lambda = 1/(\tilde{p}n)$, where $\tilde{p}$ is a slight overestimate of the probability of error. For example, if the analyst believes that subjects have approximately a 5% probability of error, then the proposed approach recovers three preferences under $\lambda = 0.04$.

We can also take a more agnostic approach following the discussion in Section 5.3. For each potential solution k, the expression in (9) delivers bounds on the differentiation parameter and probability of error. Figure 5 below shows the sets of $(p, d(\pi, \mu^*))$ values that corresponds to different solutions k.[30]

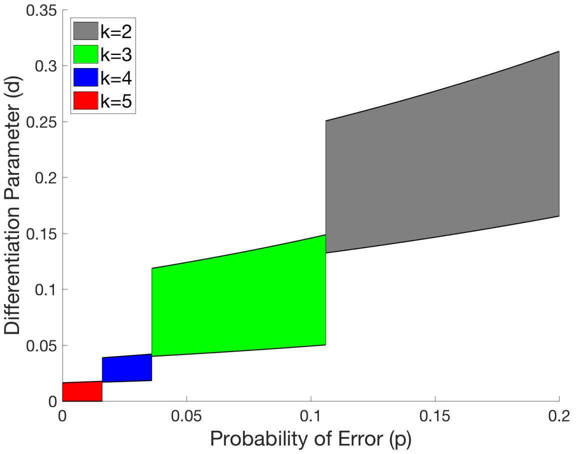


Figure 5: For each potential solution $k \in \{2, 3, 4, 5\}$, the figure shows the possible set of $(p, d(\pi, \mu^*))$ values (given the observed choice data).

We can additionally use the bounds in (10) and (11) to bound the probability of error p and the differentiation parameter d, given conjecture of the different solutions k. Table 2 reports these bounds.

[30] The set of values for $k = 1$ can be similarly computed but is not shown in Figure 5.

	$d(\pi, \mu^*)$		p	
k	Lower Bound	Upper Bound	Lower Bound	Upper Bound
1	0.25	1	0.2	1
2	0.13	0.31	0.106	0.2
3	0.04	0.15	0.036	0.106
4	0.017	0.04	0.016	0.036
5	0	0.017	0	0.016

Table 2: Bounds for the differentiation parameter $d(\pi, \mu^*)$ and probability of error p.

Notice that the regions of values corresponding to the solutions $k = 4$ and $k = 5$ are quite small. Specifically, for the candidate solution $k = 5$, the implied bounds are $p \in [0, 0.016]$ and $d(\pi, \mu^*) \in [0, 0.017]$, so that no more than 1.6% of choice observations are in error, and no more than 1.7% of choice observations provide evidence of five orderings. For the candidate solution $k = 4$, the implied bounds are $p \in [0.016, 0.036]$ and $d(\pi, \mu^*) \in [0.017, 0.04]$, so that no more than 3.6% of observations are in error, and no more than 4% of choice observations provide evidence of four orderings. To the extent that an analyst considers these restrictions too severe, the analysis suggests that the solutions $k = 4$ and $k = 5$ should not be favored. Specifically, the final two Kalai et al. (2002) preferences recovered in Crawford and Pendakur (2012) are better understood as capturing choice errors.

8 Can we Recover More?

Section 5 described conditions under which the problem in (3) recovers the number of underlying preferences with high probability. This section now asks whether it is possible to recover the preferences themselves.

8.1 Non-Identifiability of Sets of Preferences

Say that the set of preferences $\mathcal{P}$ is *identifiable* if there is at least one data set D such that $\mathcal{P}$ is the unique set of $|\mathcal{P}|$ preferences that perfectly explains every observation in D.

Definition 8.1. *The set of orderings* $\mathcal{P}$ *is* identifiable *if there exists some data set* D *such that* $\varepsilon(D, \mathcal{P}) = 0$*, and moreover* $\varepsilon(D, \mathcal{P}') > 0$ *for every* $\mathcal{P}' \neq \mathcal{P}$ *with* $|\mathcal{P}'| \leq |\mathcal{P}|$.

The following proposition says that most sets of orderings are not identifiable.

Proposition 1. *No set of orderings $\mathcal{P}$ with $|\mathcal{P}| \geq 3$ is identifiable. Suppose $\mathcal{P} = \{P_1, P_2\}$; then, $\mathcal{P}$ is identifiable if and only if the P_1-maximal alternative in X is the P_2-minimal alternative in X, and vice versa. Every singleton set $\mathcal{P} = \{P\}$ is (trivially) identifiable.*

The difficulty of recovery is not specific to the nonparametric nature of the exercise, but to basic issues concerning identifiability for multiple preferences. Consider any set that includes a pair of preferences P_1, P_2, where some alternative x_1 is ranked first under one ordering (say, P_1) and not ranked last under the other; for example,

$$x_1 P_1 x_2 P_1 x_3$$
$$x_2 P_2 x_1 P_2 x_3$$

We can construct then a new preference P_2' based on P_2, with the single difference that x_1 is ranked last:

$$x_1 P_1 x_2 P_1 x_3$$
$$x_2 P_2' x_3 P_2' x_1$$

Every choice that can be rationalized using a preference from $\{P_1, P_2\}$ can also be rationalized using a preference from the set $\{P_1, P_2'\}$; this is easily verified for the example above, and the proof of Proposition 1 shows that this follows more generally. Thus, $\{P_1, P_2'\}$ will be a solution whenever $\{P_1, P_2\}$ is.

The obstacle to recovery is loosely that sets of preferences differ in the "richness" of their choice implications. A set of preferences can (strictly) encompass all the choice implications of another set of the same size. Thus, any approach that penalizes only the *size* of a set of orderings, as both the proposed approach in (3) and the approach suggested in Kalai et al. (2002) do, will be biased towards elicitation of sets with richer choice implications.

I outline below two paths forward. Section 8.2 suggests a modification of the proposed approach which allows for recovery of the *choice implications* of the set of preferences. This is generally understood to be the real content of a set of preferences. Section 8.3 considers a richer kind of data set, which includes auxiliary information on the choice contexts active during different observations. I show that with this additional information, we can (under conditions that I characterize) recover the set of preferences.

8.2 Recovery of Choice Implications

One approach is to change the "complexity" penalty from *number of preferences* to *number of choice implications*.

Formally, define the function $l : \mathscr{P} \to \mathbb{Z}_+$ such that $l(\mathcal{P})$ counts the number of unique choice observations (x, A) that are consistent with $\mathcal{P}$; that is,

$$l(\mathcal{P}) = \#\{(x, A) : x \text{ is } P\text{-maximal in } A \text{ for some } P \in \mathcal{P}\}.$$

Some properties of l can be found in Appendix 8.2. The following definition modifies the proposed approach in Definition 4.1 by replacing the metric $|\mathcal{P}|$ with the metric $l(\mathcal{P})$. This solution minimizes a weighted average of the number of choice implications and the number of unexplained choice observations.

Definition 8.2. *For every $\lambda \in \mathbb{R}_+$, define*

$$\mathcal{P}^*_\lambda(D) = \operatorname*{argmin}_{\mathcal{P} \in \mathscr{P}} \left[l(\mathcal{P}) + \lambda \varepsilon(D, \mathcal{P}) \right]. \tag{14}$$

In this case, $1/\lambda$ can be understood as the cost of each new choice implication, so that a choice implication is attributed to the DM only if it appears at least $1/\lambda$ times in the data. As $\lambda \to 0$, the cost of errors becomes increasingly small relative to the cost of new choice implications, so that the analyst prefers to attribute to the DM as few choice implications as possible. As $\lambda \to \infty$, the cost of choice errors becomes increasingly large relative to the cost of new choice implications, so that the analyst prefers to ascribe to the DM as many choice implications as is necessary to perfectly rationalize the data.

The proposition below says that for certain choices of λ, we can recover the choice implications of $\mathcal{P}$ (denoted $\mathbb{C}(\mathcal{P})$ as in Definition 5.3) with probability arbitrarily close to 1 as the quantity of data increases.

Proposition 2. *Define α to be the smallest nonzero frequency with which any choice observation occurs under ν^*.*[31] *Choose any $\tilde{p} \in [p, \alpha(1-p)]$ and set $\lambda = 1/(\tilde{p}n)$. Then,*

$$\Pr(\mathbb{C}(\mathcal{P}^*_\lambda(D_n)) = \mathbb{C}(\mathcal{P})) \to 1 \quad \text{as } n \to \infty.$$

Above, the cost $1/\lambda$ of recovering an additional choice observation is chosen to satisfy $1/\lambda = \tilde{p}n < \alpha(1-p)n$. Thus, every choice implication $(x, A) \in \mathbb{C}(\mathcal{P})$ is observed sufficiently many times to not be mistaken as error. To see that no choice implication $(x, A) \notin \mathbb{C}(\mathcal{P})$ will be incorrectly recovered, observe that the number of instances of $(x, A) \notin \mathbb{C}(\mathcal{P})$ in the data is upper bounded by the number of choice errors. With high probability, the number of choice errors concentrates below $pn < 1/\lambda$, from which it follows that $\mathbb{C}(\mathcal{P}^*_\lambda)$ will include as few choice implications outside of $\mathbb{C}(\mathcal{P})$ as possible.

[31] That is, let $\alpha = \min_{(x,A) : \nu^*(x,A) > 0} \nu^*(x, A)$.

8.3 Auxiliary Data on Contexts

As an alternative approach to recovery of preference, we may turn to richer data sets. Specifically, suppose that there is a set of *observable* choice contexts $\mathscr{C} = \{1, \ldots, M\}$. Each context is associated with a preference, and multiple contexts may be associated with the same preference. (For example, suppose that choice data is aggregated over various kinds of financial decisions, and the contexts correspond to different financial domains. Individuals may have the same risk preference over all types of insurance decisions, but a different risk preference over 401(k) savings.)

Formally, there is an unknown map

$$m : \mathscr{C} \to \mathscr{P},$$

which assigns each choice context a preference. The key assumption is that the image of m is a sparse set of preferences $\mathcal{P} \subset \mathscr{P}$.

For each choice set A, let the empirical distribution over contexts be given by $\phi_A \in \Delta(\mathscr{C})$. The RUM μ_A^* introduced in Section 3.3 can be given the following foundation:

$$\mu_A^*(P) = \phi_A(m^{-1}(P)).$$

That is, the probability with which P is sampled (in the absence of choice error) is the probability that a context emerges which cues preference P. As before, there is a sampling distribution $\pi \in \Delta(2^X)$ over choice sets. When there is no choice error, we can write

$$\nu^*((x, A), C) = \begin{cases} \pi(A)\phi_A(C) & \text{if } x \text{ is } m(C)\text{-optimal in } A \\ 0 & \text{otherwise} \end{cases}$$

for the probability that choice observation (x, A) is observed in context C. Notice that in this case, the outcome x is deterministic conditional on the choice context C and choice set A.

In the main model, which allows for choice errors, observations are instead sampled from

$$\nu((x, A), C) = \pi(A)\phi_A(C)q_C(x|A)$$

where each $q_C(\cdot|A) \in \Delta(A)$ is a distribution over choice alternatives, associated with the context C. Assume that each $q_C(\cdot|A)$ assigns probability at least $1 - p$ to the $m(C)$-optimal alternative in A.

Example 10. There are four kinds of subjects: males over the age of 65, males under the age of 65, females over the age of 65, and females under the age of 65. These categories are indexed (in that order) using $\mathscr{C} = \{1, 2, 3, 4\}$. In reality, only age determines preference, so that $m(1) = m(3) = P$ and $m(2) = m(4) = P'$. The analyst does not know this. He observes tuples $((x, A), C)$ indicating that alternative x was chosen from choice set A by a subject of type i. The goal is to back out from the data that there are only two active preferences, and to determine which these are.

For this framework, we can modify the approach in Section 4 to search for assignments of preferences to contexts. Define

$$\varepsilon(D, m) = \#\{((x, A), C) \in D \,:\, x \text{ is not } m(C)\text{-maximal in } A\}$$

to be the number of implied choice errors when m is the mapping from contexts to preferences, and define

$$m^*_\lambda := \min_{m:\mathscr{C}\to\mathscr{P}} [|m(\mathscr{C})| + k\varepsilon(D, m)] \tag{15}$$

where $|m(\mathscr{C})|$ is the number of (unique) preferences assigned under m, and the dependence of m^*_λ on the data D is suppressed for notational convenience. Thus, m^*_λ is the assignment of preferences to contexts that minimizes the number of distinct preferences, and also the associated number of choice errors.

In the proposition below, define α to be the smallest nonzero frequency with which any observation $((x, A), C)$ occurs under ν^*.[32]

Proposition 3. *Choose any $\lambda = 1/(\tilde{p}n)$ where $\tilde{p} \in [p, \alpha(1-p)/2]$. Then,*

$$\Pr(m^*_\lambda = m) \to 1 \qquad \textit{as } n \to \infty.$$

Thus, recovery of the exact set of preferences is possible if there is auxiliary information about choice contexts, and sufficient observation of the choice implications for each choice context. See Appendix A for an application of this approach.

9 Related Literature

9.1 Nonparametric Preference Recovery

This paper builds on a literature regarding nonparametric identification of multiple preferences from choice data. Most directly, it extends Kalai et al. (2002), which defines a set of orderings $\mathcal{P}$ to be a *rationalization by multiple rationales* if for every observation (x, A), the choice alternative x is P-maximal in A for some ordering $P \in \mathcal{P}$. Kalai et al. (2002) search for the smallest L such that some set $\mathcal{P}$ with

[32] That is, $\alpha = \min_{C\in\mathscr{C},\,(x,A)\in\mathbb{C}(\{m(C)\})} \nu^*((x, A), i)$.

$|\mathcal{P}| = L$ is a rationalization by multiple rationales. Any such set of preferences $\mathcal{P}$ is a perfect rationalization of the data, but it may not correspond to a "best" rationalization of the data as defined in (3). In particular, I suggest that the analyst may prefer an imperfect rationalization of the data using some $K < L$ orderings. The key conceptual difference is that Kalai et al. (2002) is agnostic towards the degree of evidence for orderings, whereas the approach in this paper insists on sufficient evidence for each ordering in order to separate error from preference variation.

Ambrus and Rozen (2013) study multiple preference models in which choice is determined through maximization of a choice-set independent aggregation rule over preferences. They find that without prior restriction on the number of selves involved in a decision, many multiple preference ("multi-self") models have no testable implications. Although the class of models considered in their paper is different from the class studied in the present paper,[33] their lesson that restricting the number of preferences is necessary for recovery holds here as well (relating especially to the results in Section 8.1), and motivates in part the suggested criterion in (3).

Other nonparametric approaches for preference identification include Houtman and Maks (1985) and Varian (1982). These approaches differ from the present paper, and from Kalai et al. (2002), in finding a *single* best-fit ordering. A separate and sizable literature studies related questions under various parametric assumptions—see, for example, Quandt (1956), McFadden and Richter (1970), and Train (1986)). Finally, Crawford and Pendakur (2012) and Dean and Martin (2010) apply the approaches of Kalai et al. (2002) and Houtman and Maks (1985) towards recovery of preferences from real choice data.[34]

9.2 Testing Rationality

When choice data is *inconsistent*—meaning that it is incompatible with perfect rationalization by a single ordering—how should we measure the inconsistency of the observed choices? Solutions have been proposed by Afriat (1967), Varian (1982), Echenique et al. (2011), Houtman and Maks (1985), Gross (1989), Famulari (1995), Apesteguia and Ballester (2012), and Dean and Martin (2016) among others. See Apesteguia and Ballester (2012) for a summary and comparison of these approaches.

In view of this literature, one goal of the present paper is to distinguish between choice data that is inconsistent because of choice error, and choice data that is inconsistent because of multiplicity in preference. These two sources for error are confounded in many of the measures above.

[33] In the present paper, the aggregation rule varies across choice problems.

[34] See Deb (2009) and Dean and Martin (2010) for computationally efficient approaches for approximating the Kalai et al. (2002) solution.

The proposed approach offers a new perspective on this question. If the choice data D is generated via approximately perfect rationalization of multiple preferences, the recovered number of preferences $K^*_\lambda(D)$ will be large for most choices of λ (and the number of unexplained choice observations $\varepsilon_{K^*_\lambda}(D)$ will be small); see Section 5.4 for details.

Finally, Halevy et al. (2015) proposes a novel decomposition of error into two sources: *choice inconsistency* and *preference misspecification*. The present paper is related to these ideas at a high level; in particular, the proposed approach also provides a decomposition of error. The sources of error considered in the two papers are different, however. For example, Halevy et al. (2015) considers a *single* (continuous) utility function and measures misspecification that arises from parametric restrictions. The present paper takes a nonparametric approach and considers multiple preferences.

10 Conclusion

Inconsistencies in choice data may emerge *both* from (unintentional) choice error and also from (intentional) maximization of different preferences. Classic approaches such as the Houtman and Maks (1985) and Kalai et al. (2002) solution focus on either of these sources of error, but—for reasons of welfare evaluation, and out-of-sample prediction—we may prefer interpretations of the data that accommodate both.

This paper proposes identification of underlying "structural" preferences that are maximized in the data. The proposed approach looks for the multiple preference rationalization of the data that simultaneously minimizes the number of preferences and also the number of unexplained observations. Different choices of tradeoffs between these objectives yield different solutions, and the main results relate the optimal choice of tradeoff to primitives of an underlying choice model.

Some of the techniques proposed in this paper may be useful towards other goals as well. For example, suppose an analyst wants to know whether choice inconsistencies in the data can be explained via preference indifferences. Specifically, the analyst hypothesizes that the DM has a single partial ordering, and chooses uniformly from the those alternatives (potentially multiple) that he most prefers. Such a model is outside the scope of the current paper. Nevertheless, we can test the analyst's hypothesis by examining the error-preference tradeoff graph, as we did in this paper. If all choice inconsistencies can be attributed to indifferences, then the error-preference tradeoff graph should take a particular shape—similar to Figure 3, and different from Figure 1. This discussion suggests that the basic analysis of the choice data used in this paper may be amenable for evaluating other models of choice as well.

Appendix

A Out-of-Sample Prediction Accuracy

One reason the number of preferences attributed to a data set matters is that it affects prediction of choice behaviors. As mentioned in the main text, the Houtman and Maks (1985) approach can underfit the data, missing important structural preferences, while the Kalai et al. (2002) approach can overfit the data, interpreting errors as preference. These problems can result in substantial reductions in prediction accuracy, and I use two simple examples below to illustrate this. Since out-of-sample predictions are of interest, these examples follow the extension described in Section 15, where choice contexts are observed, and the goal is to correctly assign contexts to preferences.

In both examples below, a *training set* (of 20 choice observations) is generated from a fixed choice rule, and the respective approaches are applied to this data. I then generate a new *test set* (of 100 choice observations) on which to evaluate the estimated models.[35] Prediction accuracy is measured as the fraction of test observations for which the chosen alternative is correctly predicted. This procedure is repeated ten times (with new training and test data), and I report an average of the out-of-sample prediction accuracies.

In Example 11, a DM maximizes a different preference in each of two choice contexts, and I show that application of the proposed approach improves on Houtman and Maks (1985). In Example 12, the DM imperfectly maximizes a single preference in all choice contexts, and I show that application of the proposed approach improves upon Kalai et al. (2002). In both cases, the magnitude of the gain in prediction accuracy from use of the proposed approach is substantial.

To keep these examples as simple as possible, both examples have the feature that *one of* Houtman and Maks (1985) or Kalai et al. (2002) would achieve the performance of the proposed approach. More complex examples can be constructed in which both perform poorly out-of-sample (in the spirit of Section 2).

Example 11. There are four choice alternatives $X = \{x_1, x_2, x_3, x_4\}$. Choice sets $A \subseteq X$ are generated uniformly at random (excluding singleton choice sets). There are two different choice contexts, indexed $\mathcal{C} = \{1, 2\}$, both observable.

The DM maximizes a different choice rule in each context. In the first, his prob-

[35] That is, new observations (x, A) are generated, and the estimated models are applied to predict the chosen alternative x given the choice set A.

ability of choosing alternative x from choice set A is given by

$$c(x|A) = \frac{e^{-\gamma u(x)}}{\sum_{x' \in A} e^{-\gamma u(x')}}$$

where $\gamma = 10$ and $u(x_k) = k$ (so that higher indexed alternatives are more preferred). In the second, his probability of choosing alternative x from choice set A is given by the same expression, also with $\gamma = 10$ but with a different utility function $u(x_k) = 5 - k$ (so that higher indexed alternatives are less preferred).

Consider two approaches for making out-of-sample predictions: First, following Houtman and Maks (1985), find the best *single* preference that maximizes the number of rationalized choice observations. Second, following the proposed approach in this paper, identify the solution that solves (15)—that is, the assignment that minimizes a weighted average of the *number of unique preferences* assigned to contexts, and also the *number of unexplained observations*. For robustness, I show below the prediction accuracies for a couple of different choices for the tradeoff parameter λ.

Out-of-sample performance accuracies are reported and compared in the table below. The proposed approach improves upon Houtman and Maks (1985) by up to 48%.

	Prediction accuracy
Kalai et al. (2002)	47%
	(5.8)
Proposed approach using $\lambda = 1/(0.1n)$	95%
	(5.6)
Proposed approach using $\lambda = 1/(0.2n)$	93%
	(6.2)
Proposed approach using $\lambda = 1/(0.3n)$	78%
	(2.2)

Example 12. Again, there are four choice alternatives $X = \{x_1, x_2, x_3, x_4\}$, and choice sets $A \subseteq X$ are generated uniformly at random (excluding singleton choice sets). There are five different choice contexts indexed $\mathcal{C} = \{1, 2, \ldots, 5\}$, all observable.

In all contexts, the DM uses the same choice rule in which the probability of choosing alternative x from choice set A is given by

$$c(x|A) = \frac{e^{-\gamma u(x)}}{\sum_{x' \in A} e^{-\gamma u(x')}}$$

where $\gamma = 1.5$ and $u(x_k) = k$ (so that higher indexed alternatives are more preferred).

I consider two approaches for making out-of-sample predictions. First, following Kalai et al. (2002), I find an assignment of preferences to contexts that minimizes the number of unexplained choice observations.[36,37] Second, following the proposed approach in this paper, I identify the solution that solves (15)—that is, the assignment that minimizes a weighted average of the *number of unique preferences* assigned to contexts, and also the *number of unexplained observations*. For robustness, orediction accuracies are shown for a couple of different choices for the tradeoff parameter λ.

The table below reports and compares out-of-sample performance accuracies for these two approaches. Again, the proposed approach leads to significant improvements in predictive accuracy (up to 20%).

	Prediction accuracy
Kalai et al. (2002)	64% (9.8)
Proposed approach using $\lambda = 1/(0.1n)$	77% (10.7)
Proposed approach using $\lambda = 1/(0.2n)$	83% (7.1)
Proposed approach using $\lambda = 1/(0.3n)$	84% (6.0)

B Proofs from Main Text

B.1 Preliminaries

Following, I collect definitions and results that are used in the proofs of Theorem 1 and Proposition 1.

Let $G : D \mapsto G_D$ be a map that identifies every data set D with a (hyper-) graph $G_D = (V_D, E_D)$, where $V_D = \{1, \ldots, n\}$ indexes choice observations, and E_D consists of every set $T \subseteq V_D$ with the property that: (1) the observations $\{(x_i, A_i)\}_{i \in T}$ are not 1-rationalizable, and (2) every proper subset of $\{(x_i, A_i)\}_{i \in T}$ is 1-rationalizable. These concepts are related to our problem as follows.

[36] Note that unlike in the usual setting in which Kalai et al. (2002) is applied, here it may not be possible to achieve a perfect rationalization, because we require preferences to be constant within choice contexts.

[37] When there are multiple assignments that achieve the minimal error, we select from these uniformly at random.

Observation 4. *D is k-rationalizable $\iff$ G_D is k-colorable.*

Take each color class to represent consistency with a distinct ordering, and the equivalence follows directly.

This observation further implies that ε_k is the minimum number of vertices that must be removed from G_D in order for the graph to be k-colorable. From here on, I will refer to the vertices of G_D and the observations they represent interchangeably.

B.2 Proof of Theorem 1

First, observe that K is the (unique) solution to the problem in (3) if and only if

$$K + \lambda \varepsilon_K(D) < k + \lambda \varepsilon_k(D) \qquad \text{for every integer } k \neq K.$$

We can break this condition up into two parts. For $k > K$, the inequality

$$\varepsilon_K(D) - \varepsilon_k(D) < (k - K)/\lambda$$

requires that preferences beyond the best set of K each rationalize *no more than* $1/\lambda$ choice observations (beyond what would already be rationalized using the best K). For $k < K$, the inequality

$$\varepsilon_k(D) - \varepsilon_K(D) > (k - K)/\lambda$$

requires that each of the best K preferences uniquely rationalizes *at least* $1/\lambda$ choice observations (beyond what would be rationalized using the best k). Lemmas 1 and 2 bound the probability of each of these events.

Lemma 1. *There exists a constant $c_1 > 0$ (uniform across n) such that*

$$\nu^n \left(\left\{ D_n : \varepsilon_K(D_n) - \varepsilon_k(D_n) < \frac{k - K}{\lambda} \quad \forall\, k > K \right\} \right) \geq 1 - e^{-c_1 n} \quad \forall\, n$$

Proof. Suppose some $k > K$ is selected as the solution to (3), so that there are $k - K$ preferences (beyond the best set of K preferences) that together rationalize $(k-K)/\lambda$ additional choice observations. Clearly, a necessary condition for such a k to exist is that the best K preferences leave at least $1/\lambda$ observations unexplained; that is, $\varepsilon_K(D) \geq 1/\lambda$. Equivalently, a *sufficient* condition for

$$\varepsilon_K(D) - \varepsilon_k(D) > (k - K)/\lambda \quad \forall k > K \tag{16}$$

is that the best K preferences leave *fewer* than $1/\lambda$ observations unexplained; that is, $\varepsilon_K(D) < 1/\lambda$. The probability of this event lower bounds the probability of (16).

Recall that $\mathcal{P}$ is the DM's "true" set of K preferences, and define $T(D)$ to be the number of realized choice observations in data set D that cannot be rationalized by any preference in $\mathcal{P}$. Removing these $T(D)$ "bad" observations from D results in a K-rationalizable data set; thus, $\varepsilon_K(D)$ cannot exceed $T(D)$. By assumption, the probability of error in any given observation is no more than p. Thus, the random variable $Y_n \sim \text{Bin}(n, p)$ first-order stochastically dominates $T(D_n)$.

These observations imply that

$$\begin{aligned}\nu^n(\{D_n : \varepsilon_K(D) - \varepsilon_k(D) > (k-K)/\lambda \; \forall\, k > K\}) &\geq \nu^n(\{D_n \,:\, \varepsilon_K(D) < 1/\lambda\}) \\ &\geq \nu^n(T(D_n) < 1/\lambda) \\ &\geq \Pr(Y_n < 1/\lambda)\end{aligned}$$

Since λ is chosen to satisfy $1/\lambda < pn$, it follows from Hoeffding's Inequality that

$$\Pr(Y_n \leq 1/\lambda) = \Pr(Y_n - pn \leq 1/\lambda - pn) \leq \exp\left(-\frac{2(1/\lambda - pn)^2}{n}\right) = e^{-2(\tilde{p}-p)^2 n}$$

Thus

$$\nu^n\left(\left\{D_n \,: \varepsilon_K(D_n) - \varepsilon_k(D_n) < \frac{k-K}{\lambda} \quad \forall\; k > K\right\}\right) \geq 1 - e^{-c_1 n}$$

with $c_1 := 2\left(\tilde{p} - p\right)^2 > 0$, and we are done. □

Lemma 2. *$\nu^n\left(\left\{D_n \,: \varepsilon_k(D_n) - \varepsilon_K(D_n) > \frac{K-k}{\lambda} \quad \forall\; k < K\right\}\right) \to 1$ as $n \to \infty$.*

Proof. The basic approach is to study the induced graph G_D, and lower bound the number of disjoint complete K-partite subgraphs.[38] With some imprecision, a sufficient condition for

$$\varepsilon_k(D_n) - \varepsilon_K(D_n) > \frac{K-k}{\lambda} \quad \forall\; k < K \tag{17}$$

is that the number of disjoint complete K-partite subgraphs in the induced graph G_D is at least $1/\lambda$. Roughly, the argument is as follows: Consider use of any $k < K$ colors to color the graph G_D. Clearly, only k nodes in each complete K-partite graph can be colored. Moreover, each of the additional $K - k$ colors permits the coloring of at least one more node per complete K-partite subgraph, and hence $(K-k)/\lambda$ more nodes in the graph in total. Thus, going from k colors to K colors reduces the number of uncolored nodes by at least $(K-k)/\lambda$. Using Observation 4, this is equivalent to the statement that going from k preferences to K preferences leaves at least $(K-k)/\lambda$ fewer observations unexplained, and this implies the desired (17).

[38] Recall that a *complete K-partite* subgraph is a graph whose nodes can be partitioned into K sets, where nodes in the same set are not connected by an edge.

Instead of directly working with the distribution over graphs induced by ν, it is simpler to first study the related "perfect maximization" graph-generating process, where choice observations are sampled from ν^*. Call the typical such graph G^*_D, and let $T(G^*_D)$ denote the number of disjoint complete K-partite subgraphs. By definition of the differentiation parameter,

$$\Pr(T(G^*_{D_n}) > d(\pi, \mu^*)) \to 1 \qquad \text{as } n \to \infty.$$

Moreover, since the measure $(\nu^*)^n$ assigns probability 1 to data sets that are K-rationalizable, any graph induced by a data set generated under sampling from $(\nu^*)^n$ is K-colorable. Thus, each additional color permits the coloring of at least $d(\pi, \mu^*)$ additional colors. By assumption that $1/\lambda < \overline{p}n$ and by the definition of $\overline{p}$, the differentiation parameter satisfies $d(\pi, \mu^*) > \frac{1}{\lambda(1-p)^K}$. So

$$\Pr\left(T(G^*_{D_n}) > \frac{1}{\lambda(1-p)^K}\right) \to 1 \qquad \text{as } n \to \infty. \tag{18}$$

Suppose now that each node in $G^*_{D_n}$ is removed from all of its edges with probability p, and call the resulting graph $\tilde{G}^*_{D_n}$. The number of complete K-partite subgraphs first-order stochastically dominates the number of complete K-partite subgraphs in G_D, generated under sampling from ν^*. Because each complete K-partite subgraph of $G_{D'}$ is preserved with independent probability $(1-p)^K$, the random variable $T(\tilde{G}^*_{D_n})$ has distribution $\text{Bin}(T(G^*_{D_n}), (1-p)^K)$, and its expectation is $\mathbb{E}(T(\tilde{G}^*_{D_n})) = \mathbb{E}(T(G^*_{D_n}) \cdot (1-p)^K)$. Thus, (18) implies

$$\Pr\left(T\left(\tilde{G}^*_{D_n}\right) > \frac{1}{\lambda}\right) \to 1 \qquad \text{as } n \to \infty. \tag{19}$$

Finally, since $T(G_{D_n})$ first-order stochastically dominates $T(\tilde{G}^*_{D_n})$, (19) further implies that

$$\Pr\left(T\left(G_{D_n}\right) > \frac{1}{\lambda}\right) \to 1 \qquad \text{as } n \to \infty$$

and we are done. □

B.3 Proof of Corollary 1

Let $\mathcal{Z}$ be the set of possible choice observations, with typical element (x, A). Write G_n for the (random) graph induced by n i.i.d. samples from the distribution $\nu \in \Delta(\mathcal{Z})$. Fix $k \in \mathbb{Z}_+$, and let c^k_n be the (random) maximum cardinality of a k-colorable subgraph of G_n.

Claim 4. *With probability 1, c^k_n/n converges.*

Proof. Let G_0 be the graph induced by the data set $\mathcal{Z}$ (which includes each distinct observation exactly once). For each k-colorable subgraph $\mathcal{H}$ of G_0, define $\mathcal{H}_n$ to be the subgraph of G_n that includes all repetitions of $z \in \mathcal{H}$, and assigns to each such observation its color in $\mathcal{H}$. By the strong law of large numbers, $|\mathcal{H}_n|/n$ converges, for each $\mathcal{H}$, to a number $c(\mathcal{H})$ that depends only on $\mathcal{H}$ and ν. □

Observe that $\varepsilon_k(D_n)/n = 1 - c_n^k/n$. Thus, Claim 4 implies that each $k + \lambda\varepsilon_k(D_n)/n$ converges, implying that $\underline{\lambda}_k(D_n)$ and $\overline{\lambda}_k(D_n)$ converge for every k.

B.4 Proof of Corollary 3

Fix any data set $\underline{D} = \{(x_i, \underline{A}_i)\}_{i=1}^n$. Define $X = \{x_i\}_{i=1}^n$ to include all chosen alternatives, and define each $A_i = X \cap \underline{A}_i$. Set $D = \{(x_i, A_i)\}_{i=1}^n$.

Lemma 3. *For every $\lambda \in \mathbb{R}_+$,*

$$\min_{k \in \mathbb{Z}_+} [k + \lambda\varepsilon_k(D)] = \min_{k \in \mathbb{Z}_+} [k + \lambda\underline{\varepsilon}_k(\underline{D})]$$

Proof. I will show that $\varepsilon(D, \mathcal{P}) = \delta$ for some set of preference orderings $\mathcal{P} = \{P_j\}_{j=1}^k$ if and only if $\underline{\varepsilon}(\underline{D}, \mathcal{U}) = \delta$ for some set of utility functions $\mathcal{U} = \{u_j\}_{j=1}^k$.

Fix any set $\mathcal{P} = \{P_j\}_{j=1}^k$ of k orderings defined on X, and take $\delta := \varepsilon(D, \mathcal{P})$. Every ordering P_j admits representation via a utility function $u_j : X \to \mathbb{R}$.[39] Moreover, we can extend each u_j to a continuous function $\underline{u}_j$ on $\underline{X}$ satisfying $\text{argmax}_{x \in \underline{X}}\, \underline{u}_j(x) = \text{argmax}_{x \in X}\, u_j(x)$. Then, $x_j = \text{argmax}_{x \in \underline{A}}\, \underline{u}_j(x)$ if and only if x_j is P_j-maximal in A. Set $\mathcal{U} = \{\underline{u}_j\}$; then,

$$\underline{\varepsilon}(\underline{D}, \mathcal{U}) = \# \left\{ (x, \underline{A}) \in \underline{D} \, : \, x \neq \underset{x' \in \underline{A}}{\text{argmax}}\, \underline{u}_j(x') \text{ for all } j = 1, \ldots, k \right\} = \delta.$$

Thus, choice error δ is attainable using a set of k utility functions. It follows that

$$\min_{k \in \mathbb{Z}_+} [k + \lambda\varepsilon_k(D)] \geq \min_{k \in \mathbb{Z}_+} [k + \lambda\underline{\varepsilon}_k(\underline{D})]. \tag{20}$$

In the other direction, fix a set $\mathcal{U} = \{u_j\}_{j=1}^k$ of k continuous functions defined on $\underline{X}$, and take $\delta = \underline{\varepsilon}(\underline{D}, \mathcal{U})$. For every utility function u_j, let P_j be the ordering on X that satisfies

$$x P_j x' \iff u_j(x) > u_j(x').$$

[39] That is, there exists a utility function u_j such that for every $x, x' \in X$, $x P_j x'$ if and only if $u_j(x) > u_j(x')$.

Then, $x_j = \text{argmax}_{x \in \underline{A}}\, u_j(x)$ if and only if x_j is P_j-maximal in A. Setting $\mathcal{P} = \{P_j\}_{j=1}^k$, we have that

$$\varepsilon(D, \mathcal{P}) = \#\{(x, A) \in D : x \text{ is not } P_j\text{-maximal in } A \text{ for any } j = 1, \dots, k\} = \delta.$$

Thus, also

$$\min_{k \in \mathbb{Z}_+} [k + \lambda \varepsilon_k(D)] \leq \min_{k \in \mathbb{Z}_+} [k + \lambda \underline{\varepsilon}_k(\underline{D})]$$

as desired. □

It follows from this lemma that $\underline{K}^*_\lambda(\underline{D}) = K^*_\lambda(D)$ for every choice of λ, so the problem posed in Section 6.2 can be mapped directly into a corresponding problem involving choice over a discrete set. Apply Theorem 1, and the desired result follows.

B.5 Proof of Claim 2

It can be shown that for every n, the single preference that minimizes the expected value of $\varepsilon_1(D_n)$ is $x_3 P x_2 P x_1$. The pair of preferences that minimizes the expected value of $\varepsilon_2(D_n)$ is $\{P, P'\}$ with $x_3 P x_2 P x_1$ and $x_1 P' x_2 P' x_3$. The associated expected errors are

$$\begin{aligned} \mathbb{E}(\varepsilon_1(D_n)) &= \frac{1}{4}\left(\frac{e^\gamma + e^{2\gamma}}{e^\gamma + e^{2\gamma} + e^{3\gamma}}\right) + \frac{1}{4}\left(\frac{e^\gamma}{e^\gamma + e^{2\gamma}}\right) \\ &\qquad + \frac{1}{4}\left(\frac{e^\gamma}{e^\gamma + e^{3\gamma}}\right) + \frac{1}{4}\left(\frac{e^{2\gamma}}{e^{2\gamma} + e^{3\gamma}}\right) \\ \mathbb{E}(\varepsilon_2(D_n)) &= \frac{1}{4}\left(\frac{e^{2\gamma}}{e^\gamma + e^{2\gamma} + e^{3\gamma}}\right) \end{aligned} \tag{21}$$

while $\varepsilon_3(D_n) = 0$ for every data set D_n. From this, we see that the (expected) error-preference tradeoff curve is convex, and thus it is sufficient to show that $\mathbb{E}(\varepsilon_1(D_n)) - \mathbb{E}(\varepsilon_2(D_n))$ is decreasing in γ.

Using the above displays,

$$\begin{aligned} \mathbb{E}(\varepsilon_1(D_n)) - \mathbb{E}(\varepsilon_2(D_n)) &= \frac{1}{4}\left(\frac{e^\gamma}{e^\gamma + e^{2\gamma} + e^{3\gamma}}\right) + \frac{1}{4}\left(\frac{e^\gamma}{e^\gamma + e^{2\gamma}}\right) \\ &\qquad + \frac{1}{4}\left(\frac{e^\gamma}{e^\gamma + e^{3\gamma}}\right) + \frac{1}{4}\left(\frac{e^{2\gamma}}{e^{2\gamma} + e^{3\gamma}}\right) \end{aligned}$$

Each component of this sum is decreasing in γ, so the sum is as well. The desired result follows.

B.6 Proof of Claim 3

From the proof of Claim 2 above, we already have that $\mathbb{E}(\varepsilon_1(D_n)) - \mathbb{E}(\varepsilon_2(D_n))$ is decreasing in γ. It will be sufficient to show additionally that $\mathbb{E}(\varepsilon_2(D_n)) - \mathbb{E}(\varepsilon_3(D_n))$ is decreasing in γ. Using (21), and that $\varepsilon_3(D_n) = 0$ for every data set D_n, we have

$$\mathbb{E}(\varepsilon_2(D_n)) - \mathbb{E}(\varepsilon_3(D_n)) = \frac{1}{4}\left(\frac{e^{2\gamma}}{e^{\gamma} + e^{2\gamma} + e^{3\gamma}}\right)$$

which is indeed decreasing in γ for all $\gamma > 0$.

B.7 Proof of Proposition 1

First, I will show the following:

Claim 5. *If there exist orderings $P_1, P_2 \in \mathcal{P}$ such that a choice alternative x is ranked first according to P_1, and not last according to P_2, then $\mathcal{P}$ is not identifiable.*

Proof. Consider any set of orderings $\mathcal{P}$ including the preferences P_1, P_2, where x_1 is ranked first according to P_1, x_2 is ranked last according to P_2, and the alternatives x_1 and x_2 are not the same. Fix any data set D that can be perfectly rationalized using $\mathcal{P}$ (that is, $\varepsilon(D, \mathcal{P}) = 0$). I will show by construction that there exists another set of preferences $\mathcal{P}' \neq \mathcal{P}$ with $|\mathcal{P}'| \leq |\mathcal{P}|$ such that also $\varepsilon(D, \mathcal{P}') = 0$.

Define the ordering P_2' to agree with P_2 everywhere, except that it ranks x_1 last. Let $\mathcal{P}' = \mathcal{P} - \{P_2\} + \{P_2'\}$, where the operators denote set addition and subtraction. I will now show that $\varepsilon(D, \mathcal{P}') = 0$.

Suppose towards contradiction that there is some choice observation $(x, A) \in D$ where x is not P-maximal in A for any $P \in \mathcal{P}'$. By assumption, there is some ordering $P^* \in \mathcal{P}$ such that x is P^*-maximal in A. If $P^* \neq P_2$, then also $P^* \in \mathcal{P}'$, which yields a contradiction. So it must be that x is P_2-maximal in A. Now, there are two possibilities: if $x \neq x_1$, then x must also be P_2'-maximal in A, and we are done. If instead $x = x_1$, then x is P_1-maximal in A (by definition of x_1). So we are again done. □

Since every set of three or more orderings satisfies the condition in Claim 5, it immediately follows that every set $\mathcal{P}$ with $|\mathcal{P}| \geq 3$ fails to be identifiable.

Now let us consider an arbitrary set $\mathcal{P} = \{P_1, P_2\}$. Index the alternatives such that x_1 is ranked first according to P_1 and x_2 is ranked first according to P_2. If the condition in Claim 5 is satisfied, it again follows that $\mathcal{P}$ is not identifiable. Suppose otherwise, so that x_1 is ranked last according to P_2 and x_2 is ranked last according to P_1. Define

$$D = \left\{(x, A) \,:\, x \text{ is } P_1\text{-maximal in } A,\ A \in 2^X\right\} \cup \\ \left\{(x, A) \,:\, x \text{ is } P_2\text{-maximal in } A,\ A \in 2^X\right\}.$$

Clearly there is no singleton set $\mathcal{P}'$ such that $\varepsilon(D, \mathcal{P}') = 0$. Suppose towards contradiction that there exists some set $\mathcal{P}' = \{P_1', P_2'\} \neq \mathcal{P}$ such that $\varepsilon(D, \mathcal{P}') = 0$. It must be that x_1 is ranked first according to P_1' and last according to P_2', and that x_2 is ranked first according to P_2' and last according to P_1' (otherwise relabel P_1' and P_2').

Without loss of generality, suppose that $P_1' \neq P_1$.[40] Then there exist alternatives x_i, x_j such that x_i is preferred to x_j under P_1 but not under P_1':

$$x_i P_1 x_j \quad \text{and} \quad x_j P_1' x_i.$$

Take $A := \{x : x_i P_1 x\} \cup \{x_i\}$ to be the set of all alternatives that P_1 ranks weakly lower than x_i. Then $(x_i, A) \in D$. But x_i cannot be P_1'-maximal in A, since $x_j \in A$, and x_1 cannot be P_2'-maximal in A, since $x_2 \in A$. So $\varepsilon(D, \mathcal{P}') > 0$ as desired.

Finally, every singleton set $\mathcal{P} = \{P\}$ is trivially identifiable, taking

$$D = \left\{(x, A) \,:\, \text{ is } P\text{-maximal in } A,\ A \in 2^X\right\}.$$

B.8 Proof of Proposition 2

Consider any choice observation $(x, A) \in \mathbb{C}(\mathcal{P})$. By definition of α, $\nu^*(x, A) > \alpha$, so also $\nu(x, A) > \alpha(1-p)$. Thus, the random variable $Z_n \sim \text{Bin}\,(n, \alpha(1-p))$ first-order stochastically dominates the number of occurrences of (x, A) in the observed data. Since by assumption, $1/\lambda = \tilde{p}n < \alpha(1-p)n$, it follows from Hoeffding's inequality that

$$\begin{aligned} \Pr\left(Z_n < \frac{1}{\lambda}\right) &= \Pr\left(Z_n - \alpha(1-p)n < 1/\lambda - \alpha(1-p)n\right) \\ &\leq \exp\left(-2\left(\tilde{p} - \alpha(1-p)\right)^2 n\right) \end{aligned}$$

Setting $c_1 = 2\left(\tilde{p} - \alpha(1-p)\right)^2 > 0$, it follows that

$$\Pr\left(Z < \frac{1}{\lambda}\right) \leq e^{-c_1 n}.$$

so the probability that (x, A) appears fewer than $1/\lambda$ times in the realized data set is no more than $e^{-c_1 n}$. Taking a union bound, the probability that any $(x, A) \in \mathbb{C}(\mathcal{P})$

[40] Otherwise, $P_2' \neq P_2$ and the remainder of the proof is mirrored.

appears fewer than $1/\lambda$ times in the data is no more than $|\mathbb{C}(\mathcal{P})|e^{-c_1 n}$. Thus, the probability that every $(x, A) \in \mathbb{C}(\mathcal{P})$ appears at least $1/\lambda$ times in the realized data set is at least

$$1 - |\mathbb{C}(\mathcal{P})|e^{-c_1 n}$$

which converges to 1 as the quantity of data n increases. This immediately implies that

$$\Pr\left(\mathbb{C}\left(\mathcal{P}\right) \subseteq \mathbb{C}(\mathcal{P}^*_\lambda(D_n))\right) \to 1 \text{ as } n \to \infty. \tag{22}$$

In the other direction, the random variable $Y_n \sim \text{Bin}(n, p)$ first-order stochastically dominates the number of observations of all $(x, A) \notin \mathbb{C}(\mathcal{P})$. (Informally, Y_n is the number of choice observations "in error.") Thus, $\Pr(Y_n \leq 1/\lambda)$ is an upper bound on the probability that there are $1/\lambda$ realized observations outside of the set $\mathbb{C}(\mathcal{P})$. Since by assumption $1/\lambda = \tilde{p}n > pn$,

$$\Pr(Y_n \geq 1/\lambda) \to 0. \tag{23}$$

Combining this with (22), when feasible, it is optimal to recover preferences whose choice implications are exactly $\mathbb{C}(\mathcal{P})$. But by construction, $\mathbb{C}(\mathcal{P})$ is the set of choice implications corresponding to the set of preferences $\mathcal{P}$, and $\mathcal{P}$ is a valid solution to the problem in (14). Thus,

$$\Pr(\mathbb{C}(\mathcal{P}^*_\lambda(D_n)) = \mathbb{C}(\mathcal{P})) \to 1 \text{ as } n \to \infty$$

as desired.

B.9 Proof of Proposition 3

The following lemma demonstrates a sufficient condition given which the recovered mapping m^*_λ is the true mapping m.

Lemma 4. *Suppose that*

(a) for every context C and every $(x, A) \in \mathbb{C}(m(C))$, the observation $((x, A), C)$ appears at least $2/\lambda$ times in the data, and

(b) there are fewer than $1/\lambda$ observations $((x, A), C)$ where $(x, A) \notin \mathbb{C}(m(C))$.

Then $m^*_\lambda = m$.

Proof. Suppose to the contrary that (a) and (b) are satisfied, but the recovered mapping is some $m' \neq m$. First consider the case in which

$$|m'(\mathscr{C})| \geq |m(\mathscr{C})|. \tag{24}$$

There is some choice context C for which the preference P assigned by mapping m is different from the preference P' assigned by mapping m'. Consider any $(x, A) \in \mathbb{C}(\{P\}) \backslash \mathbb{C}(\{P'\})$. By (a), the observation $((x, A), C)$ appears at least $2/\lambda$ times in the choice data. These observations can be rationalized using m but not by m'. Moreover, by (b), the number of observations that can be rationalized using m' but not by m is no more than $1/\lambda$. So $\varepsilon(D, m) < \varepsilon(D, m')$. Combining this with (24), we have that $|m'(\mathscr{C})| + \lambda\varepsilon(D, m') > |m(\mathscr{C})| + \lambda\varepsilon(D, m)$, and hence m' is not the recovered mapping.

Now suppose that

$$|m'(\mathscr{C})| < |m(\mathscr{C})|.$$

But then there must be at least $|m(\mathscr{C})| - |m'(\mathscr{C})|$ contexts where the preference assigned by m is different from the preference assigned by m'. Call the set of such contexts $\mathscr{C}^*$. For each $C \in \mathscr{C}^*$, there is some $(x, A) \in \mathbb{C}(m(C))$ but $(x, A) \notin \mathbb{C}(m'(C))$. Thus, by (a), there are at least $2/\lambda \cdot (|m(\mathscr{C})| - |m'(\mathscr{C})|)$ observations that can be rationalized under m but not under m'. So $\varepsilon(D, m) - \varepsilon(D, m') < -2/\lambda \cdot |m(\mathscr{C})| - |m'(\mathscr{C})|$. Moreover, again by (b), the number of observations that can be rationalized using m' but not by m is no more than $1/\lambda$. Thus,

$$\begin{aligned}\varepsilon(D, m') - \varepsilon(D, m) &> \frac{2}{\lambda}\left(|m(\mathscr{C})| - |m'(\mathscr{C})|\right) - \frac{1}{\lambda} \\ &> \frac{1}{\lambda}\left(|m(\mathscr{C})| - |m'(\mathscr{C})|\right)\end{aligned}$$

using that $|m(\mathscr{C})| - |m'(\mathscr{C})| \geq 1$. This implies that $|m'(\mathscr{C})| + \lambda\varepsilon(D, m') > |m(\mathscr{C})| + \lambda\varepsilon(D, m)$, so m' is not the recovered mapping. □

Now I will show that conditions (a) and (b) are satisfied with probability converging to 1 as the number of observations gets large. By assumption, every choice observation $((x, A), C)$ with $(x, A) \in \mathbb{C}(m(C))$ is observed with probability at least $\alpha(1 - p)$. Let $E_{((x,A),C)}$ be the event that $((x, A), C)$ is observed fewer than $1/\lambda = \tilde{p}n$ times. Then, application of Hoeffding's inequality gives

$$\Pr(E_{((x,A),C)}) \leq e^{-2(\alpha(1-p)-\tilde{p})^2 n}.$$

Using a union bound, the probability that *any* $E_{((x,A),C)}$ occurs is no more than $\kappa \cdot e^{-2(\alpha(1-p)-\tilde{p})^2 n}$, where $\kappa = |\mathscr{C}| \cdot |\bigcup_{P \in \mathcal{P}} \mathbb{C}(P)|$ is a constant. Or equivalently, the probability that every $((x, A), C)$ is observed at least $1/\lambda$ times is

$$1 - \kappa \cdot e^{-2(\alpha(1-p)-\tilde{p})^2 n}. \tag{25}$$

Thus, the probability that the condition in (a) is satisfied converges to 1 as $n \to \infty$.

To show that the condition in (b) also converges to 1, define Z_n to be the number of observations of tuples $((x, A), C)$ where $(x, A) \notin \mathbb{C}(m(C))$. By assumption, Z_n is first-order stochastically dominated by the random variable $Z \sim \mathrm{Bin}(n, p)$. Since $\mathbb{E}(Z) = pn > \tilde{p}n = 1/\lambda$, it follows that

$$\Pr(Z \geq 1/\lambda) \to 0 \text{ as } n \to \infty \tag{26}$$

implying the same for Z_n. Combining (25) and (26), the desired result directly follows from Lemma 4.

References

Afriat, S. N. (1967): "The Construction of a Utility Function from Expenditure Data," *International Economic Review*, 8, 67–77.

Ambrus, A. and K. Rozen (2013): "Rationalizing Choice with Multi-Self Models," *Economic Journal*.

Apesteguia, J. and M. A. Ballester (2012): "A Measure of Rationality and Welfare," Working Paper.

Bernheim, B. D. and A. Rangel (2009): "Beyond Revealed Preference: Choice Theoretic Foundations for Behavioral Welfare Economics," *Quarterly Journal of Economics*.

Crawford, I. and K. Pendakur (2012): "How Many Types Are There?" *Economic Journal*.

Dean, M. and D. Martin (2010): "How Consistent are your Choice Data?" Working Paper.

——— (2016): "Measuring Rationality with the Minimum Cost of Revealed Preference Violations," *Review of Economics and Statistics*.

Deb, R. (2009): "A testable model of consumption with externalities," *Journal of Economic Theory*, 144.

Echenique, F., S. Lee, and M. Shum (2011): "The Money Pump as a Measure of Revealed Preference Violations," *Journal of Political Economy*, 119.

Einav, L., A. Finkelstein, I. Pascu, and M. Cullen (2012): "How General are Risk Preferences? Choices under Uncertainty in Different Domains," *American Economic Review*.

Famulari, M. (1995): "A Household-Based, Nonparametric Test of Demand Theory," *Review of Economics and Statistics*, 77, 372–383.

Fudenberg, D. and D. Levine (2006): "A Dual-Self Model of Impulse Control," *American Economic Review*, 96, 1449–1476.

Green, J. and D. Hojman (2007): "Choice, Rationality and Welfare Measurement," .

Gross, A. (1989): "Determining the number of violators of the weak axiom." Tech. rep., University of Wisconsin–Milwaukee.

Halevy, Y., D. Persitz, and L. Zrill (2015): "Parametric Recoverability of Preferences," Working Paper.

Houtman, M. and J. Maks (1985): "Determining all Maximal Data Subsets Consistent with Revealed Preference," *Kwantitatieve Methoden*, 19, 89–104.

Kalai, G., A. Rubinstein, and R. Spiegler (2002): "Rationalizing Choice Functions by Multiple Rationales," *Econometrica*, 70, 2481–2488.

Luce, D. and H. Raiffa (1957): *Games and Decisions: Introduction and Critical Survey*, New York: Wiley.

Manzini, P. and M. Mariotti (2007): "Sequentially Rationalizble Choice," *American Economic Review*.

——— (2009): "Categorize Then Choose: Boundedly Rational Choice and Welfare," .

McFadden, D. and M. Richter (1970): "Revealed Stochastic Preference," .

Quandt, R. E. (1956): "A Probabilistic Theory of Consumer Behavior," *Quaterly Journal of Economics*, 70, 507–536.

Rubinstein, A. and Y. Salant (2006): "A model of choice from lists," *Theoretical Economics*, 3.

——— (2008): "(A,f): Choice with Frames," *The Review of Economics Studies.*

Sen, A. (1993): "Internal Consistency of Choice," *Econometrica*, 61.

Train, K. (1986): *Qualitative Choice Analysis*, Cambridge University Press.

Varian, H. R. (1982): "The Nonparametric Approach to Demand Analysis," *Econometrica*, 50, 945–73.