Choice Set Confounding in Discrete Choice

Kiran Tomlinson
Cornell University
kt@cs.cornell.edu

Johan Ugander
Stanford University
jugander@stanford.edu

Austin R. Benson
Cornell University
arb@cs.cornell.edu

ABSTRACT

Standard methods in preference learning involve estimating the parameters of discrete choice models from data of selections (choices) made by individuals from a discrete set of alternatives (the *choice set*). While there are many models for individual preferences, existing learning methods overlook how choice set assignment affects the data. Often, the choice set itself is influenced by an individual's preferences; for instance, a consumer choosing a product from an online retailer is often presented with options from a recommender system that depend on information about the consumer's preferences. Ignoring these assignment mechanisms can mislead choice models into making biased estimates of preferences, a phenomenon that we call *choice set confounding*. We demonstrate the presence of such confounding in widely-used choice datasets.

To address this issue, we adapt methods from causal inference to the discrete choice setting. We use covariates of the chooser for inverse probability weighting and/or regression controls, accurately recovering individual preferences in the presence of choice set confounding under certain assumptions. When such covariates are unavailable or inadequate, we develop methods that take advantage of structured choice set assignment to improve prediction. We demonstrate the effectiveness of our methods on real-world choice data, showing, for example, that accounting for choice set confounding makes choices observed in hotel booking and commute transportation more consistent with rational utility maximization.

CCS CONCEPTS

• **Mathematics of computing** → **Probabilistic inference problems**; • **Information systems** → **Recommender systems**.

KEYWORDS

discrete choice; causal inference; preference learning

ACM Reference Format:
Kiran Tomlinson, Johan Ugander, and Austin R. Benson. 2021. Choice Set Confounding in Discrete Choice. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '21), August 14–18, 2021, Virtual Event, Singapore.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3447548.3467378

KDD '21, August 14–18, 2021, Virtual Event, Singapore

ACM ISBN 978-1-4503-8332-5/21/08…$15.00
https://doi.org/10.1145/3447548.3467378

1 INTRODUCTION

Individual choices drive the success of businesses and public policy, so predicting and understanding them has far-reaching applications in, e.g., environmental policy [12], marketing [3], Web search [22], and recommender systems [57]. The central task of discrete choice analysis is to learn individual preferences over a set of available items (the *choice set*), given observations of people's choices. In recent years, machine learning approaches have enabled more accurate choice modeling and prediction [11, 38, 43, 50]. However, observational choice data analysis has thus far overlooked a crucial fact: the choice set assignment mechanism underlying a dataset can have a significant impact on the generalization of learned choice models, in particular their validity on *counterfactuals*.

Understanding how new choice sets affect preferences in such counterfactuals is key to many applications, such as determining which alternative-fuel vehicles to subsidize or which movies to recommend. In particular, chooser-dependent choice set assignment coupled with heterogeneous preferences can severely mislead choice models, as they do not model the influence of preferences on choice set assignment. Recommender systems are one extreme case, where items are selected specifically to appeal to a user. Such situations also arise in transportation decisions, online shopping, and personalized Web search, resulting in widespread (but often invisible) error in choice models learned from this data.

Drawing on connections with causal inference [24], we term the issue of chooser-dependent choice set assignment *choice set confounding*. Choice set confounding is a major issue for recent machine learning methods whose success is due to capturing deviations from the traditional principles of rational utility maximization that underlie the workhorse multinomial logit model [31]. (Unlike older econometric models of "irrational" behavior [52, 56], these recent methods are practical for modern, large-scale datasets.) These deviations are known as *context effects*, and occur whenever the choice set has an influence on a chooser's preferences. Examples include the *asymmetric dominance effect* [21], where superior options are made to look even better by including inferior alternatives, and the *compromise effect* [45], where intermediate options are preferred (e.g., choosing a medium-priced bottle of wine). While context effects are widespread and worth capturing, choice set confounding can result in spurious effects and over-fitting, and it is unclear if recent machine learning models are learning true effects or simply being misled by chooser-dependent choice set assignment.

In this paper, we formalize when choice set confounding is an issue and show that it can result in arbitrary systems of choice probabilities, even if choosers are rational utility-maximizers (in contrast, tractable choice models only describe a tiny fraction of possible choice systems). We also provide strong evidence of choice set confounding in two transportation datasets commonly used to demonstrate the presence of context effects and to test new models [8, 26, 36, 43]. Then, to manage choice set confounding, we first

adapt two causal inference methods—inverse probability weighting (IPW) and regression controls—to train choice models in the presence of confounding. These methods require chooser covariates satisfying certain assumptions that differ from the traditional causal inference setting. For instance, given access to the same covariates used by a recommender system to construct choice sets, we can reweight the dataset to learn a choice model as if choice sets had been user-independent. Alternatively, we can incorporate covariates into the choice model itself, recovering individual preferences as long as those covariates capture preference heterogeneity.

We also show how to manage choice set confounding without such covariates, as many observational datasets have little information about the individuals making choices. We demonstrate a link between models accounting for context effects and models for choice systems induced by choice set confounding. For example, we derive the context-dependent random utility model (CDM) [44] from the perspective of choice set confounding, by treating the choice set as a vector of substitute covariates (e.g., "someone who is offered item i") in a multinomial logit model.

We develop spectral clustering methods typically used for co-clustering [14] that exploit choice set assignment as a signal for chooser preferences, as a way to improve counterfactual predictions for observed choosers. To show why and when this can work, we frame the problem of finding sufficient chooser covariates as a problem of recovering latent cluster membership in a stochastic block model (SBM) of the bipartite graph that connects choosers to the items in their choice sets.

In addition to theoretical analysis, we demonstrate the efficacy of our methods on real-world choice data. We provide evidence that IPW reduces confounding when modeling hotel booking data, making the choice system more consistent with utility-maximization and making inferred parameters more plausible. For example, the confounded data overweights the importance of price, since many users are shown hotels matching their preferences and select the cheapest one. Factors such as star rating would play a more important role in counterfactuals. We also evaluate our clustering approach on online shopping data. By training separate models for different chooser clusters, we outperform a mixture model that attempts to discover preference heterogeneity from choices alone, ignoring the signal from choice set assignment.

All of our code, results, and links to our data are available at https://github.com/tomlinsonk/choice-set-confounding.

1.1 Additional related work

This research is inspired by recent computational advances in learning context-dependent preferences [11, 35, 38, 43, 50]. These methods exhibit strong gains by exploiting context effects but are often evaluated on data with possible choice set confounding. Similar confounding issues are well-studied in rating and ranking data within recommender systems [30, 42, 53, 54], but these approaches do not directly apply to choice data. The causal inference ideas that we develop are based on long-standing methods [23, 24]; the challenge we address is how to adapt them for discrete choice data.

The role of choice set assignment does ocassionally appear in the choice literature. For instance, Manski used choice set assignment probabilities to derive random utility models [28]. More often, traditional choice theory has focused on latent *consideration sets*, which are subsets of alternatives that are actually considered by choosers [6, 10] where non-uniform choice set probabilities play a key role. In another setting, Manski and Lerman [29] used an approach similar to our inverse probability weighting. They were concerned with "choice-based samples," where we first sample an item and then get an observation of a chooser who selected that item (usually, we sample a chooser and then observe their choice) [29].

The use of regression controls in discrete choice (i.e., including chooser covariates in the utility function) is standard in econometrics [9, 47, 51]. However, in these settings, regression aims to understand how the attributes of an individual affect decision-making, which can unknowingly and accidentally help with confounding. This may explain why choice set confounding has not been widely recognized (additionally, in an interview, Manksi discusses that choice set generation has been under-explored [48]). We formalize when and how regression adjusts for choice set confounding.

2 DISCRETE CHOICE BACKGROUND

We start with some notation and the basics of discrete choice models. Let $\mathcal{U}$ denote a universe of n items and $\mathcal{A}$ a population of individuals. In a discrete choice setting, a *chooser* $a \in \mathcal{A}$ is presented a nonempty *choice set* $C \subseteq \mathcal{U}$ and they choose one item $i \in C$. Specifically, a is sampled with probability $\Pr(a)$, then C is presented to a with probability $\Pr(C \mid a)$, and finally a selects i with probability $\Pr(i \mid a, C)$. Most discrete choice analysis focuses only on $\Pr(i \mid a, C)$ or $\Pr(i \mid C)$, but we consider this entire process. A discrete choice dataset $\mathcal{D}$ is a collection of tuples (C, i) generated by this process. We use $C_{\mathcal{D}}$ to denote the set of unique choice sets in $\mathcal{D}$.

Discrete choice models posit a parametric form for choice probabilities, with parameters learned from data. The *universal logit* [33] can express any system of choice probabilities (called a *choice system*). Under a universal logit, each chooser a has a scalar utility $u_i(C, a)$ for item i in choice set C. Choice probabilities are then a softmax of these utilities: $\Pr(i \mid a, C) = \exp(u_i(C,a)) / \sum_{j \in C} \exp(u_j(C,a))$. This arises from a notion of rational utility-maximization [51]. Specifically, these are the choice probabilities if a observes random utilities $u_i(C, a) + \epsilon$ (where the ϵ are i.i.d. Gumbel-distributed for each item and choice) and selects the item with maximum observed utility. The above model has too many degrees of freedom to be practical (e.g., it has entirely separate parameters for every chooser a), and typically one assumes utilities are fixed across sets and individuals. This is the *logit* model [31], where $u_i(C, a) = u_i, \forall C, \forall a$.

Other discrete choice models come from different assumptions on $u_i(C, a)$, trading off descriptive power for ease of inference and interpretation. For example, we may have access to a vector of covariates $\boldsymbol{x_a} \in \mathbb{R}^{d_x}$ for person a. Similarly, an item i may be described by a vector of features $\boldsymbol{y_i} \in \mathbb{R}^{d_y}$. We can write $u_i(C, a)$ as a function of $\boldsymbol{x_a}, \boldsymbol{y_i}$, or both, yielding several choice models (Table 1)—the multinomial logit (MNL), conditional logit (CL), and conditional multinomial logit (CML).[1] All of these models obey a common assumption, the *independence of irrelevant alternatives*

[1] "MNL" sometimes refers to logit and conditional logit. Here, we follow the convention [20] that "multinomial" means chooser covariates are used and "conditional" means item features are used. Additionally, for CML, we assume $\boldsymbol{\gamma_i} = B^T \boldsymbol{x_a}$, which reduces the number of parameters from $d_y + nd_x$ to $d_y(d_x + 1)$, allowing us to use the model when the number of items is prohibitively large.

Table 1: Discrete choice models. The item and chooser feature vectors $\boldsymbol{y_i}$ and $\boldsymbol{x_a}$ are part of the dataset, while $u_i \in \mathbb{R}, \boldsymbol{\theta} \in \mathbb{R}^{d_y}, \boldsymbol{\gamma_i} \in \mathbb{R}^{d_x}$, and $B \in \mathbb{R}^{d_y \times d_x}$ are learned parameters.

Model	$u_i(C, a)$	# Parameters
logit	u_i	n
multinomial logit (MNL)	$u_i + \boldsymbol{x_a}^T \boldsymbol{\gamma_i}$	$n(d_x + 1)$
conditional logit (CL)	$\boldsymbol{y_i}^T \boldsymbol{\theta}$	d_y
cond. mult. logit (CML)	$\boldsymbol{y_i}^T (\boldsymbol{\theta} + B\boldsymbol{x_a})$	$d_y(d_x + 1)$

Table 2: Context effect models. $p_{ij} \in \mathbb{R}, \boldsymbol{\gamma_i} \in \mathbb{R}^{d_x}, \boldsymbol{\theta} \in \mathbb{R}^{d_y}, A \in \mathbb{R}^{d_y \times d_y}, B \in \mathbb{R}^{d_y \times d_x}$ are learned parameters.

Model	$u_i(C, a)$	# Parameters
CDM [43]	$\sum_{j \in C \setminus i} p_{ij}$	$n(n-1)$
mult. CDM (MCDM)	$\sum_{j \in C \setminus i} p_{ij} + \boldsymbol{x_a}^T \boldsymbol{\gamma_i}$	$n(n + d_x)$
LCL [50]	$\boldsymbol{y_i}^T (\boldsymbol{\theta} + A\boldsymbol{y_C})$	$d_y(d_y + 1)$
mult. LCL (MLCL)	$\boldsymbol{y_i}^T (\boldsymbol{\theta} + A\boldsymbol{y_C} + B\boldsymbol{x_a})$	$d_y(d_y + d_x + 1)$

(IIA) [51]. IIA states that relative choice probabilities are conserved across choice sets: $\Pr(i|a,C)/\Pr(j|a,C) = \Pr(i|a,C')/\Pr(j|a,C')$. To be precise, this is individual-level rather than group-level IIA. Among the models in Table 1, the latter is only obeyed by the logit and conditional logit. In general, models obey individual-level IIA if utility is independent of C, i.e., $u_i(C, a) = u_i(a)$ and obey group-level IIA if $u_i(C, a)$ is independent of both C and a.

While the IIA assumption is convenient, it is commonly violated through context effects [8, 21, 46]. Due to the ubiquity of context effects, models incorporating information from the choice set have become increasingly popular and have shown considerable success [11, 38, 43, 50]. Other models allow IIA violations without explicitly modeling effects of the choice set [8, 32, 36].

We briefly introduce two of these context effect models, the *context-dependent random utility model* (CDM) [43] and the *linear context logit* (LCL) [50], that we use extensively. In the CDM, each item in the choice set exerts a pull on the utility of every other item: $u_i(C, a) = \sum_{j \in C \setminus i} p_{ij}$. The CDM can be derived as a second-order approximation to universal logit (where plain logit is the first-order approximation) [43]. The LCL instead operates in settings with item features, adjusting the conditional logit parameter $\boldsymbol{\theta}$ according to a linear transformation of the choice set's mean feature vector: $u_i(C, a) = \boldsymbol{y_i}^T (\boldsymbol{\theta} + A\boldsymbol{y_C})$, where $\boldsymbol{y_C} = 1/|C| \sum_{j \in C} \boldsymbol{y_j}$. To incorporate chooser covariates, we define multinomial versions of these models (Table 2). For this paper, LCL and CDM should be thought of as the simplest context effect models with and without item features.

In contrast, *mixed logit* [32] accounts for group-level rather than individual-level IIA violations and has a different structure than any of the models introduced thus far. A (discrete) mixed logit is a mixture of K logits with mixing proportions $\pi_1, \ldots, \pi_K$ such that $\sum_{k=1}^{K} \pi_k = 1$. With $u_i(a_k)$ denoting the utility of the kth component for item i, a mixed logit has choice probabilities

$$\Pr(i \mid C) = \sum_{k=1}^{K} \pi_k \frac{\exp(u_i(a_k))}{\sum_{j \in C} \exp(u_j(a_k))}. \tag{1}$$

This can result in a choice system violating IIA but not because any individual chooser experiences context effects. Rather, the aggregation of several choosers each obeying IIA can result in IIA violations.

3 CHOICE SET CONFOUNDING

The traditional approach to choice modeling is to learn a single model for $\Pr(i \mid C)$ (such as a logit) and assume it represents overall choice behavior, namely, that the model accurately reflects average choice probabilities $\mathrm{E}_a[\Pr(i \mid a, C)]$. However, $\Pr(i \mid C)$ need not represent average choice behavior at all, as this is only guaranteed under restrictive independence assumptions.

Observation 1. *If, for all $a \in \mathcal{A}, C \in \mathcal{C}_\mathcal{D}, i \in C$, at least one of*

(1) $\Pr(C) = \Pr(C \mid a)$ *(chooser-independent choice sets) or*
(2) $\Pr(i \mid a, C) = \Pr(i \mid C)$ *(chooser-independent preferences)*

holds, then $\Pr(i \mid C) = \mathrm{E}_a[\Pr(i \mid a, C)$. *If both conditions are violated, then this equality can fail.*

Appendix A has a proof for this fact and other theoretical statements presented later. When we have both chooser-dependent sets and preferences, observed choice probabilities $\Pr(i \mid C)$ can differ significantly from true aggregate choice probabilities $\mathrm{E}_a[\Pr(i \mid a, C)]$. We call this phenomenon *choice set confounding*, and provide the following toy example as an illustration.

Example 1. Let $\mathcal{U} = \{\mathsf{cat}, \mathsf{dog}, \mathsf{fish}\}$. Choosers are either cat people or dog people choosing a pet, with choice probabilities

	$\{\mathsf{cat}, \mathsf{dog}\}$	$\{\mathsf{cat}, \mathsf{dog}, \mathsf{fish}\}$
cat person	3/4, 1/4	3/4, 1/4, 0
dog person	1/4, 3/4	1/4, 3/4, 0

Note that the preferences of cat and dog people do not change when fish are included in the choice set. Choice sets are assigned non-independently: cat people see $\{\mathsf{cat}, \mathsf{dog}\}$ w.p. 3/4 and $\{\mathsf{cat}, \mathsf{dog}, \mathsf{fish}\}$ w.p. 1/4 (vice-versa for dog people). Let the population consist of 1/4 cat people and 3/4 dog people. If we only observe samples (C, i) without knowing who is a cat person and who is a dog person,

$$\Pr(\mathsf{dog} \mid \{\mathsf{cat}, \mathsf{dog}\}) = 1/2 \cdot 1/4 + 1/2 \cdot 3/4 = 1/2$$
$$\Pr(\mathsf{dog} \mid \{\mathsf{cat}, \mathsf{dog}, \mathsf{fish}\}) = 1/10 \cdot 1/4 + 9/10 \cdot 3/4 = 7/10.$$

However,

$$\mathrm{E}_a[\Pr(\mathsf{dog} \mid a, \{\mathsf{cat}, \mathsf{dog}\})] = 1/4 \cdot 1/4 + 3/4 \cdot 3/4 = 5/8$$
$$\mathrm{E}_a[\Pr(\mathsf{dog} \mid a, \{\mathsf{cat}, \mathsf{dog}, \mathsf{fish}\})] = 1/4 \cdot 1/4 + 3/4 \cdot 3/4 = 5/8.$$

This mismatch is especially problematic for models that use choice-set dependent utilities $u_i(C)$, such as those designed to account for context effects. From the above data, we might conclude that the presence of a fish causes a dog to become a more appealing option. This spurious context effect would be seized upon by context-based models and even result in improved predictive performance on test data drawn form the same distribution. However, these models would make biased predictions on counterfactual examples where sets are chosen from a different distribution.

In reality, no one's choice would be affected by adding fish to their choice set—it's a red herring. This is a *causal inference* problem. We want to know the cause of a choice, but we are being misled as to whether the change in preferences between the $\{\mathsf{cat}, \mathsf{dog}\}$ and $\{\mathsf{cat}, \mathsf{dog}, \mathsf{fish}\}$ choice sets is due to the presence of fish or to a hidden confounder: the underlying preferences of cat and dog people, coupled with chooser-dependent choice set assignment.

Extending this idea, the equality in Observation 1 can fail dramatically. If the population consists of individuals each of whom obeys

IIA (i.e., chooses according to a logit), then $\mathrm{E}_a[\Pr(i \mid a, C)]$ is exactly the mixed logit choice probability. On the other hand, $\Pr(i \mid C)$ can express an arbitrary choice system with choice set confounding.

Theorem 2. *Mixed logit with chooser-dependent choice sets is powerful enough to express any system of choice probabilities.*

Arbitrary choice systems are much more powerful than mixed logit (even ones with continuous mixtures). For example, it is impossible for mixed logit to violate *regularity*, the condition that $\Pr(i \mid C) \geq \Pr(i \mid C \cup \{j\})$ for all $C \subseteq \mathcal{U}, i \in C, j \in \mathcal{U}$, as choice probabilities for i can only go down in each mixture component when we include j. On the other hand, even Example 1 has a regularity violation (picking a dog is more likely when a fish is available), despite there being only two types of choosers, both adhering to IIA.

We have shown that choice set confounding is an issue in theory, and we now demonstrate it to be a problem in practice. We present evidence of choice set confounding in two transportation choice datasets, SF-WORK and SF-SHOP [26]. These datasets consist of San Francisco (SF) resident surveys for preferred transportation mode to work or shopping, where the choice set is the set of modes available to a respondent. The SF datasets are common testbeds for choice models violating IIA [8, 26, 36, 43] and in choice applications [2, 49].

The SF data have regularity violations (Table 5 in Appendix C), ruling out the possibility that the IIA violations in these datasets are just due to mixtures of choosers obeying IIA. Thus, these datasets either have (1) true context effects or (2) choice set confounding. So far, the literature has focused on (1), but we argue that (2) is more likely. We compare the likelihoods of logit, MNL, CDM, and MCDM (recall Tables 1 and 2) on these datasets through likelihood-ratio tests (Table 3). MNL and MCDM both account for chooser-dependent preferences through covariates, while CDM and MCDM both account for context effects. With true context effects, we would expect CDM to be significantly more likely than logit and MCDM to be significantly more likely than MNL. However, this is not the case. While CDM is significantly more likely than logit, MCDM is not significantly more likely than MNL in both SF datasets. Thus, context effects only appear significant before controlling for preference heterogeneity through covariates. This is exactly what we would expect if the IIA violations in these datasets are due to choice set confounding rather than context effects. In contrast, we see significant context effects in the EXPEDIA hotel-booking dataset [25] even after controlling for covariates (this dataset uses item features, hence the different models in Table 3), so context effects are likely. This dataset consists of search results (choice sets) and hotel bookings (choices), and we explore it further in Section 4.4.

The choice set confounding leads to a key question: how were choice sets constructed in SF-WORK and SF-SHOP? According to Koppelman and Bhat [26], choice sets were imputed based on chooser covariates from the survey. For instance, walking was included as an option if a respondent's distance to the destination was < 4 miles and driving was included if they had a driver's license and at least one car in their household [26]. This choice set assignment is highly chooser-dependent, resulting in strong choice set confounding.

Example 1 and the SF datasets highlight how confounding can lead to spurious context effects and incorrect average choice probabilities. Next, in Section 4, we adapt methods from causal inference so that chooser covariates can correct choice probability estimates.

Table 3: Likelihood gains in SF-WORK, SF-SHOP, and EXPEDIA from covariates and context with likelihood ratio test (LRT) p-values. $\Delta\ell$ denotes improvement in log-likelihood.

Comparison	Testing	Controlling	$\Delta\ell$	LRT p
SF-WORK				
Logit to MNL	covariates	—	883	$< 10^{-10}$
Logit to CDM	context	—	85	$< 10^{-10}$
CDM to MCDM	covariates	context	819	$< 10^{-10}$
MNL to MCDM	context	covariates	20	0.08
SF-SHOP				
Logit to MNL	covariates	—	343	$< 10^{-10}$
Logit to CDM	context	—	96	$< 10^{-10}$
CDM to MCDM	covariates	context	276	$< 10^{-10}$
MNL to MCDM	context	covariates	29	0.36
EXPEDIA				
CL to CML	covariates	—	1218	$< 10^{-10}$
CL to LCL	context	—	2345	$< 10^{-10}$
LCL to MLCL	covariates	context	1167	$< 10^{-10}$
CML to MLCL	context	covariates	2294	$< 10^{-10}$

And in Section 5, we address what can be done without covariates if we want to (1) make predictions under chooser-dependent choice set assignment mechanisms or (2) make counterfactual predictions for previously observed choosers.

4 CAUSAL INFERENCE METHODS

In traditional causal inference [23, 24, 39], we wish to estimate the causal effect of an intervention (e.g., a medical treatment) from observational data. However, we cannot simply compare the outcomes of the treated and untreated cohorts if treatment was not randomly assigned—confounders might affect both whether someone was treated and their outcome. There are many methods to debias treatment effect estimation, including matching [37, 39], inverse probability weighting (IPW) [19], and regression [40]. One can also combine methods, such as IPW and regression, which is the basis for *doubly robust* estimators [5].

Here, we adapt causal inference methods to estimate unbiased discrete choice models from data with choice set confounding. First, we adapt IPW to learn unbiased models that do not use chooser covariates in the utility function. After, we show an equivalence between incorporating chooser covariates in the utility function and regression for causal inference. Finally, we combine these methods for doubly robust choice model estimation. For discrete choice, these methods require new assumptions and have different guarantees. We first provide a brief introduction to causal inference terminology in the binary treatment setting, such as an observational medical study (in contrast, we will think of choice sets as treatments).

In potential outcomes notation [41], each person i has covariates X_i and is either treated ($T_i = 1$) or untreated ($T_i = 0$). At some point after treatment, we measure the *outcome* $Y_i(T_i)$. A typical goal of the causal inference methods above is to estimate the *average treatment effect* $\mathrm{E}_i[Y_i(1) - Y_i(0)]$. All of these methods rely on untestable assumptions; in particular, they rely on *strong ignorability* [23, 37] (also called *unconfoundedness* or *no unmeasured confounders*), which

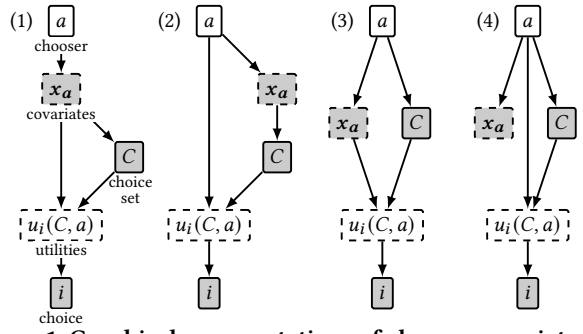

Figure 1: Graphical representations of chooser covariate assumptions: (1) ignorability; (2) choice set ignorability; (3) preference ignorability; (4) no ignorability. Shaded nodes are observed, dashed nodes are deterministic.

requires that the treatment is independent from the outcome, conditioned on observed covariates: $\Pr(T_i \mid X_i, Y_i) = \Pr(T_i \mid X_i), \forall i$.

4.1 Inverse probability weighting

IPW estimation commonly requires estimating *propensity scores* describing the probability of each treatment assignment given individual covariates. The true probabilities $\Pr(T_i \mid X_i)$ are unknown, so estimated "propensities" $\widehat{\Pr}(T_i \mid X_i)$ are learned from observed data, typically via logistic regression [4]. Propensities can then be used to estimate average treatment effects or, as in our case, to re-weight a model's training data [16]. By weighting each sample by the inverse of its propensity, we effectively construct a *pseudo-population* where treatment is assigned independently from covariates. In addition to ignorability, IPW requires *positivity*, the assumption that all propensities satisfy $0 < \Pr(T_i \mid X_i) < 1$.

In the discrete choice setting, we think of choice sets as treatments. By Observation 1, we need chooser-independent choice sets in order to learn an unbiased choice model. Our idea of IPW for discrete choice is to create a pseudo-dataset in which this is true and to learn a choice model over that pseudo-dataset. To do this, we model choice set assignment probabilities $\Pr(C \mid a)$. We can then replace each sample (i, a, C) with $1/[|C_{\mathcal{D}}| \Pr(C \mid a)]$ copies, creating a pseudo-dataset $\tilde{\mathcal{D}}$ with uniformly random choice sets (note that we allow "fractional samples," since we don't explicity construct $\tilde{\mathcal{D}}$). However, we cannot hope to learn $\Pr(C \mid a)$ in datasets with only a single observation per chooser (which is very often the case). We instead need to rely on observed covariates $\boldsymbol{x_a}$. We thus learn $\Pr(C \mid \boldsymbol{x_a})$ and use these propensities to construct $\tilde{\mathcal{D}}$. For the analysis, we assume we know the true propensities, but a correctly specified choice set assignment model learned from data is sufficient.

To learn a choice model from $\tilde{\mathcal{D}}$, we can simply add weights to the model's log-likelihood function, resulting in

$$\ell(\theta; \tilde{\mathcal{D}}) = \sum_{(i,C,a) \in \mathcal{D}} \frac{\log \Pr_\theta(i \mid C)}{|C_{\mathcal{D}}| \Pr(C \mid \boldsymbol{x_a})}. \tag{2}$$

In order for $\Pr(C \mid \boldsymbol{x_a})$ to be an effective stand-in for $\Pr(C \mid a)$, we need the following assumption (see Figure 1).

Definition 3. Choice set ignorability is satisfied if choice sets are independent of choosers, conditioned on chooser covariates: $\Pr(C \mid a, \boldsymbol{x_a}) = \Pr(C \mid \boldsymbol{x_a})$.

Just as in standard IPW, we also need positivity (of choice set propensities). Under these assumptions, IPW guarantees that empirical choice probabilities in the pseudo-dataset $\tilde{\mathcal{D}}$ reflect aggregate choice probabilities in the true population. To formalize this, we introduce $\mathcal{D}^*$, an idealized dataset with uniformly random choice set assignment for every chooser (of the same size as $\mathcal{D}$). $\mathcal{D}^*$ consists of $|\mathcal{D}|$ independent samples (a, C, i) each occuring with probability $\Pr(a) \frac{1}{|C_{\mathcal{D}}|} \Pr(i \mid a, C)$. We now show that the IPW-weighted log-likelihood (eq. (2)) is, in expectation, the same as the log-likelihood function over $\mathcal{D}^*$. Since $\mathcal{D}^*$ has chooser-independent choice sets, we can train a model for $\Pr(i \mid C)$ using eq. (2) and expect it to capture unbiased aggregate choice probabilities (by Observation 1).

Theorem 4. *If, for all* $a \in \mathcal{A}, C \in C_{\mathcal{D}}$,

(1) $0 < \Pr(C \mid \boldsymbol{x_a}) < 1$ *(positivity), and*

(2) $\Pr(C \mid a, \boldsymbol{x_a}) = \Pr(C \mid \boldsymbol{x_a})$ *(choice set ignorability),*

then $E_{\mathcal{D}}[\ell(\theta; \tilde{\mathcal{D}})] = E_{\mathcal{D}^*}[\ell(\theta; \mathcal{D}^*)]$.

Choice set ignorability is crucial to the success of IPW, so we should assess when this assumption is reasonable. If choice sets are generated by an exogenous process (such as a recommender system, as in the EXPEDIA dataset), then as long as we have access to the same covariates as that process, choice set ignorability holds, although learning the propensities may still be a challenge. However, in other datasets, choice sets are formed through self-directed browsing (e.g., clicking around an online shop, as in the YOOCHOOSE dataset we examine later). In those cases, basic covariates (age, gender, etc.) are unlikely to fully capture choice set generation, since sets result from the complexities of human behavior rather than the simpler algorithmic behavior of a recommender system. As in traditional causal inference, the validity of choice set ignorability must be determined by the practitioner applying the method.

4.2 Regression

An alternative to using chooser covariates to learn choice set propensities is to incorporate covariates directly into the utility formulation, as in the multinomial or conditional multinomial logit models. If chooser covariates fully capture their preferences and the choice model is correctly specified, then the model that we learn is consistent. We formalize the first condition as follows (see Figure 1).

Definition 5. Preference ignorability is satisfied if choice probabilities are independent of choosers, conditioned on chooser covariates: $\Pr(i \mid a, \boldsymbol{x_a}, C) = \Pr(i \mid \boldsymbol{x_a}, C)$.

Given correct specification and preference ignorability, the choice model will be consistent in terms of aggregate choice probabilities and result in accurate individual choice probability estimates.

Theorem 6. *If* $\Pr(i \mid a, \boldsymbol{x_a}, C) = \Pr(i \mid \boldsymbol{x_a}, C)$ *for all* $a \in \mathcal{A}, C \in C_{\mathcal{D}}, i \in C$ *(preference ignorability), then the MLE of a correctly specified (and well-behaved, in the standard MLE sense [55, Theorem 9.13]) choice model that incorporates chooser covariates* $\boldsymbol{x_a}$ *is consistent:* $\lim_{|\mathcal{D}| \to \infty} \widehat{\Pr}(i \mid \boldsymbol{x_a}, C) = \Pr(i \mid a, C)$.

While the guarantee of regression is stronger than IPW, preference ignorability is more challenging to satisfy in practice. Instead of needing all covariates used to generate choice sets, we need covariates to fully describe choice behavior.

4.3 Doubly robust estimation

A constraint of both IPW and regression is correct model specification, either of the choice set propensity model or of the choice model. In traditional causal inference, one can combine both methods to provide guarantees if either model is correctly specified, producing *doubly robust* estimators [5, 17]. In the same way, we can combine IPW and regression for choice models and achieve their respective guarantees if their respective conditions are satisfied. In other words, the two methods do not interfere with each other. However, this increases the variance of estimates, so it may be advisable to only use one method if we are confident in one of the assumptions.

4.4 Empirical analysis of IPW and regression

We begin by evaluating regression and IPW adjustments in synthetic data, and then apply our methods to the EXPEDIA dataset (training details in Appendix C).

Counterfactual evaluation in synthetic data. We generate synthetic data with heterogeneous preferences, CDM-style context effects, and choice set confounding. Specifically, we use 20 items with embeddings $\boldsymbol{y_i} \in \mathbb{R}^2$ sampled uniformly from the unit circle. We also generate embeddings $\boldsymbol{x_a}$ in the same way for each chooser a. Each chooser a picks items according to an MCDM, where the utility for i is a sum of $\boldsymbol{x_a^T y_i}$ plus a CDM term shared by all choosers, with each "push/pull" term $p_{ij} \sim \text{Uniform}(-1, 1)$. To generate a choice set for a, we sample a uniformly random set with probability 0.25 (to satisfy positivity) and otherwise include each item with probability $1/(1+e^{-c\boldsymbol{x_a^T y_i}})$, where c is the *confounding strength* (we condition on having at least two items in the choice set). Higher confounding strength results in sets containing items more preferred by a. Each trial consists of 10000 samples. Item embeddings are unobserved, but chooser embeddings are used as covariates. We train models on a confounded portion of the data and measure prediction quality on a held-out confounded subset as well as a counterfactual portion with uniformly random choice sets. For IPW, we estimate choice set propensities via per-item logistic regression, multiplying item propensities to get set propensities.

To measure prediction quality, we use the mean relative position of the true choice in the list of predictions sorted in descending probability order. A value of 1 says that the true choices were all predicted as most likely. As confounding strength increases, prediction quality increases in the confounded data for logit, MNL, and CDM, while decreasing on counterfactual data (Figure 2). For logit and MNL, IPW leads to models that generalize better to counterfactual data. For CDM, IPW correctly prevents the illusion of increased performance with more confounding (although variance caused by IPW appears to result in a small dip in performance at low confounding). Since preference ignorability is satisfied, IPW is unnecessary for MCDM: regression with the correctly specified model successfully generalizes despite confounding.

Empirical data with chooser covariates. We now consider the EXPEDIA hotel choice dataset [25] from Section 3, using five hotel features: star rating, review score, location score, price, and promotion status. This allows us to use feature-based choice models (CL, CML, LCL, and MLCL; Tables 1 and 2). The dataset includes information about chooser searches, such as the number of adults and children in their party, which likely have strong effects on choice sets

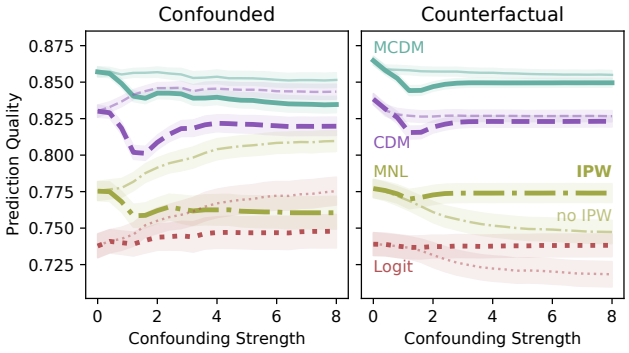

Figure 2: Mean prediction quality of models on synthetic data with both context effects and choice set confounding, with IPW (bold) and without IPW (light). Left: out-of-sample predictions on data with confounding. Right: counterfactual predictions of models trained on confounded data. Shaded regions show standard error over 16 trials.

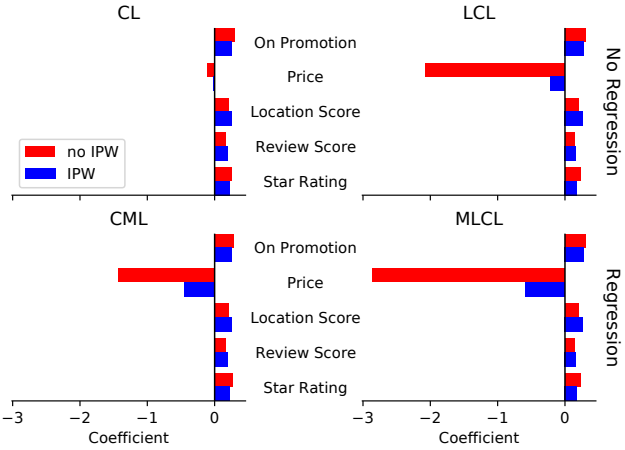

Figure 3: Preference coefficients θ in EXPEDIA for CL and LCL (top row, no regression); and CML and MLCL (bottom row, with regression), with and without IPW. A higher coefficient means choosers prefer higher values of the feature.

(i.e., search results). This is an excellent testbed for IPW since these covariates are likely informative about choice sets, making choice set ignorability more reasonable than preference ignorability.

We do not have counterfactual choices for the EXPEDIA data, but we still consider several types of analysis. First, we recall the results from Table 3 to see if apparent context effects are accounted for by chooser covariates. There, in contrast to the SF datasets, context effects still appear significant after controlling for covariates. In fact, context effects provide a larger likelihood boost than the chooser covariates. Thus, either (1) there are true context effects or (2) the chooser covariates in EXPEDIA do not satisfy preference ignorability (or both). Based on the nature of the covariates, (2) seems very likely: the number of children in the chooser's party and the length of their stay are unlikely to fully describe hotel preferences.

Since regression is inconclusive, we also apply IPW. To learn choice set propensities, we use a probabilistic model of the mean feature vectors of choice sets. We assume these vectors follow a multivariate Gaussian conditioned on chooser covariates, with

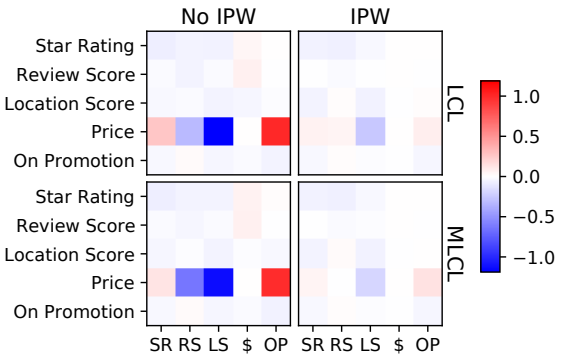

Figure 4: LCL and MLCL context effect matrix A in EXPEDIA with and without IPW. A higher value means choosers prefer a row feature more in a set where the mean column feature (abbreviated) is high; 0 indicates no context effect.

Table 4: Log-likelihoods and estimated random-set log-likelihoods with IPW on EXPEDIA. After adjusting for confounding, the data is far easier to explain.

Model	Confounded	IPW-adjusted
CL	−839499	−786653
CML	−838281	−785753
LCL	−837154	−784770
MLCL	−835986	−783928

mean $W\boldsymbol{x_a}+\boldsymbol{z}$ for some $W \in \mathbb{R}^{d_y \times d_x}$, $\boldsymbol{z} \in \mathbb{R}^{d_y}$. Given observed mean choice set vectors $\boldsymbol{y_C}$ and corresponding chooser covariates $\boldsymbol{x_a}$, we compute the maximum-likelihood W, $\boldsymbol{z}$, and covariance matrix (see Appendix B). This model gives us propensities for any (a, C) pair.

Using IPW with these propensities dramatically decreases the negative impact of high price in all four models (Figure 3). After adjusting for confounding, the models indicate that users are more willing to book more expensive hotels. This makes sense if Expedia is recommending relevant hotels: among a set of hotels matching a user's desired characteristics (such as location and star rating), we would expect them to select the cheapest option. On the other hand, if we presented users with a set of random hotels, location and star rating might play a stronger role in determining their choice, since a random set might have many cheap hotels that are undesirable for other reasons. In addition to the preference coefficients, IPW affects the context effect matrix A in the LCL and MLCL (Figure 4). In both models, IPW decreases (but does not entirely eliminate) the strong price context effects. This is evidence that some of the apparent context effects in the dataset are due to choice set confounding.

Finally, the estimated likelihoods of the models under IPW are significantly better than without IPW (Table 4). We normalize the IPW-weighted log-likelihood by the sum of the IPW weights, which provides an estimate of what the IPW-trained model's log-likelihood would be given random sets. The gap between likelihood with no IPW and estimated likelihood with IPW dwarfs the gaps between different choice models, indicating that accounting for choice set confounding makes the data much more consistent with the random utility maximization principle underlying all four models. (By Theorem 2, choice set confounding can result in choice systems far from rational behavior, even when choosers are rational.)

5 MANAGING WITHOUT COVARIATES

So far, we have used chooser covariates to correct for choice set confounding. However, in some choice data, there are no covariates available, or we are not willing to make ignorability assumptions. Here, we show what can be done in this setting.

5.1 Within-distribution prediction

Unfortunately, by Theorem 2, it is impossible to determine whether IIA violations are caused by choice set confounding or true context effects in the absence of chooser information. Nonetheless, we can still exploit IIA violations—whatever their origin—to improve prediction, as long as we are careful not to make counterfactual predictions. This is essentially what researchers developing context effect models [11, 36, 38, 43, 50] have been doing (without a framework for understanding the possibility of choice set confounding and the associated risks for counterfactual prediction). Beyond emphasizing a need for caution, we also establish a duality between models accounting for context effects and models accounting for choice set confounding; specifically, we show that a model equivalent to the CDM—which was designed with context effects in mind—can be derived purely from the perspective of choice set confounding.

In a multinomial logit (MNL), we learn a latent parameter vector $\boldsymbol{\gamma_i}$ for each item $i \in \mathcal{U}$ and model utilies as $u_i(a) = \boldsymbol{x_a}^T\boldsymbol{\gamma_i}$ (omitting the intercept term). Suppose we don't have any chooser covariates, but we know choice set assignment depends on choosers. We could then use the choice set itself as a surrogate for user covariates (e.g., one covariate could be "someone who is offered item i"). Let $\mathbf{1}_{C_a}$ be a binary encoding of the choice set C_a of a chooser a (a length $|\mathcal{U}|$ vector with a 1 in position i if $i \in C_a$). Consider treating $\mathbf{1}_{C_a}$ as a substitute for the user covariates $\boldsymbol{x_a}$. Then the MNL model is

$$\Pr(i \mid C_a) = \frac{\exp(\mathbf{1}_{C_a}^T\boldsymbol{\gamma_i})}{\sum_{j \in C_a}\exp(\mathbf{1}_{C_a}^T\boldsymbol{\gamma_j})}.$$

The utility of i in set C_a is $\sum_{j \in C_a}\gamma_{ij}$, which is exactly the CDM (with self-pulls, since the sum is over C_a rather than $C_a \setminus i$), a model designed to capture choice-set-dependent utilities. Thus, the CDM can either be thought of as accounting for pairwise interactions between items or using the choice set as a stand-in for user covariates.

One natural question this duality raises is how the set of choice systems expressible by CDM (or other context-effect models) compares to the choice systems induced by mixed populations of IIA choosers with choice set confounding, which take the form

$$\Pr(i \mid C) = \sum_{a \in \mathcal{A}}\Pr(a \mid C)\frac{\exp(u_i(a))}{\sum_{j \in C}\exp(u_j(a))}. \tag{3}$$

Mixtures of logits such as eq. (3) are notoriously hard to analyze (even in two-component case [13]), so no simple equivalence between a context-effect model and such a mixture is likely. In fact, eq. (3) is even trickier than standard mixed logit (eq. (1)), since the mixture weights depend on the choice set.

Nonetheless, some progress in this direction is possible. Here, we provide an instance where the LCL approximates a choice system induced by choice set confounding (of the form of eq. (3)). Recall that the LCL has utilities $u_i(a, C) = (\boldsymbol{\theta}+A\boldsymbol{y_C})^T\boldsymbol{y_i}$, where $\boldsymbol{y_C}$ is the mean feature vector over the choice set. If we make Gaussian assumptions on the distribution of features and on choice set assignment, and if chooser utilities are inner products of chooser and item vectors,

then the LCL is a mean-field approximation to the induced choice system. In particular, we assume choice sets are generated to be similar to items the chooser a would like (as in a recommender system) by sampling items from a Gaussian with mean $\boldsymbol{x_a}$.

Theorem 7. *Let items and choosers both be represented by vectors in $\mathbb{R}^d$. Suppose chooser covariates $\boldsymbol{x_a}$ are distributed in the population according to a multivariate Gaussian $\mathcal{N}(\boldsymbol{\mu}, \Sigma_0)$, and a choice set for chooser a is constructed by sampling k items from the multivariate Gaussian $\mathcal{N}(\boldsymbol{x_a}, \Sigma)$. Additionally, assume choosers have the utility function $u_i(a, C) = \boldsymbol{x_a^T y_i}$. Then the expected chooser given a choice set C, $\boldsymbol{x_a^*} = \mathrm{E}[\boldsymbol{x_a} \mid C]$, has LCL choice probabilities, with*

$$\boldsymbol{\theta} = \tfrac{1}{k}\Sigma(\Sigma_0 + \tfrac{1}{k}\Sigma)^{-1}\boldsymbol{\mu}, \quad A = \Sigma_0(\Sigma_0 + \tfrac{1}{k}\Sigma)^{-1}.$$

Thus, the LCL can either be thought of as a context effect model, or as an approximation to the choice system induced by recommender-style preferred item overrepresentation.

5.2 Counterfactuals for known choosers

To make counterfactual predictions without chooser covariates or insufficiently descriptive covariates (preventing us from applying IPW or regression), we develop a clustering method for the challenge of choice set confounding. Suppose a recommender system suggests two sets of movies to two users: {Romance A, Romance B} to a_1 and {Drama A, Drama B} to a_2. While we know nothing about a_1 or a_2, we might be inclined to think a_1 is likely to pick Romance A from {Romance A, Drama A}, while a_2 is likely to pick Drama A from the same choice set. Similar to the CDM derivation in the previous section, the choice set is a signal for chooser preferences. We can also apply collaborative filtering principles, with the distinction that instead of thinking that similar users like similar items, we assume similar choosers are shown similar choice sets. There is a limitation, though, as this approach only lets us make predictions for choosers who appear in the original dataset. While there are many ways of using information from choice set assignment, we highlight an approach for the case where we have corresponding types of choosers and items (e.g., "romance fans" for "romance movies").

Suppose that choosers are more likely to have an item in their choice set if it matches their type. Define the $m \times n$ matrix R, where $R_{ij} = 1$ if the ith choice set includes item j and $R_{ij} = 0$ otherwise. We can think of R as the upper right block of the adjacency matrix of a bipartite graph between choosers and items, in which an edge (a, i) means that i is in a's choice set. With fixed choice set inclusion probabilities for each type, clustering choosers into types based on their choice sets is then an instance of the bipartite stochastic block model (SBM) recovery problem [1, 27].

In Theorem 8, we apply a classic exact recovery result due to McSherry [34] to show how a choice system with discrete types can be deconfounded without access to chooser covariates (i.e., knowledge of type membership), but any bipartite SBM clustering algorithm could be used (see Abbe [1] for a survey of SBM results).

Theorem 8. *Suppose items and choosers are jointly split into k types. Let s be the smallest number of items or choosers of any type and let $n = |A| + |\mathcal{U}|$. Suppose that for each chooser $a \in A$, $i \in \mathcal{U}$ is included in a's choice set with probability p if a and i are of the same type and with probability q otherwise.*

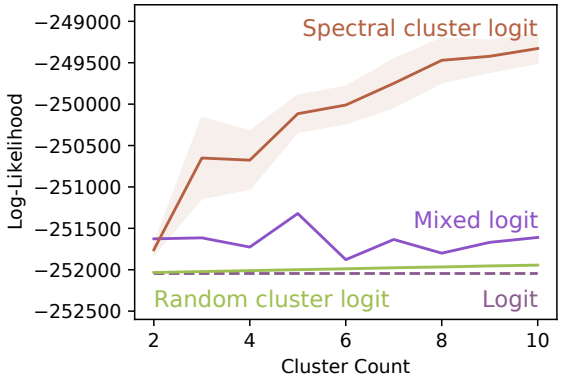

Figure 5: YOOCHOOSE log-likelihood comparison. Spectral and random cluster results are averaged over eight trials, with one standard deviation shaded.

There exists a constant C such that for large enough n, if

$$s(p-q)^2 > Ck\left(n/s + \log n/\delta\right), \tag{4}$$

then w.p. $1-\delta$, we can efficiently learn the type of every item and every chooser given a dataset $\mathcal{D}$ with one choice from each $a \in A$.

While McSherry's algorithm has strong theoretical guarantees, a more practical implementation is spectral co-clustering [14], which performs well for our purposes. Once we recover type memberships, we train separate models for each type of chooser and use the model for a chooser's type for deconfounded counterfactual predictions.

5.3 Empirical data without chooser covariates

We apply our spectral method to the YOOCHOOOSE online shopping dataset [7]. The dataset consists of all items clicked on in a session and an indicator of whether each item was purchased. We consider each purchase to be a choice from the set of all items viewed in the session. We group items by category (e.g., sports equipment) removing those with fewer than 100 purchases, leaving 29 categories.

We then perform spectral co-clustering [14] on the choice set matrix R with 2 to 10 chooser clusters and train a separate logit on each cluster. We ignore the item clusters. We compare against random clustering with the cluster sizes found by spectral clustering and mixed logit with the same number of components.

Spectral clustered logit describes the data much better than random clustering or even mixed logit (Figure 5). Note that the clusters are based only on choice set assignment, not choice behavior. In contrast, mixed logit bases its mixture components solely based on choices. The strong performance of spectral clustering indicates that choice sets are informative about preferences, and our use of this information is much easier than learning a mixture model.

6 DISCUSSION

Choice set confounding is widespread and can affect choice probability estimates, alter or introduce context effects, and lead to poor generalization. Existing models ignoring chooser covariates are particularly susceptible, but plugging in covariates is not a universal solution. We saw that covariates may be more informative about choice sets than preferences, making IPW more viable than regression. An important contribution is formalizing and demonstrating

choice set confounding, as it has significant implications for discrete choice modeling. For instance, initial research on the SF transportation data used extensive nested logit modeling to account for IIA violations [26], which we can manage with choice set confounding.

Our methods are a first step in addressing confounding. A challenge was learning choice set propensities for IPW. Simple logistic regression can work for binary treatments, but estimating exponentially many choice set propensities is difficult. In EXPEDIA, we learned a distribution over mean choice set feature vectors as an approximation. Other methods for learning set assignment probabilities would be valuable. Instrumental variables are another causal inference approach [18] that could be used in our setting, but identifying instruments for choice data is difficult. Alternatively, a matching approach [23] could compare pairs of similar choosers with different choice sets. Other directions for future investigation include rigorous methods of detecting choice set confounding, or verifying that it has been successfully accounted for, and of testing assumptions.

ACKNOWLEDGMENTS

This research was supported by ARO MURI, ARO Awards W911NF19-1-0057 and 73348-NS-YIP, NSF Award DMS-1830274, the Koret Foundation, and JP Morgan Chase & Co. We thank Spencer Peters for helpful discussions.

REFERENCES

[1] Emmanuel Abbe. 2017. Community detection and stochastic block models: recent developments. *JMLR* 18, 1 (2017).
[2] Arpit Agarwal, Prathamesh Patil, and Shivani Agarwal. 2018. Accelerated spectral ranking. In *ICML*.
[3] Greg M Allenby and Peter E Rossi. 1998. Marketing models of consumer heterogeneity. *J. Econometrics* 89, 1-2 (1998).
[4] Peter C Austin. 2011. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res* (2011).
[5] Heejung Bang and James M Robins. 2005. Doubly robust estimation in missing data and causal inference models. *Biometrics* 61, 4 (2005).
[6] Moshe Ben-Akiva and Bruno Boccara. 1995. Discrete choice models with latent choice sets. *Int. Journal of Research in Marketing* 12, 1 (1995).
[7] David Ben-Shimon et al. 2015. RecSys challenge 2015 and the YOOCHOOSE dataset. In *RecSys*.
[8] Austin R Benson, Ravi Kumar, and Andrew Tomkins. 2016. On the relevance of irrelevant alternatives. In *WWW*.
[9] Chandra R Bhat and Rachel Gossen. 2004. A mixed multinomial logit model analysis of weekend recreational episode type choice. *Transp Res Part B* (2004).
[10] Michel Bierlaire, Ricardo Hurtubia, and Gunnar Flötteröd. 2010. Analysis of implicit choice set generation using a constrained multinomial logit model. *Transportation Research Record* 2175, 1 (2010).
[11] Amanda Bower and Laura Balzano. 2020. Preference Modeling with Context-Dependent Salient Features. In *ICML*.
[12] David Brownstone et al. 1996. A transactions choice model for forecasting demand for alternative-fuel vehicles. *Res. Transp. Econ.* 4 (1996).
[13] Flavio Chierichetti, Ravi Kumar, and Andrew Tomkins. 2018. Learning a mixture of two multinomial logits. In *International Conference on Machine Learning*. PMLR, 961–969.
[14] Inderjit S Dhillon. 2001. Co-clustering documents and words using bipartite spectral graph partitioning. In *KDD*.
[15] Richard O Duda, Peter E Hart, and David G Stork. 2001. *Pattern Classification*.
[16] David A Freedman and Richard A Berk. 2008. Weighting regressions by propensity scores. *Evaluation Rev.* 32, 4 (2008).
[17] Michele Jonsson Funk et al. 2011. Doubly robust estimation of causal effects. *American Journal of Epidemiology* 173, 7 (2011), 761–767.
[18] Miguel A Hernán and James M Robins. 2006. Instruments for causal inference: an epidemiologist's dream? *Epidemiology* (2006).
[19] Keisuke Hirano, Guido W Imbens, and Geert Ridder. 2003. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71, 4 (2003).
[20] Saul D Hoffman and Greg J Duncan. 1988. Multinomial and conditional logit discrete-choice models in demography. *Demography* 25, 3 (1988), 415–427.
[21] Joel Huber, John W Payne, and Christopher Puto. 1982. Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. *Journal of Consumer Research* 9, 1 (1982).
[22] Samuel Ieong, Nina Mishra, and Or Sheffet. 2012. Predicting preference flips in commerce search. In *ICML*.
[23] Guido W Imbens. 2004. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Econ. and Stat.* 86, 1 (2004).
[24] Guido W Imbens and Donald B Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.
[25] Kaggle. 2013. Personalize Expedia Hotel Searches — ICDM 2013. https://www.kaggle.com/c/expedia-personalized-sort.
[26] Frank S Koppelman and Chandra Bhat. 2006. A self instructing course in mode choice modeling: multinomial and nested logit models. (2006).
[27] Daniel B Larremore, Aaron Clauset, and Abigail Z Jacobs. 2014. Efficiently inferring community structure in bipartite networks. *Phys. Rev. E* 90, 1 (2014).
[28] Charles F Manski. 1977. The structure of random utility models. *Theory and Decision* 8, 3 (1977).
[29] Charles F Manski and Steven R Lerman. 1977. The estimation of choice probabilities from choice based samples. *Econometrica* (1977).
[30] Benjamin M Marlin, Richard S Zemel, Sam Roweis, and Malcolm Slaney. 2007. Collaborative filtering and the missing at random assumption. In *UAI*.
[31] Daniel McFadden. 1974. Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics* (1974).
[32] Daniel McFadden and Kenneth Train. 2000. Mixed MNL models for discrete response. *Journal of Applied Econometrics* 15, 5 (2000).
[33] Daniel McFadden, William B Tye, and Kenneth Train. 1977. An Application of Diagnostic Tests for the Independence From Irrelevant Alternatives Property of the Multinomial Logit Model. *Transp Res Rec* (1977).
[34] Frank McSherry. 2001. Spectral partitioning of random graphs. In *FOCS*.
[35] Karlson Pfannschmidt, Pritha Gupta, and Eyke Hüllermeier. 2019. Learning choice functions: Concepts and architectures. *arXiv* (2019).
[36] Stephen Ragain and Johan Ugander. 2016. Pairwise choice Markov chains. In *NeurIPS*.
[37] Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 1 (1983).
[38] Nir Rosenfeld, Kojin Oshiba, and Yaron Singer. 2020. Predicting Choice with Set-Dependent Aggregation. In *ICML*.
[39] Donald B Rubin. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Edu. Psych.* 66, 5 (1974), 688.
[40] Donald B Rubin. 1977. Assignment to treatment group on the basis of a covariate. *J. Educational Stat.* 2, 1 (1977).
[41] Donald B Rubin. 2005. Causal inference using potential outcomes: Design, modeling, decisions. *J. Amer. Statist. Assoc.* 100, 469 (2005), 322–331.
[42] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. 2016. Recommendations as treatments: debiasing learning and evaluation. In *ICML*.
[43] Arjun Seshadri, Alex Peysakhovich, and Johan Ugander. 2019. Discovering Context Effects from Raw Choice Data. In *ICML*.
[44] Arjun Seshadri, Stephen Ragain, and Johan Ugander. 2020. Learning Rich Rankings. *NeurIPS* (2020).
[45] Itamar Simonson. 1989. Choice based on reasons: The case of attraction and compromise effects. *J. Consumer Research* 16, 2 (1989).
[46] Itamar Simonson and Amos Tversky. 1992. Choice in context: Tradeoff contrast and extremeness aversion. *Journal of Marketing Research* 29, 3 (1992), 281–295.
[47] Leslie S Stratton, Dennis M O'Toole, and James N Wetzel. 2008. A multinomial logit model of college stopout and dropout behavior. *Econ. Edu. Rev.* 27, 3 (2008).
[48] Elie Tamer. 2019. The ET Interview: Professor Charles Manski. *Econometric Theory* 35, 2 (2019).
[49] Kiran Tomlinson and Austin Benson. 2020. Choice Set Optimization Under Discrete Choice Models of Group Decisions. In *ICML*.
[50] Kiran Tomlinson and Austin R Benson. 2021. Learning Interpretable Feature Context Effects in Discrete Choice. In *KDD*.
[51] Kenneth E Train. 2009. *Discrete choice methods with simulation*.
[52] A Tversky. 1972. Elimination by aspects: A theory of choice. *Psych. Rev.* (1972).
[53] Xiaojie Wang, Rui Zhang, Yu Sun, and Jianzhong Qi. 2019. Doubly robust joint learning for recommendation on data missing not at random. In *ICML*.
[54] Yixin Wang, Dawen Liang, Laurent Charlin, and David M Blei. 2020. Causal Inference for Recommender Systems. In *RecSys*.
[55] Larry Wasserman. 2013. *All of Statistics: A Concise Course in Statistical Inference*. Springer Science & Business Media.
[56] Chieh-Hua Wen and Frank S Koppelman. 2001. The generalized nested logit model. *Transportation Research Part B: Methodological* 35, 7 (2001).
[57] Shuang-Hong Yang et al. 2011. Collaborative competitive filtering: learning recommender using context of user choice. In *SIGIR*.

A PROOFS

Proof of Observation 1. Conditioning over choosers yields $\Pr(i \mid C) = \sum_a \Pr(i \mid a, C) \Pr(a \mid C)$. Meanwhile, $\mathrm{E}_a[\Pr(i \mid a, C)] = \sum_a \Pr(i \mid a, C) \Pr(a)$. These are equal if condition (1) holds (since independence also implies $\Pr(a) = \Pr(a \mid C)$). If condition (2) holds, then we directly have $\mathrm{E}_a[\Pr(i \mid a, C)] = \Pr(i \mid C)$. See Example 1 for an instance where this equality fails when neither (1) nor (2) hold. □

Proof of Theorem 2. First, notice that any universal logit choice probabilities aggregated over a population can be expressed by set-dependent utilities $u_i^*(C)$ for each $C \subseteq \mathcal{U}, i \in C$. For every choice set $C \subseteq \mathcal{U}$, construct a chooser a_C with fixed utilities $u_i(a_C) = u_i^*(C)$. Let $\Pr(C \mid a_C) = 1$ and $\Pr(C' \mid a_C) = 0$ for all other $C' \neq C$. The choice probabilities of this mixture with chooser-dependent sets is the same as in the original system, and the mixture has finitely many $(2^{|\mathcal{U}|} - 1)$ components, one for each nonempty choice set C. □

Proof of Theorem 4. We need (1) in order for IPW (and therefore $\tilde{D}$) to be well-defined. Fix i and C. Consider the coefficient of $\log \Pr_\theta(i \mid C)$ in $\ell(\theta; \mathcal{D}^*)$. In expectation, this term appears $|\mathcal{D}| \Pr(i, C)$ times. Expanding this:

$$\begin{aligned}|\mathcal{D}| \Pr(i, C) &= |\mathcal{D}| \sum_{a \in A} \Pr(i, C \mid a) \Pr(a) \\ &= |\mathcal{D}| \sum_{a \in A} \Pr(i \mid C, a) \Pr(C \mid a) \Pr(a) \\ &= \frac{|\mathcal{D}|}{|C_{\mathcal{D}}|} \sum_{a \in A} \Pr(i \mid C, a) \Pr(a),\end{aligned}$$

where the last step follows from $\mathcal{D}^*$ having uniformly random choice sets. Now consider the coefficient of $\log \Pr_\theta(i \mid C)$ in $\ell(\theta; \tilde{\mathcal{D}})$. By IPW, this coefficient is

$$\sum_{\substack{(a,C',i) \in \mathcal{D} \\ i'=i, C'=C}} \frac{1}{|C_{\mathcal{D}}| \Pr(C \mid \boldsymbol{x_a})} = \frac{1}{|C_{\mathcal{D}}|} \sum_{a \in A} \sum_{\substack{(a',C',i') \in \mathcal{D} \\ a'=a, C'=C, i'=i}} \frac{1}{\Pr(C \mid \boldsymbol{x_a})}.$$

In expectation, the sample (a, C, i) occurs $|\mathcal{D}| \Pr(a, C, i)$ times. Additionally, by choice set ignorability, $\Pr(C \mid \boldsymbol{x_a}) = \Pr(C \mid a)$. We thus have that the expected coefficient is

$$\begin{aligned}&\frac{1}{|C_{\mathcal{D}}|} \sum_{a \in A} \frac{1}{\Pr(C \mid \boldsymbol{x_a})} |\mathcal{D}| \Pr(a, C, i) \\ &= \frac{|\mathcal{D}|}{|C_{\mathcal{D}}|} \sum_{a \in A} \frac{1}{\Pr(C \mid a)} \Pr(a) \Pr(C \mid a) \Pr(i \mid C, a) \\ &= \frac{|\mathcal{D}|}{|C_{\mathcal{D}}|} \sum_{a \in A} \Pr(a) \Pr(i \mid C, a),\end{aligned}$$

which matches the coefficient in $\ell(\theta; \mathcal{D}^*)$. Since the expected coefficients agree for all i and C, we then have the equality. □

Proof of Theorem 6. By the consistency of the MLE, as $|\mathcal{D}| \to \infty$, parameter estimates for a correctly specified choice model converge to the true parameters. Thus, estimated choice probabilities also converge:

$$\begin{aligned}\lim_{|\mathcal{D}| \to \infty} \hat{\Pr}(i \mid \boldsymbol{x_a}, C) &= \Pr(i \mid \boldsymbol{x_a}, C) \\ &= \Pr(i \mid a, \boldsymbol{x_a}, C) \\ &\quad \text{(by preference ignorability)} \\ &= \Pr(i \mid a, C).\end{aligned}$$

□

Proof of Theorem 7. Observing the choice set gives us a noisy measurement of $\boldsymbol{x_a}$, which we can adjust using our knowledge of the distribution of $\boldsymbol{x_a}$. The posterior of a Gaussian with a Gaussian prior is also Gaussian—in particular, $\boldsymbol{x_a} \mid C$ is Gaussian, with mean

$$\mathrm{E}[\boldsymbol{x_a} \mid C] = \Sigma_0 \left(\Sigma_0 + \frac{1}{k}\Sigma\right)^{-1} \boldsymbol{y_C} + \frac{1}{k}\Sigma \left(\Sigma_0 + \frac{1}{k}\Sigma\right)^{-1} \boldsymbol{\mu}$$

[15, Section 3.4.3]. Thus, the expected chooser $\boldsymbol{x_a^*}$ has utilities

$$u_i(a^*, C) = \left[\Sigma_0 \left(\Sigma_0 + \frac{1}{k}\Sigma\right)^{-1} \boldsymbol{y_C} + \frac{1}{k}\Sigma \left(\Sigma_0 + \frac{1}{k}\Sigma\right)^{-1} \boldsymbol{\mu}\right]^T \boldsymbol{y_i}.$$

This is exactly an LCL with $\boldsymbol{\theta}$ and A as claimed. □

Proof of Theorem 8. Consider the bipartite graph whose left nodes are choosers and whose right nodes are items, each split into blocks according to their type. The choice set assignment process above defines a bipartite SBM on this graph with intra-type probabilities p and inter-type probabilities q (between chooser nodes and item nodes). Recovering types from choice sets can then be viewed as an instance of the planted partition problem [34].

We can thus directly[2] apply Theorem 4 of McSherry [34] to achieve the desired result given Equation (4), with the caveat that algorithm is random and succeeds with probability $1/k$.

Repeating the algorithm ck times achieves failure probability $(1 - \frac{1}{k})^{ck} \leq 1/e^c$, which is smaller than δ if $c > \log(1/\delta)$. We can thus make δ smaller by a factor of 2 (absorbing this into the constant C in eq. (4)) and we are left with the guarantee as stated, only increasing the running time by a factor $k \log(1/\delta)$. □

B AFFINE-MEAN GAUSSIAN CHOICE SET MODEL

For estimating choice set propensities in expedia, we model the distribution of mean choice set features using an affine-mean Gaussian. Here, we show how this model can be easily estimated from data.

Proposition 9. *Given a dataset $\mathcal{D}$, the model $\boldsymbol{y_C} \sim \mathcal{N}(W\boldsymbol{x_a} + \boldsymbol{z}, \Sigma)$ is identifiable iff there are $m+1$ choosers in $\mathcal{D}$ with affinely independent covariates. If the model is identified, the maximum likelihood parameters $W^*, \boldsymbol{z}^*$ are the solution to the least-squares problem*

$$(W^*, \boldsymbol{z}^*) = \operatorname*{arg\,min}_{\substack{W \in \mathbb{R}^{n \times m} \\ \boldsymbol{z} \in \mathbb{R}^n}} \sum_{(a,C) \in \mathcal{D}} \|\boldsymbol{y_C} - (W\boldsymbol{x_a} + \boldsymbol{z})\|_2^2, \tag{5}$$

[2] Notice that $s(p-q)^2$ is a lower bound on the squared 2-norm of the columns of the SBM edge probability matrix required by [34, Theorem 4]. Additionally, we use the crude variance upper bound $\sigma^2 = 1$ for simplicity.

which have the closed form:

$$W^* = \left[\sum_{(a,C)\in\mathcal{D}} (\boldsymbol{y}_C - \boldsymbol{y}_\mathcal{D})\boldsymbol{x}_a^T\right]\left[\sum_{(a,C)\in\mathcal{D}} (\boldsymbol{x}_a - \boldsymbol{x}_\mathcal{D})\boldsymbol{x}_a^T\right]^{-1} \quad (6)$$

$$\boldsymbol{z}^* = \boldsymbol{y}_\mathcal{D} - W^*\boldsymbol{x}_\mathcal{D}, \quad (7)$$

where $\boldsymbol{x}_\mathcal{D} = 1/|\mathcal{D}| \sum_{(a,C)\in\mathcal{D}} \boldsymbol{x}_a$ *and* $\boldsymbol{y}_\mathcal{D} = 1/|\mathcal{D}| \sum_{(a,C)\in\mathcal{D}} \boldsymbol{y}_C$.

Additionally, the maximum likelihood covariance matrix is the sample covariance:

$$\Sigma^* = \frac{1}{|\mathcal{D}|} \sum_{(a,C)\in\mathcal{D}} (\boldsymbol{y}_C - W^*\boldsymbol{x}_a - \boldsymbol{z}^*)(\boldsymbol{y}_C - W^*\boldsymbol{x}_a - \boldsymbol{z}^*)^T. \quad (8)$$

Proof sketch. This can be derived following the same steps as the standard Gaussian MLE proof (with a bit of extra matrix calculus): (1) take partial derivatives of the log-likelihood with respect to W and $\boldsymbol{z}$, (2) set them to zero, (3) solve for $\boldsymbol{z}$, (4) plug this in to solve for W, (5) do the same to solve for Σ in its partial derivative. This works since the log-likelihood is still convex after adding in the affine trasformation. We omit the details as they are tedious and unenlightening. □

C EXPERIMENT DETAILS

We implemented all choice models with PyTorch and (except mixed logit) train them using Rprop with no minibatching to optimize the log-likelihood for 500 epochs or until convergence (squared gradient norm $< 10^{-8}$), whichever comes first. We use ℓ_2 regularization with coefficient $\lambda = 10^{-4}$ for all models to ensure identifiability. For mixed logit, we use an expectation-maximization (EM) algorithm [51] with a one hour timeout. Our code, results, and links to data are available at

https://github.com/tomlinsonk/choice-set-confounding.

Table 5: Regularity violations in SF-WORK and SF-SHOP, impossible under mixed logit. Including additional item(s) appears to increase the probability that DA or DA/SR is chosen. The differences are significant according to Fisher's exact test (SF-WORK: $p = 6.5 \times 10^{-9}$, SF-SHOP: $p = 0.005$).

SF-WORK **Choice set (C)**	Pr(DA | C)	N
{DA, SR 2, SR 3+, Transit}	0.72	1661
{DA, SR 2, SR 3+, Transit, Bike}	0.83	829

SF-SHOP **Choice set (C)**	Pr(DA/SR | C)	N
{DA, DA/SR, SR 2, SR 3+, SR 2/SR 3+, Transit}	0.17	534
{DA, DA/SR, SR 2, SR 3+, SR 2/SR 3+, Transit, Bike, Walk}	0.23	1315

DA: drive alone. SR: shared ride, number indicates car occupancy.
Slashes indicate different mode used for outbound and inbound trips.