European Journal of Operational Research xxx (xxxx) xxx

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/eor

Decision support

Robust binary and multinomial logit models for classification with data uncertainties

Baichuan Mo [a,b], Yunhan Zheng [b,c,*], Xiaotong Guo [b], Ruoyun Ma [d], Jinhua Zhao [e]

[a] Department of Civil Engineering, Tsinghua University, Beijing, 100084, China
[b] Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, 02139, United States of America
[c] Singapore–MIT Alliance for Research and Technology Centre (SMART), Singapore
[d] Department of Management Science and Engineering, Stanford University, Stanford, CA 94305, United States of America
[e] Department of Urban Studies and Planning, Massachusetts Institute of Technology, Cambridge, MA 20139, United States of America

ARTICLE INFO

Keywords:
Discrete choice models for classification
Robust optimization
Data uncertainty

ABSTRACT

Binary logit (BNL) and multinomial logit (MNL) models are the two most widely used discrete choice models for travel behavior modeling and prediction. However, in many scenarios, the collected data for those models are subject to measurement errors. Previous studies on measurement errors mostly focus on "better estimating model parameters" with training data. In this study, we focus on using BNL and MNL for classification problems, that is, to "better predict the behavior of new samples" when measurement errors occur in testing data. To this end, we propose a robust BNL and MNL framework that is able to account for data uncertainties in both features and labels. The models are based on robust optimization theory that minimizes the worst-case loss over a set of uncertainty data scenarios. Specifically, for feature uncertainties, we assume that the ℓ_p-norm of the measurement errors in features is smaller than a pre-established threshold. We model label uncertainties by limiting the number of mislabeled choices to at most Γ. Based on these assumptions, we derive a tractable robust counterpart. The derived robust-feature BNL and the robust-label MNL models are exact. However, the formulation for the robust-feature MNL model is an approximation of the exact robust optimization problem. An upper bound of the approximation gap is provided. We prove that the robust estimators are inconsistent but with a higher trace of the Fisher information matrix. They are preferred when out-of-sample data has errors due to the shrunk scale of the estimated parameters. The proposed models are validated in a binary choice data set and a multinomial choice data set, respectively. Results show that the robust models (both features and labels) can outperform the conventional BNL and MNL models in prediction accuracy and log-likelihood. We show that the robustness works like "regularization" and thus has better generalizability.

1. Introduction

Binary logit (BNL) and multinomial logit (MNL) models are the two most widely used discrete choice models (DCMs) (Ben-Akiva & Lerman, 1985). They are widely used to describe, explain, and predict choices between two or more discrete alternatives, such as entering or not entering the labor market, or choosing between modes of transport. The models specify the probability that a person chooses a particular alternative, with the probability expressed as a function of observed variables that relate to the alternatives and the person.

1.1. Preliminaries

BNL and MNL models can be derived from random utility theory. Particularly, let $\boldsymbol{x}_{n,i}$ be the vector of observed factors for person n with respective alternative i, defined as

$$\boldsymbol{x}_{n,i} := \left(x_{n,i}^{(k)}\right)_{k\in\mathcal{K}_i}, \quad \forall i \in C_n, \tag{1}$$

where $x_{n,i}^{(k)}$ is the kth element of $\boldsymbol{x}_{n,i}$. C_n is the set of available alternatives for person n (for example, public transit, walking, car). The set of all alternatives is $C := \cup_{n\in\mathcal{N}} C_n$. $\mathcal{K}_i$ is the set of features for alternative

* Corresponding author at: Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, 02139, United States of America.
E-mail address: yunhan@mit.edu (Y. Zheng).

https://doi.org/10.1016/j.ejor.2025.05.013
Received 16 June 2024; Accepted 6 May 2025
Available online 22 May 2025

i. The set of all features is $\mathcal{K} := \cup_{i\in C}\mathcal{K}_i$. For example, features may include travel time, travel cost, and a person's income. C and $\mathcal{K}$ are both finite.

Let $U_{n,i}$ be the utility that person n obtains from choosing alternative i. The person's utility depends on many factors, some of which we observe and some not. Hence, $U_{n,i}$ can be decomposed into a part that depends on observed variables and another part with unobserved factors. In a linear form, the utility is expressed as

$$U_{n,i} = \boldsymbol{\beta}_i^\top \boldsymbol{x}_{n,i} + \epsilon_{n,i}, \tag{2}$$

where $\boldsymbol{\beta}_i \in \mathbb{R}^{|\mathcal{K}_i|}$ is the corresponding vector of coefficients to be estimated. It represents the contribution of a feature (e.g., cost, travel time) to the total utility. $\epsilon_{n,i}$ is a random variable that captures the impact of all unobserved factors that affect the person's choice and zero-mean measurement errors. $U_{n,i}$ is also a random variable.

The choice of the person is designated by dummy variables $y_{n,i}$. $\boldsymbol{y_n} = (y_{n,i})_{i\in C_n}$ is the associated vector. $y_{n,i} = 1$ indicating person n choosing alternative i and $y_{n,i} = 0$ otherwise. Based on the utility maximization assumption (i.e., people will choose the highest utility alternative), the probability of $y_{n,i} = 1$ can be modeled as

$$\begin{aligned} P_{n,i} := \mathbb{P}(y_{n,i} = 1) &= \mathbb{P}(U_{n,i} \geq U_{n,j},\ \forall j \in C_n \setminus \{i\}) \\ &= \mathbb{P}\big(\epsilon_{n,i} - \epsilon_{n,j} \geq (\boldsymbol{\beta}_j^\top \boldsymbol{x}_{n,j} - \boldsymbol{\beta}_i^\top \boldsymbol{x}_{n,i}),\ \forall j \in C_n \setminus \{i\}\big). \end{aligned} \tag{3}$$

In BNL and MNL models, $\epsilon_{n,i}$ is assumed to be independent and identically Gumbel distributed. Then $\epsilon_{n,i} - \epsilon_{n,j}$ will follow the logistic distribution with cumulative density function (CDF) $F_{\Delta\epsilon_{i,j}}(z) = \frac{1}{1+e^{-z}}$. When $|C_n| = 2$ for all people, this formulation represents the BNL model. Eq. (3) has a closed-form formulation (Train, 2009):

$$\begin{aligned} P_{n,i} &= F_{\Delta\epsilon_{i,j}}(\boldsymbol{\beta}_i^\top \boldsymbol{x}_{n,i} - \boldsymbol{\beta}_j^\top \boldsymbol{x}_{n,j}) \\ &= \frac{1}{1 + e^{\boldsymbol{\beta}_j^\top \boldsymbol{x}_{n,j} - \boldsymbol{\beta}_i^\top \boldsymbol{x}_{n,i}}} = \frac{e^{\boldsymbol{\beta}_i^\top \boldsymbol{x}_{n,i}}}{e^{\boldsymbol{\beta}_i^\top \boldsymbol{x}_{n,i}} + e^{\boldsymbol{\beta}_j^\top \boldsymbol{x}_{n,j}}}, \quad \forall i, j \in C_n, \end{aligned} \tag{4}$$

which is equivalent to the canonical logistic regression if $\boldsymbol{x}_{n,j} = \boldsymbol{x}_{n,i}$.

For $|C_n| > 3$, the formulation represents the MNL model and Eq. (3) is equivalent to $\mathbb{P}(U_{n,i} - \max_{j\in C_n\setminus\{i\}} U_{n,j} \geq 0)$. Given the property of the Gumbel distribution, $U_n^* := \max_{j\in C_n\setminus\{i\}} U_{n,j}$ is also Gumbel distributed with the location parameter equal to $\sum_{j\in C_n\setminus\{i\}} \boldsymbol{\beta}_j^\top \boldsymbol{x}_{n,j}$. Then the CDF of $(U_{n,i} - \max_{j\in C_n\setminus\{i\}} U_{n,j})$ will be logistic distributed (Ben-Akiva & Lerman, 1985) with CDF $F_{\Delta U_{i,*}}(z) = \frac{1}{1+\exp(-z+\boldsymbol{\beta}_i^\top \boldsymbol{x}_{n,i} - \sum_{j\in C_n\setminus\{i\}} \boldsymbol{\beta}_j^\top \boldsymbol{x}_{n,j})}$. Therefore, the probability for the MNL model is

$$P_{n,i} = \mathbb{P}(U_{n,i} - \max_{j\in C_n\setminus\{i\}} U_{n,j} \geq 0) = 1 - F_{\Delta U_{i,*}}(0) = \frac{e^{\boldsymbol{\beta}_i^\top \boldsymbol{x}_{n,i}}}{\sum_{j\in C_n} e^{\boldsymbol{\beta}_j^\top \boldsymbol{x}_{n,j}}}, \tag{5}$$

Maximum likelihood estimation (MLE) is usually used to estimate the BNL and MNL models to get the parameter. Define $\boldsymbol{\beta} := (\boldsymbol{\beta}_j)_{j\in\mathcal{C}}$. The MLE can be expressed as the following optimization problem:

$$\max_{\boldsymbol{\beta}} \sum_{n\in\mathcal{N}} \sum_{i\in C_n} y_{n,i} \cdot \log(P_{n,i}(\boldsymbol{\beta})) = \max_{\boldsymbol{\beta}} \sum_{n\in\mathcal{N}} \sum_{i\in C_n} y_{n,i} \cdot \log\left(\frac{\exp(\boldsymbol{\beta}_i^\top \boldsymbol{x}_{n,i})}{\sum_{j\in C_n} \exp(\boldsymbol{\beta}_j^\top \boldsymbol{x}_{n,j})}\right), \tag{6}$$

where $\mathcal{N}$ is the set of all persons.

1.2. Uncertainties in data

BNL and MNL models are usually used as classifiers to predict an individual's behavior based on the estimated parameters from real-world data (such as surveys). However, in many scenarios, the collected data are subject to uncertainties (such as erroneous responses, dictation errors, etc.), which are known as measurement errors (Hausman, 2001). Beyond measurement errors, there are other sources of data uncertainty in transportation. For instance, Li and Xu (2023) identify three key factors: (a) perceived data used to construct models can be perturbed due to errors in measurement or recording; (b) while nominal distributions are used for random parameters in transportation networks, actual distributions may deviate, such as when roads unexpectedly close or maintenance work extends longer than planned; and (c) future validation data may differ from past data on which the model was built. Additionally, Mo et al. (2022) highlighted other sources of data uncertainty in traffic state estimation, including inherent stochasticity in driving behavior and random initial or boundary conditions. This research specifically focuses on measurement error issues within the context of DCMs.

Data errors may lead to biased or inconsistent estimates of model parameters, deteriorating the model's predictive power. Note that although the error term $\epsilon_{n,i}$ in Eq. (2) is assumed to capture some zero-mean measurement errors, an important assumption for the error term is that $\epsilon_{n,i}$ is uncorrelated with observed factors $\boldsymbol{x}_n$. However, in reality, measurement errors are often caused by misreporting some specific factors (like travel time, income, etc.), which are inevitably related to $\boldsymbol{x}_n$. Therefore, relying solely on $\epsilon_{n,i}$ to capture measurement errors will break the independence assumption and cause endogeneity.

Errors can happen in features (i.e., $\boldsymbol{x}_n$, left-hand side variables) and labels (i.e., $\boldsymbol{y}_n$, right-hand side variables), which require different ways to address. The typical way to deal with Measurement errors in the literature is instrumental variables (Hausman, 2001). The instrumental variables are assumed to be correlated with the "true value" of the mismeasured variables but uncorrelated with error terms.

Previous literature on measurement errors usually focuses on "better estimating model parameters" with training data. However, in the real world when using the trained (or estimated) DCMs to **predict** new users (or samples) behavior (i.e., a classification problem), measurement errors also exist in the testing data set. This implies that an "unbiased estimation" after correcting bias may end up performing worse in the travel behavior prediction task. Our paper will focus on this task, which differentiates us from previous econometrics studies, as shown in Table 1. Specifically, our study assumes uncertainties in the testing data set and aims to improve prediction accuracy under uncertainties. While previous studies assume uncertainties in the training data set and focus on estimating unbiased model parameters with the training data. This makes our paper closely related to developing robust classifiers (Bertsimas et al., 2019).

1.3. Organization and contributions

In this paper, we propose a robust discrete choice model framework that is able to account for data uncertainties in both features and labels. The objective is to provide a more accurate prediction of an individual's behavior for new samples (i.e., testing data set) as a classification task when there existing data errors. The model is based on a robust optimization framework that minimizes the worst-case loss over a set of uncertainty data scenarios. Specifically, for feature uncertainties, we assume that the ℓ_p-norm of the measurement errors in features is smaller than a pre-established threshold. We model label uncertainties by limiting the number of mislabeled choices to at most Γ. Based on these assumptions, we derive a tractable robust counterpart for robust-feature and robust-label DCM models. The derived robust-feature BNL and the robust-label MNL models are exact. However, the formulation for the robust-feature MNL model is an approximation of the exact robust optimization problem. The proposed models are validated in a binary choice data set and a multinomial choice data set, respectively. Results show that the robust models (both features and labels) can outperform the conventional BNL and MNL models in prediction accuracy and log-likelihood. We show that the robustness works like "regularization" and thus has better generalizability.

The main contribution of the paper is as follows:

- We drive the closed-form robust counterparts for the robust BNL and MNL models. Specifically, the formulations for robust-feature BNL robust-label MNL models are exact (i.e., not an approximation). For robust MNL, due to the difficulties in maximizing

Table 1
Comparison of this study and literature.

	Task	Performance	Data uncertainties
Literature	Estimate unbiased parameters	Interpretability	Training data
This study	Predict new samples	Prediction accuracy	Training and testing data

a convex function, we use Jensen's inequality to approximate the original objective function, yielding a lower bound of the original maximization problem. We prove that the gap of the approximation will not be extreme.

- We prove that the proposed robust-feature and robust-label estimators are inconsistent, but they have a higher trace of the Fisher information matrix compared to the MLE estimator (usually imply lower variance). When predicting new samples with data errors, robust estimators are preferred due to the shrinking scale of the estimated parameters.
- We explain the good performance of robust DCM models from the aspects of both theories and experiments. We show that robust models tend to shrink the scale of the estimated parameters, which lowers the expected prediction errors when testing data has perturbations. Experiment results also validate the theories.

Our work is connected to the robust classification methods proposed by Bertsimas et al. (2019), where robust optimization and logistic regression are integrated to construct classifiers that are capable of handling uncertainties in both features and labels. We draw inspiration from these approaches in terms of constructing suitable uncertainty sets and deriving robust counterparts in robust classifications. However, our work goes beyond their exclusive focus on binary classification. Instead, we contribute by developing robust DCMs for more than two categories and providing a theoretical analysis of the statistical properties of the robust estimators. For more than categories, we show that no exact robust counterpart exists, and an outer approximation is needed. We also analyze the quality of the approximation and prove that the gap of the approximation will not be extreme.

The remainder of this paper is organized as follows. The literature review is presented in Section 2. In Section 3, we describe formulations and derivations of the robust-feature DCMs for binary and multinomial cases. In Section 4, we elaborate on the robust-label DCMs that are suitable for both binary and multinomial cases. Section 5 discusses the theoretical statistical properties and the out-of-sample prediction performance for the robust estimators. We apply the proposed framework to two different data sets as case studies in Section 6. Conclusions and discussions are presented in Section 7.

2. Literature review

2.1. Measurement errors in DCM

Transportation planning and policy analysis heavily rely on travel survey data, which includes information about activity patterns, travel behaviors, and comprehensive socio-demographic profiles of the surveyed populations. However, it is crucial to acknowledge that the presence of measurement errors in survey data is not uncommon. These errors can affect various travel-related variables, including mode choice, trip duration, and travel costs, as well as socio-demographic factors such as income (Paleti & Balan, 2019). These errors present a significant challenge when utilizing DCM for the analysis of travel surveys. For instance, a Monte Carlo simulation conducted by Hausman et al. (1998) revealed that even a small rate of outcome misclassifications (e.g., 2%) can result in DCM estimates that exhibit biases ranging from 15% to 25% when the outcome is binary.

The origins of measurement errors in household surveys are multifaceted and can be attributed to several distinct sources. First, respondents may consciously choose to provide misleading information due to various motivations, including the desire to conceal certain details or to offer socially acceptable responses (Kreuter et al., 2008). For instance, research has revealed that self-employed individuals, in particular, may deliberately underreport their income by as much as 25% when participating in household surveys (Hurst et al., 2014). Second, measurement errors can stem from inadvertent misreporting by respondents. This occurs when individuals encounter difficulties in comprehending survey questions, struggle to recollect specific details from memory, or employ inappropriate decision-making heuristics (Campanelli et al., 1991). Furthermore, the precision of survey data can be intricately connected to the particular survey methods and tools utilized for data collection. For instance, individuals engaged in surveys conducted through Computer Assisted Telephone Interviewing (CATI) methodologies have exhibited systematic tendencies to underreport travel, underestimate travel distances, and overstate travel times, in comparison to surveys that leverage GPS-based tracking technology (Stopher et al., 2007).

DCM has been widely used for both understanding people's travel decisions and conducting activity-travel planning (Ben-Akiva & Bierlaire, 1999; Bowman & Ben-Akiva, 2001). Consequently, the presence of measurement errors within the data can potentially distort policy decisions that rely on estimation outcomes derived from compromised data. For instance, when essential travel decision factors like travel cost and time are inaccurately measured, it can yield erroneous estimates of their influence on travelers' preferences. Likewise, if mode choice is subject to mismeasurement, it may result in investments in transportation infrastructure that do not align with the genuine preferences of travelers, potentially leading to suboptimal resource allocation and inefficient utilization of public funds. Therefore, it is imperative to develop methods that account for data mismeasurements within traditional DCM.

Data perturbation and uncertainty are not limited to measurement errors in DCM but extend to other fields of transportation modeling. For instance, in the field of traffic modeling, Li and Xu (2023) identify several potential sources of data uncertainty, such as measurement and recording errors, deviations between nominal and actual distributions of random parameters, and discrepancies between future validation data and the historical data used for model development. They demonstrate the robustness of their Modified Late Arrival Penalized User Equilibrium (MLAPUE) model in addressing these data perturbations. Shao et al. (2021) emphasize the critical impact of sensor measurement error, noting that magnetic field interference can compromise the accuracy of inductive loop detectors, while factors such as visibility, lighting, and weather conditions can significantly affect video detector efficiency—ultimately influencing traffic flow estimation. In addition to measurement error, traffic stochasticity is another source of uncertainty. Mo et al. (2022) underscore the inherent randomness in driving behaviors and propose a Physics-Informed Generative Adversarial Network (GAN) to effectively quantify uncertainty in traffic state estimation. Musunuru and Porter (2019) highlight the serious implications of data measurement errors in road safety modeling. These errors often stem from transportation agencies extrapolating short-term traffic counts over time and space, leading to biases in regression coefficient estimates and increased model dispersion. To address this, they propose robust measurement error correction techniques, including regression calibration and simulation extrapolations.

The widespread occurrence of data uncertainty in transportation modeling underscores its critical impact on the reliability and robustness of analytical results. While numerous studies have explored the implications of measurement errors and uncertainty in traffic modeling, the challenges posed by data uncertainty in DCM have received

relatively limited attention. This study addresses this gap by focusing specifically on the effects of measurement errors in DCM, thereby contributing to the advancement of robust modeling practices within the broader transportation research field.

2.2. Methods for addressing measurement errors

To address the issue of feature mismeasurement (or "uncertainty"), researchers have primarily employed two main coping strategies (Schennach, 2016). The first approach centers on the recovery of accurate, error-free values from data tainted by measurement errors. However, this method assumes prior knowledge of the measurement error distribution, which may not always align with real-world situations. For instance, some earlier studies employed Fourier transform algorithms to mitigate measurement errors while assuming the availability of known error distributions (Schennach, 2019; Wang & Wang, 2011).

The second strategy involves correcting measurement error biases by incorporating readily available auxiliary variables, which can include repeated measurements (Schennach, 2004), indicators (Ben-Moshe, 2014), or instrumental variables (Hu, 2008). The instrumental variable approach, in particular, has been widely adopted to mitigate mismeasurement problems in linear specifications (Hausman, 2001). In this approach, instrumental variables are carefully selected to serve as proxies for the imperfectly measured feature variables. They are chosen based on their lack of correlation with the measurement errors, allowing them to effectively separate the measurement errors from the estimation process for the dependent variable (Baiocchi et al., 2014). However, a significant challenge with this group of approaches is the difficulty of obtaining suitable auxiliary variables in many practical applications.

In the context of addressing label uncertainties within DCM, previous research has employed modified maximum likelihood estimators (Hausman, 2001; Hausman et al., 1998; Paleti & Balan, 2019). This approach involves the direct codification and estimation of the proportion of misclassified data using maximum likelihood estimation. Nevertheless, this method assumes that the proportion of misclassified data is fixed and can be applied to the out-of-sample data. In many scenarios, the extent of misclassification is not deterministic but falls within specific ranges.

To overcome the aforementioned limitations, we propose the application of robust optimization to handle feature and label uncertainties in DCM. Robust optimization offers a novel approach by accounting for data uncertainties within a predefined range. Unlike conventional econometric methods, robust optimization does not require prior knowledge of the error distribution or the collection of auxiliary data. Specifically, it assumes that uncertain parameters, such as measurement errors and the number of mislabeled choices, belong to an uncertainty set of possible outcomes. The optimization process is based on identifying and addressing the worst-case scenario within this uncertainty set (Bertsimas et al., 2010; Gorissen et al., 2015).

Robust optimization techniques have provided researchers with valuable tools to tackle problems involving data uncertainty across a wide array of domains, including transportation routing and scheduling (Guo et al., 2024; Shi et al., 2019; Sungur et al., 2008), path recommendation (Mo et al., 2023), healthcare resource allocation (Wang et al., 2019), and portfolio optimization (Fernandes et al., 2016). Notably, to the best of our knowledge, no existing studies have employed robust optimization to tackle the challenges of measurement errors in travel behavior modeling with DCMs.

The robust optimization is solved under the worst-case scenario to provide a conservative decision. It has been shown in many studies that though the actual testing case is not the worst case, the decisions made by robust optimization still work well in general compared to the nominal model. In this paper, we derive the estimated parameters assuming some "worst-case" pattern of data errors that deteriorates the likelihood function. In the testing data, though the actual error patterns are not "worst-case", the estimated parameters may still perform better than the nominal DCM models. The intuition behind this can be understood using the analogy of robustness and regularization. Robustness can be seen as a way to avoid the model from overfitting.

2.3. Robust optimization for classification problems

Robust optimization has emerged as a key methodology for tackling classification tasks in the presence of uncertainty. The theoretical groundwork for robust optimization was established by Ben-Tal and Nemirovski (Ben-Tal & Nemirovski, 1998, 1999), who developed practical reformulations for optimization problems involving uncertainty. These foundational concepts were later applied to classification, particularly through the extension of support vector machines (SVMs) to more robust formulations. For example, Xu et al. (2009) introduced a robust SVM approach that accounts for input uncertainty using ellipsoidal uncertainty sets, showing that such models naturally incorporate regularization and can enhance generalization performance.

Building on these initial ideas, later studies explored robustness in both linear and nonlinear classification frameworks. Shivaswamy et al. (2006) introduced robust kernel-based classifiers, while Lanckriet et al. (2002) proposed semidefinite programming formulations that enabled robustness in more complex, nonlinear decision boundaries. These early advancements laid the groundwork for viewing robust classification as a problem of optimizing performance under worst-case perturbations in the input space.

More recent developments have focused on distributionally robust optimization, which seeks to minimize classification loss under the most adverse distribution drawn from a specified ambiguity set. Key contributions in this area include Duchi et al. (2021), who utilized the Wasserstein distance to define neighborhoods around empirical distributions, and Duchi and Namkoong (2019), who introduced ambiguity sets based on chi-squared divergence to improve both generalization and fairness. Building on these foundations, Sinha et al. (2017) proposed scalable methods for training deep learning models within the DRO framework, emphasizing the critical role that the choice of divergence metric plays in shaping classifier performance.

Another influential strand of literature connects robust optimization with adversarial machine learning. Goodfellow et al. (2014) introduced adversarial examples, and revealed deep classifiers' vulnerabilities to small perturbations. Building on this, Madry et al. (2017) formulated adversarial training as a robust optimization problem, where the classifier is optimized to withstand worst-case perturbations within a norm-bounded set. TRADES (Zhang et al., 2019) refined this approach by introducing a trade-off between natural accuracy and robustness, grounded in a distributionally robust perspective.

Existing work on robust optimization for classification has predominantly focused on improving parameter estimation from training data with noisy features, often limited to binary classification settings. In contrast, we develop a unified framework that explicitly accounts for uncertainties in both features and labels, and extends beyond binary cases to include multinomial logit models. We derive closed-form robust counterparts for the BNL and MNL models—exact in some cases and accompanied by provable approximation bounds in others. Additionally, we analyze the statistical properties of the proposed estimators, showing that they yield higher Fisher information and exhibit improved out-of-sample generalization.

3. Robustness against uncertainties in features

In this section, we consider perturbations (i.e., measurement errors) $\boldsymbol{\Delta x}_n$ for person n in his/her features. Without loss of generality, let us assume all alternatives use full features (i.e., $\boldsymbol{x}_{n,i} = \boldsymbol{x}_{n,j} = \boldsymbol{x}_n$ for all $i, j \in C_n$). For any model specification, we can set corresponding parameters

to zero so as to filter out undesired features in an alternative. This is equivalent to defining a parameter domain $\boldsymbol{\beta}_i \in \mathcal{B}_i$, where:

$$\mathcal{B}_i = \{\boldsymbol{\beta}_i \in \mathbb{R}^{|\mathcal{K}|} : \beta_i^{(k)} = 0 \text{ if } k\text{-th feature not used in mode } i\}, \quad \forall i \in C. \tag{7}$$

The overall parameter domain is thus $\mathcal{B} = \cup_{i\in C}\mathcal{B}_i$. With uncertainty in features, we have

$$U_{n,i} = \boldsymbol{\beta}_i^\top(\boldsymbol{x}_n + \boldsymbol{\Delta x}_n) + \epsilon_{n,i}, \tag{8}$$

where $\boldsymbol{\Delta x}_n \in \mathcal{Z}_n(\rho_n)$ and $\mathcal{Z}_n(\rho_n) = \{\boldsymbol{\Delta x}_n : \|\boldsymbol{\Delta x}_n\|_p \le \rho_n\}$ is the uncertainty set (we consider an ℓ_p-norm uncertainty). ρ_n is a hyper-parameter that represents the largest error of the perturbations $\boldsymbol{\Delta x}_n$.

Assume the uncertainties are independent across individuals:

$$\mathcal{Z}(\boldsymbol{\rho}) = \prod_{n\in\mathcal{N}} \mathcal{Z}_n(\rho_n) = \prod_{n\in\mathcal{N}} \{\boldsymbol{\Delta x}_n : \|\boldsymbol{\Delta x}_n\|_p \le \rho_n\}, \tag{9}$$

where $\boldsymbol{\rho} := (\rho_n)_{n\in\mathcal{N}}$.

Then, the robust-feature MNL can be formulated as:

$$\max_{\boldsymbol{\beta}\in\mathcal{B}} \min_{\boldsymbol{\Delta x}\in\mathcal{Z}(\rho)} \sum_{n\in\mathcal{N}} \sum_{i\in C_n} y_{n,i} \cdot \log\left(\frac{\exp(\boldsymbol{\beta}_i^\top(\boldsymbol{x}_n + \boldsymbol{\Delta x}_n))}{\sum_{j\in C_n}\exp(\boldsymbol{\beta}_j^\top(\boldsymbol{x}_n + \boldsymbol{\Delta x}_n))}\right). \tag{10}$$

The selection of $\boldsymbol{\rho}$ depends on the data patterns. In practice, we can follow the typical hyper-parameter tuning ideas in machine learning by reserving some of the training data as a validation data set or using cross-validation to select the best uncertainty sets.

3.1. Binary logit model

The derivation of robust BNL is adapted from Bertsimas et al. (2019)'s work on robust binary logistic regression (i.e., not the original work of the paper). This is because the BNL model is equivalent to logistic regression in the binary classification case. Consider a binary logit model (BNL) with at most two alternatives for each individual (i.e., $C = \{1, 2\}$). Define $I_n \in C$ as the choice for individual n, and $J_n \in C$ as the counterpart (i.e., non-choice), where $y_{n,I_n} = 1, y_{n,J_n} = 0, \ \forall n \in \mathcal{N}$. Then the robust-feature BNL model can be reformulated as:

$$\max_{\boldsymbol{\beta}\in\mathcal{B}} \min_{\boldsymbol{\Delta x}\in\mathcal{Z}} \sum_{n\in\mathcal{N}} \log\left(\frac{1}{1+\exp\left(-(\boldsymbol{\beta}_{I_n} - \boldsymbol{\beta}_{J_n})^\top(\boldsymbol{x}_n + \boldsymbol{\Delta x}_n)\right)}\right). \tag{11}$$

The inner minimization problem is:

$$\min_{\boldsymbol{\Delta x}\in\mathcal{Z}} \sum_{n\in\mathcal{N}} -\log\left(1+\exp\left(-(\boldsymbol{\beta}_{I_n} - \boldsymbol{\beta}_{J_n})^\top(\boldsymbol{x}_n + \boldsymbol{\Delta x}_n)\right)\right) \tag{12}$$

$$\Longleftrightarrow \sum_{n\in\mathcal{N}} \min_{\boldsymbol{\Delta x}_n\in\mathcal{Z}_n} -\log\left(1+\exp\left(-(\boldsymbol{\beta}_{I_n} - \boldsymbol{\beta}_{J_n})^\top(\boldsymbol{x}_n + \boldsymbol{\Delta x}_n)\right)\right). \tag{13}$$

Let $s_n = (\boldsymbol{\beta}_{I_n} - \boldsymbol{\beta}_{J_n})^\top(\boldsymbol{x}_n + \boldsymbol{\Delta x}_n)$, and define $g(s_n) = -\log(1+\exp(-s_n))$. Notice that $g(s_n)$ is strictly increasing with the increase in s_n. Hence, for each $n \in \mathcal{N}$, to minimize the objective function in Eq. (13), we only need to minimize the following:

$$\min_{\|\boldsymbol{\Delta x}_n\|_p\le\rho_n} -(\boldsymbol{\beta}_{I_n} - \boldsymbol{\beta}_{J_n})^\top(\boldsymbol{x}_n + \boldsymbol{\Delta x}_n) \quad \forall n \in \mathcal{N}. \tag{14}$$

Lemma 1 (*Dual Norm*). *Let $\boldsymbol{x}$ be a vector. $\|\boldsymbol{x}\|_p$ is the ℓ_p norm of $\boldsymbol{x}$. Then, for any given vector $\boldsymbol{v}$, the dual norm problem is:*

$$\max_{\|\boldsymbol{x}\|_p\le\rho}\{\boldsymbol{v}^\top\boldsymbol{x}\} = \rho\cdot\|\boldsymbol{v}\|_q, \qquad \min_{\|\boldsymbol{x}\|_p\le\rho}\{\boldsymbol{v}^\top\boldsymbol{x}\} = -\rho\cdot\|\boldsymbol{v}\|_q, \tag{15}$$

where $\|\cdot\|_q$ is called the dual norm of $\|\cdot\|_p$ and $\frac{1}{q}+\frac{1}{p}=1$

Lemma 1 is a direct result of Hölder's inequality (Hölder, 1889). According to Lemma 1, the optimal objective function of Eq. (14) is

$$\min_{\|\boldsymbol{\Delta x}_n\|_p\le\rho_n} -(\boldsymbol{\beta}_{I_n} - \boldsymbol{\beta}_{J_n})^\top(\boldsymbol{x}_n+\boldsymbol{\Delta x}_n) = -(\boldsymbol{\beta}_{I_n} - \boldsymbol{\beta}_{J_n})^\top\boldsymbol{x}_n + \rho_n\|\boldsymbol{\beta}_{I_n} - \boldsymbol{\beta}_{J_n}\|_q, \tag{16}$$

where $\frac{1}{q}+\frac{1}{p}=1$.

Substituting the optimal value into Eq. (13), the robust binary logit model becomes:

$$\max_{\boldsymbol{\beta}\in\mathcal{B}} \sum_{n\in\mathcal{N}} -\log\left(1+\exp\left(-(\boldsymbol{\beta}_{I_n} - \boldsymbol{\beta}_{J_n})^\top\boldsymbol{x}_n + \rho_n\|\boldsymbol{\beta}_{I_n} - \boldsymbol{\beta}_{J_n}\|_q\right)\right) \tag{17}$$

$$\Longleftrightarrow \max_{\boldsymbol{\beta}\in\mathcal{B}} \sum_{n\in\mathcal{N}} \log\left(\frac{\exp\left(\boldsymbol{\beta}_{I_n}^\top\boldsymbol{x}_n\right)}{\exp\left(\boldsymbol{\beta}_{I_n}^\top\boldsymbol{x}_n\right) + \exp\left(\boldsymbol{\beta}_{J_n}^\top\boldsymbol{x}_n + \rho_n\|\boldsymbol{\beta}_{I_n} - \boldsymbol{\beta}_{J_n}\|_q\right)}\right). \tag{18}$$

Remark 1. Compared to the nominal BNL, the robust-feature counterpart of the BNL model has an additional $\rho_n\|\boldsymbol{\beta}_{I_n} - \boldsymbol{\beta}_{J_n}\|_q$ term in the exponent of the logit function (Eq. (17)). It resembles the ℓ_q-regularization term in typical machine learning problems (such as logistic regression). However, the additional term from robustness penalizes model complexity in the log-odds ratio, whereas the typical regularization term is a linear penalty on the entire likelihood.

To see the connections between the robust BNL and the typical regularization in machine learning, we can take the first-order Taylor approximation of Eq. (17). Define:

$$h_n(z) = -\log\left(1+\exp\left(-(\boldsymbol{\beta}_{I_n} - \boldsymbol{\beta}_{J_n})^\top\boldsymbol{x}_n + z\right)\right). \tag{19}$$

Then the first-order Taylor approximation of $h_n(z)$ at $z = 0$ is

$$h_n(z) \approx -\log\left(1+\exp\left(-(\boldsymbol{\beta}_{I_n} - \boldsymbol{\beta}_{J_n})^\top\boldsymbol{x}_n\right)\right) - \frac{\exp\left(-(\boldsymbol{\beta}_{I_n} - \boldsymbol{\beta}_{J_n})^\top\boldsymbol{x}_n\right)}{1+\exp\left(-(\boldsymbol{\beta}_{I_n} - \boldsymbol{\beta}_{J_n})^\top\boldsymbol{x}_n\right)}\cdot z. \tag{20}$$

Substitute $z = \rho_n\|\boldsymbol{\beta}_{I_n} - \boldsymbol{\beta}_{J_n}\|_q$ we have:

$$\begin{aligned}
&\sum_{n\in\mathcal{N}} -\log\left(1+\exp\left(-(\boldsymbol{\beta}_{I_n} - \boldsymbol{\beta}_{J_n})^\top\boldsymbol{x}_n + z\right)\right)\\
&\approx \sum_{n\in\mathcal{N}} \log\left(\frac{\exp\left(\boldsymbol{\beta}_{I_n}^\top\boldsymbol{x}_n\right)}{\exp\left(\boldsymbol{\beta}_{I_n}^\top\boldsymbol{x}_n\right) + \exp\left(\boldsymbol{\beta}_{J_n}^\top\boldsymbol{x}_n\right)}\right)\\
&\quad - \frac{\exp\left((\boldsymbol{\beta}_{J_n} - \boldsymbol{\beta}_{I_n})^\top\boldsymbol{x}_n\right)}{1+\exp\left((\boldsymbol{\beta}_{J_n} - \boldsymbol{\beta}_{I_n})^\top\boldsymbol{x}_n\right)}\cdot\rho_n\|\boldsymbol{\beta}_{I_n} - \boldsymbol{\beta}_{J_n}\|_q\\
&= \sum_{n\in\mathcal{N}} \log\left(P_{n,I_n}\right) - \sum_{n\in\mathcal{N}}\left(1-P_{n,I_n}\right)\cdot\rho_n\|\boldsymbol{\beta}_{I_n} - \boldsymbol{\beta}_{J_n}\|_q.
\end{aligned} \tag{21}$$

Therefore, when $\rho_n\|\boldsymbol{\beta}_{I_n} - \boldsymbol{\beta}_{J_n}\|_q$ is small, the robust BNL is approximately equivalent to the ℓ_q regularization in machine learning problems with penalty coefficients as $\sum_{n\in\mathcal{N}}\left(1-P_{n,I_n}\right)\cdot\rho_n$. This means that unlike typical machine learning regularization, this regularization effect will not diminish with increasing sample size.

Remark 2. When $\rho_n = 0$, the robust-feature BNL will fall back to the conventional BNL model. When $\rho_n = +\infty$ (which represents an extreme scenario where the data errors could go to infinity), the optimal value will be achieved when $\|\boldsymbol{\beta}_{I_n} - \boldsymbol{\beta}_{J_n}\|_q = 0, \forall n \in \mathcal{N}$ (i.e., $\boldsymbol{\beta}_{I_n} = \boldsymbol{\beta}_{J_n}$). In DCM, a feature is actually only put in one alternative for estimation purposes (i.e., $\beta_i^{(k)} = 0$ or $\beta_j^{(k)} = 0$ for a feature $k \in \mathcal{K}$, $i, j \in C$). Therefore, $\rho_n = +\infty$ will force the estimated $\boldsymbol{\beta}$ to be 0.

An extension of the robust BNL model is to consider a more general uncertainty set $\tilde{\mathcal{Z}}$ with multiple norm constraints. Let the set of all norm constraints be $\mathcal{P}$, then

$$\tilde{\mathcal{Z}} = \prod_{n\in\mathcal{N}}\prod_{i\in\mathcal{P}}\{\boldsymbol{\Delta x}_n : \|\boldsymbol{\Delta x}_n\|_{p_i} \le \rho_n^{(i)}\}. \tag{22}$$

Lemma 2 (*Dual Norm with Multiple Constraints*). *Let $\boldsymbol{x}$ be a vector. $\|\boldsymbol{x}\|_p$ is the ℓ_p norm of $\boldsymbol{x}$. Then, for any given vector $\boldsymbol{v}$, the dual norm problem with multiple constraints is:*

$$\max_{\boldsymbol{x}\in\prod_{i\in\mathcal{P}}\{\|\boldsymbol{x}\|_{p_i}\le\rho^{(i)}\}}\{\boldsymbol{v}^\top\boldsymbol{x}\} = \min_{\boldsymbol{v}_i}\sum_{i\in\mathcal{P}}\rho^{(i)}\cdot\|\boldsymbol{v}_i\|_{q_i} \quad s.t.\sum_{i\in\mathcal{P}}\boldsymbol{v}_i = \boldsymbol{v}, \tag{23}$$

$$\min_{\boldsymbol{x}\in\prod_{i\in\mathcal{P}}\{\|\boldsymbol{x}\|_{p_i}\le\rho^{(i)}\}}\{\boldsymbol{v}^\top\boldsymbol{x}\} = \max_{\boldsymbol{v}_i}\sum_{i\in\mathcal{P}}-\rho^{(i)}\cdot\|\boldsymbol{v}_i\|_{q_i} \quad s.t.\sum_{i\in\mathcal{P}}\boldsymbol{v}_i = \boldsymbol{v}, \tag{24}$$

where $\|\cdot\|_{q_i}$ *is called the dual norm of* $\|\cdot\|_{p_i}$ *and* $\frac{1}{q_i}+\frac{1}{p_i}=1,\ \forall i\in\mathcal{P}$

Lemma 2 is a direct result of Lemma 9 in Ben-Tal et al. (2015). Therefore, Eq. (14) with multiple uncertainty constraints can be reformulated as

$$\begin{aligned}&\prod_{i\in\mathcal{P}}\{\boldsymbol{\Delta x}_n:\|\boldsymbol{\Delta x}_n\|_{p_i}\le\rho_n^{(i)}\}\ \min\ -(\boldsymbol{\beta}_{I_n}-\boldsymbol{\beta}_{J_n})^\top(\boldsymbol{x}_n+\boldsymbol{\Delta x}_n)\\&=\max_{\boldsymbol{w}_n^{(i)}}-(\boldsymbol{\beta}_{I_n}-\boldsymbol{\beta}_{J_n})^\top\boldsymbol{x}_n+\sum_{i\in\mathcal{P}}\rho_n^{(i)}\|\boldsymbol{w}_n^{(i)}\|_{q_i},\quad\text{s.t.}\sum_{i\in\mathcal{P}}\boldsymbol{w}_n^{(i)}=\boldsymbol{\beta}_{I_n}-\boldsymbol{\beta}_{J_n},\end{aligned}\tag{25}$$

where $\frac{1}{q_i}+\frac{1}{p_i}=1,\forall i\in\mathcal{P}$. The final robust binary logit model with multiple uncertainty constraints becomes:

$$\begin{aligned}&\max_{\beta\in\mathcal{B},\,\boldsymbol{w}}\sum_{n\in\mathcal{N}}\log\left(\frac{\exp(\boldsymbol{\beta}_{I_n}^\top\boldsymbol{x}_n)}{\exp(\boldsymbol{\beta}_{I_n}^\top\boldsymbol{x}_n)+\exp(\boldsymbol{\beta}_{J_n}^\top\boldsymbol{x}_n+\sum_{i\in\mathcal{P}}\rho_n^{(i)}\|\boldsymbol{w}_n^{(i)}\|_{q_i})}\right)\\&\text{s.t.}\sum_{i\in\mathcal{P}}\boldsymbol{w}_n^{(i)}=\boldsymbol{\beta}_{I_n}-\boldsymbol{\beta}_{J_n}\quad\forall n\in\mathcal{N}.\end{aligned}\tag{26}$$

Two widely used examples for multiple-norm uncertainty sets are (1) ball-box uncertainty set (i.e., $\{\boldsymbol{\Delta x}_n:\|\boldsymbol{\Delta x}_n\|_2\le\rho_n^{(2)},\|\boldsymbol{\Delta x}_n\|_\infty\le\rho_n^{(\infty)}\}$) and (2) budget (box-polyhedral) uncertainty set (i.e., $\{\boldsymbol{\Delta x}_n:\|\boldsymbol{\Delta x}_n\|_1\le\rho_n^{(1)},\|\boldsymbol{\Delta x}_n\|_\infty\le\rho_n^{(\infty)}\}$) (Bertsimas & Hertog, 2022). By including multiple norm constraints with proper hyper-parameter tuning, we could better capture the potential error patterns, and prevent the robust model from choosing unreasonable points as the worst-case scenario.

3.2. Multinomial logit model

The robustification of the multinomial logit model (MNL) is more difficult than the binary model. The inner maximization problem of the robust MNL is equivalent to "maximizing a convex function". For the robust BNL model, the convex function is monotonically increasing, which leads to a direct simplification of the problem. However, the MNL model cannot be simplified in a similar way. In this study, we approximate the robust MNL problem using Jensen's inequality, which results in a similar formulation as the robust BNL model.

Similarly, let $I_n\in C_n$ be the choice index for individual n. Replacing the objective function of Eq. (10) by an auxiliary decision variable t, we have:

$$\max_{\beta\in\mathcal{B},\,t}\quad t\tag{27a}$$

$$\text{s.t.}\min_{\boldsymbol{\Delta x}\in\mathcal{Z}}\sum_{n\in\mathcal{N}}\left[-\log\left(\sum_{j\in C_n}\exp\big((\boldsymbol{\beta}_j-\boldsymbol{\beta}_{I_n})^\top(\boldsymbol{x}_n+\boldsymbol{\Delta x}_n)\big)\right)\right]\ge t.\tag{27b}$$

Since the uncertainty sets are independent across individuals, similar to Eq. (13), we have

$$\begin{aligned}&\min_{\boldsymbol{\Delta x}\in\mathcal{Z}}\sum_{n\in\mathcal{N}}-\log\left(\sum_{j\in C_n}\exp\big((\boldsymbol{\beta}_j-\boldsymbol{\beta}_{I_n})^\top(\boldsymbol{x}_n+\boldsymbol{\Delta x}_n)\big)\right)\ge t\\&\iff\sum_{n\in\mathcal{N}}\min_{\boldsymbol{\Delta x}_n\in\mathcal{Z}_n}-\log\left(\sum_{j\in C_n}\exp\big((\boldsymbol{\beta}_j-\boldsymbol{\beta}_{I_n})^\top(\boldsymbol{x}_n+\boldsymbol{\Delta x}_n)\big)\right)\ge t.\end{aligned}\tag{28}$$

Eq. (28) can be reformulated as:

$$\sum_{n\in\mathcal{N}}\max_{\boldsymbol{\Delta x}_n\in\mathcal{Z}_n}\log\left(\sum_{j\in C_n}\exp\big((\boldsymbol{\beta}_j-\boldsymbol{\beta}_{I_n})^\top(\boldsymbol{x}_n+\boldsymbol{\Delta x}_n)\big)\right)\le -t.\tag{29}$$

Note that we eliminate the negative sign and change the formulation to a maximization problem.

Consider the inner maximization problem in Eq. (29), according to the Jensen's inequality, we have:

$$\begin{aligned}&\max_{\boldsymbol{\Delta x}_n\in\mathcal{Z}_n}\log\left(\sum_{j\in C_n}\exp\big((\boldsymbol{\beta}_j-\boldsymbol{\beta}_{I_n})^\top(\boldsymbol{x}_n+\boldsymbol{\Delta x}_n)\big)\right)\\&\le\log\left(\sum_{j\in C_n}\exp\Big(\max_{\boldsymbol{\Delta x}_n\in\mathcal{Z}_n}(\boldsymbol{\beta}_j-\boldsymbol{\beta}_{I_n})^\top(\boldsymbol{x}_n+\boldsymbol{\Delta x}_n)\Big)\right)\end{aligned}\tag{30}$$

With the same derivation as Eq. (16), we now have:

$$\max_{\|\boldsymbol{\Delta x}_n\|_p\le\rho_n}(\boldsymbol{\beta}_j-\boldsymbol{\beta}_{I_n})^\top(\boldsymbol{x}_n+\boldsymbol{\Delta x}_n)=(\boldsymbol{\beta}_j-\boldsymbol{\beta}_{I_n})^\top\boldsymbol{x}_n+\rho_n\|\boldsymbol{\beta}_j-\boldsymbol{\beta}_{I_n}\|_q.\tag{31}$$

where $\frac{1}{q}+\frac{1}{p}=1$. Therefore, Eq. (29) can be approximated as:

$$\sum_{n\in\mathcal{N}}\log\left(\sum_{j\in C_n}\exp\Big((\boldsymbol{\beta}_j-\boldsymbol{\beta}_{I_n})^\top\boldsymbol{x}_n+\rho_n\|\boldsymbol{\beta}_j-\boldsymbol{\beta}_{I_n}\|_q\Big)\right)\le -t.\tag{32}$$

And the approximation of the robust-feature DCM can be reformulated as:

$$\max_{\beta\in\mathcal{B}}\sum_{n\in\mathcal{N}}\log\left(\frac{\exp(\boldsymbol{\beta}_{I_n}^\top\boldsymbol{x}_n)}{\sum_{j\in C_n}\exp(\boldsymbol{\beta}_j^\top\boldsymbol{x}_n+\rho_n\|\boldsymbol{\beta}_j-\boldsymbol{\beta}_{I_n}\|_q)}\right).\tag{33}$$

It has a similar form as the robust BNL model (Eq. (18)). Actually, when $|C|=2$, the robust MNL problem will reduce to the robust BNL problem. However, robust BNL is an exact robust counterpart while robust MNL is an approximation.

Remark 3. The solution of Eq. (33) is a lower bound of the original robust MNL problem (Eq. (10)). The reason is that, Constraint (32) is more restricted than the original constraint (Eq. (29)), which gives a smaller feasible region. Therefore, the optimal objective function in Eq. (33) is smaller than that in Eq. (10) (i.e., lower bound for a maximization problem). The inequality in Eq. (30) is tight when (1) $(\boldsymbol{\beta}_j-\boldsymbol{\beta}_{I_n})=(\boldsymbol{\beta}_i-\boldsymbol{\beta}_{I_n})$ for all $i,j\in C_n$. (2) Or one $\boldsymbol{\beta}_i-\boldsymbol{\beta}_{I_n}$ significantly larger than others. (i.e., $\exists i^*$ such that $(\boldsymbol{\beta}_{i^*}-\boldsymbol{\beta}_{I_n})^\top(\boldsymbol{x}_n+\boldsymbol{\Delta x}_n)\gg(\boldsymbol{\beta}_j-\boldsymbol{\beta}_{I_n})^\top(\boldsymbol{x}_n+\boldsymbol{\Delta x}_n), j\ne i^*$). The first case implies $\boldsymbol{\beta}=0$, which will be achieved when $\rho_n\to\infty$. More analysis on the tightness of the inequality can be found in Appendix A. Moreover, we also derive an upper bound of the original robust MNL problem in Appendix B. Combining it with the lower bound could provide the optimality gap for the approximation.

Similar to the extension of Robust BNL, for a general uncertainty set $\tilde{\mathcal{Z}}$ (Eq. (22)), the robust MNL problem is:

$$\begin{aligned}&\max_{\beta\in\mathcal{B},\,\boldsymbol{w}}\sum_{n\in\mathcal{N}}\log\left(\frac{\exp(\boldsymbol{\beta}_{I_n}^\top\boldsymbol{x}_n)}{\sum_{j\in C_n}\exp(\boldsymbol{\beta}_j^\top\boldsymbol{x}_n+\sum_{i\in\mathcal{P}}\rho_n^{(i)}\|\boldsymbol{w}_n^{(i,j)}\|_q)}\right)\\&\text{s.t.}\sum_{i\in\mathcal{P}}\boldsymbol{w}_n^{(i,j)}=\boldsymbol{\beta}_j-\boldsymbol{\beta}_{I_n}\quad\forall n\in\mathcal{N},\ \forall j\in C_n.\end{aligned}\tag{34}$$

It is worth noting that the robust-feature MNL model is solved under the worst-case scenario. Whether it works well or not depends on how the uncertainty set is defined and what are the error patterns in the testing data. To avoid solutions from being too conservative, we need to choose a proper value of ρ_n, or integrate multiple norm constraints to define the uncertainty set.

4. Robustness against uncertainties in labels

Section 3 discusses the possible uncertainties in features (i.e., $\boldsymbol{\Delta x}$). The measurement errors may also apply to labels or individual choices (i.e., $\boldsymbol{\Delta y}$). In this study, we consider the following uncertainty set:

$$\begin{aligned}\mathcal{U}(\Gamma)=\{\boldsymbol{\Delta y}:\ &\sum_{j\in C_n}\Delta y_{n,j}=0,\ \Delta y_{n,I_n}\in\{0,-1\},\forall n\in\mathcal{N};\\&\Delta y_{n,j}\in\{0,1\}\ \forall j\in C_n\setminus\{I_n\},\ \forall n\in\mathcal{N};\\&\sum_{n\in\mathcal{N}}-\Delta y_{n,I_n}\le\Gamma\}.\end{aligned}\tag{35}$$

Specially, if $\Delta y_{n,I_n}=-1$ and $\Delta y_{n,j}=1$, it means that the individual's actual choice is $I_n\in C_n$ but the data mislabeled it as $j\in C_n\setminus\{I_n\}$. The uncertainty set $\mathcal{U}$ represents that there are at most Γ mislabeled samples. Then the robust DCM problem against feature uncertainty can be represented as:

$$\max_{\beta\in\mathcal{B}}\min_{\boldsymbol{\Delta y}\in\mathcal{U}(\Gamma)}\sum_{n\in\mathcal{N}}\sum_{i\in C_n}(y_{n,i}+\Delta y_{n,i})\cdot\log(P_{n,i}(\boldsymbol{\beta})).\tag{36}$$

The formulation is general for both BNL and MNL cases. Consider the convex hull of $\mathcal{U}(\Gamma)$:

$$\begin{aligned}\text{Conv}(\mathcal{U}(\Gamma)) = \{\boldsymbol{\Delta y} : & \sum_{j\in C_n} \Delta y_{n,j} = 0, -1 \le \Delta y_{n,I_n} \le 0, \forall n \in \mathcal{N};\\ & 0 \le \Delta y_{n,j} \le 1 \ \forall j \in C_n \setminus \{I_n\},\ \forall n \in \mathcal{N};\\ & \sum_{n\in\mathcal{N}} -\Delta y_{n,I_n} \le \Gamma\}. \end{aligned} \tag{37}$$

The inner minimization problem is linear in $\boldsymbol{\Delta y}$, and the extreme points for $\text{Conv}(\mathcal{U}(\Gamma))$ are integers. Hence, the original inner minimization problem on $\mathcal{U}(\Gamma)$ is equivalent to minimizing over its convex hull:

$$\min_{\boldsymbol{\Delta y}} \sum_{n\in\mathcal{N}} \sum_{i\in C_n} (y_{n,i} + \Delta y_{n,i}) \cdot \log(P_{n,i}(\boldsymbol{\beta})) \tag{38a}$$

$$\text{s.t. } \Delta y_{n,I_n} + \sum_{j\in C_n\setminus\{I_n\}} \Delta y_{n,j} = 0, \quad \forall n \in \mathcal{N}, \tag{38b}$$

$$0 \le -\Delta y_{n,I_n} \le 1, \quad \forall n \in \mathcal{N}, \tag{38c}$$

$$0 \le \Delta y_{n,j} \le 1, \quad \forall n \in \mathcal{N}, \forall j \in C_n \setminus \{I_n\}, \tag{38d}$$

$$\sum_{n\in\mathcal{N}} -\Delta y_{n,I_n} \le \Gamma. \tag{38e}$$

As $\boldsymbol{\Delta y}$ is bounded, the optimal solution of Eq. (38) is also bounded for a given $\boldsymbol{\beta}$. By strong duality, the optimal solution in Eq. (38) equals that of its dual problem

$$\max_{\boldsymbol{\eta},\boldsymbol{\gamma},\lambda} \sum_{n\in\mathcal{N}} \sum_{i\in C_n} y_{n,i} \cdot \log(P_{n,i}(\boldsymbol{\beta})) + \sum_{n\in\mathcal{N}} \sum_{i\in C_n} \eta_{n,i} + \Gamma \cdot \lambda \tag{39a}$$

$$\text{s.t. } -\gamma_n + \eta_{n,I_n} + \lambda \le \log(P_{n,I_n}(\boldsymbol{\beta})), \quad \forall n \in \mathcal{N}, \tag{39b}$$

$$\gamma_n + \eta_{n,j} \le \log(P_{n,j}(\boldsymbol{\beta})), \quad \forall n \in \mathcal{N}, \forall j \in C_n \setminus \{I_n\}, \tag{39c}$$

$$\eta_{n,i} \le 0, \quad \forall n \in \mathcal{N}, i \in C_n, \tag{39d}$$

$$\lambda \le 0, \tag{39e}$$

where $\boldsymbol{\gamma} = (\gamma_n)_{n\in\mathcal{N}}$, $\boldsymbol{\eta} = (\eta_{n,i})_{n\in\mathcal{N}, i\in C_n}$, and λ are dual decision variables.

The final formulation is just a combination of the inner and outer problems:

$$\max_{\boldsymbol{\beta}\in\mathcal{B},\boldsymbol{\eta},\boldsymbol{\gamma},\lambda} \sum_{n\in\mathcal{N}} \log\left(\frac{\exp(\boldsymbol{\beta}_{I_n}^\top \boldsymbol{x}_n)}{\sum_{j\in C_n}\exp(\boldsymbol{\beta}_j^\top \boldsymbol{x}_n)}\right) + \sum_{n\in\mathcal{N}} \sum_{i\in C_n} \eta_{n,i} + \Gamma \cdot \lambda \tag{40a}$$

$$\text{s.t. Constraints (39b) } \sim \text{ (39e).} \tag{40b}$$

This problem has a twice continuously differentiable concave objective function and constraints, making it tractably solvable with an interior point method.

Remark 4. When $\Gamma = 0$, the optimal solution for $\boldsymbol{\eta}$ is 0 because $\boldsymbol{\gamma}$ and λ become free variables. Then constraints (39b) and (39c) do not restrict $\boldsymbol{\eta}$ to take its maximum value. Therefore, when $\Gamma = 0$ (i.e., no uncertainty), Eq. (40) is equivalent to the nominal MNL model, which validates the formulation.

5. Statistical properties of the robust DCM estimators

In the context of econometrics, it is often of interest to understand the statistical properties of an estimator (i.e., consistency and efficiency). In this study, we first show that robust formulations (both robust feature and robust label) tend to shrink the scale of the estimated $\boldsymbol{\beta}$ compared to nominal formulations, which implies that they are biased and inconsistent estimators. However, we also show that this shrinkage reduces variances and could make the model perform better in out-of-sample predictions, especially when there are errors in the testing data.

5.1. Statistical properties of the robust feature estimators

It is worth noting that the robust BNL formulation (Eq. (18)) is a special case of the robust MNL formulation (Eq. (33)). Therefore, we only need to discuss the properties of robust MNL formulations.

Proposition 1. *Let $\hat{\boldsymbol{\beta}}^{RF}$ be the solution of Eq. (33) (robust feature MNL problem). $\hat{\boldsymbol{\beta}}^{RF}$ is an inconsistent estimate of the actual parameter.*

Proof. For simplicity of description, we first consider a single sample n. Let us rewrite the objective function of Eq. (33) for sample n as

$$\mathcal{LL}_n^{\text{RF}} = -\log \sum_{j\in C_n} \exp\left(\left(\boldsymbol{\beta}_j - \boldsymbol{\beta}_{I_n}\right)^\top \boldsymbol{x}_n + \rho_n \|\boldsymbol{\beta}_j - \boldsymbol{\beta}_{I_n}\|_q\right). \tag{41}$$

Without loss of generality, let us assume all samples' choices are $I_n = I \in C$, and define $\tilde{\boldsymbol{\beta}}_j = (\boldsymbol{\beta}_j - \boldsymbol{\beta}_I)$. Then Eq. (41) can be transformed to:

$$\mathcal{LL}_n^{\text{RF}}(\tilde{\boldsymbol{\beta}}) = -\log \sum_{j\in C_n} \exp\left(\tilde{\boldsymbol{\beta}}_j^\top \boldsymbol{x}_n + \rho_n \|\tilde{\boldsymbol{\beta}}_j\|_q\right). \tag{42}$$

Then maximizing Eq. (42) is essentially minimizing $\log\sum_{j\in C_n}\exp\left(\tilde{\boldsymbol{\beta}}_j^\top \boldsymbol{x}_n\right) \cdot \exp\left(\rho_n \|\tilde{\boldsymbol{\beta}}_j\|_q\right)$. Compared to the nominal MNL model, there is an additional bias term $\exp(\rho_n\|\tilde{\boldsymbol{\beta}}_j\|_q)$ for every sample n, and it does not diminish as $|\mathcal{N}|$ goes to infinity. Hence, the robust-feature MNL estimator is inconsistent. And by definition, it is also inefficient. □

Remark 5. Note that the value of $\hat{\boldsymbol{\beta}}^{\text{RF}}$ depends on external hyper-parameter ρ_n. Let $\hat{\tilde{\boldsymbol{\beta}}}^{\text{RF}}$ be the transformation of $\hat{\boldsymbol{\beta}}^{\text{RF}}$ as above. As ρ_n goes to infinity, the optimal solution will be achieved at $\hat{\tilde{\boldsymbol{\beta}}}^{\text{RF}} = 0$, which implies $\hat{\boldsymbol{\beta}}_j^{\text{RF}} = \hat{\boldsymbol{\beta}}_I^{\text{RF}}$ for all $j \in C$. In the specification of DCM, one of the alternatives (i.e., the base alternative) will have a fixed coefficient of feature $k \in \mathcal{K}$ to be zero (to avoid perfect co-linearity). Therefore, $\exists i \in C$ such that $\hat{\boldsymbol{\beta}}_i^{\text{RF},(k)} = 0$, $\forall k \in \mathcal{K}$. Then $\hat{\boldsymbol{\beta}}_j^{\text{RF}} = \hat{\boldsymbol{\beta}}_I^{\text{RF}}$ is equivalent to $\hat{\boldsymbol{\beta}}^{\text{RF}} = 0$. Therefore, a larger value of ρ_n will shrink $\hat{\boldsymbol{\beta}}^{\text{RF}}$ toward 0.

Remark 6. We may also consider the case when the size of the uncertainty set goes to zero (i.e., $\rho_n \to 0$). In this case, the bias term has $\lim_{\rho_n\to 0}\exp(\rho_n\|\tilde{\boldsymbol{\beta}}_j\|_q) = 1$ because the function is continuous in ρ_n. Then the objective function becomes $\log\sum_{j\in C_n}\exp\left(\tilde{\boldsymbol{\beta}}_j^\top \boldsymbol{x}_n\right)$, which reduces to the nominal MNL model. Therefore, when the size of the uncertainty set goes to zero, the robust formulation will converge to the consistent and unbiased MLE estimator. This is similar to the case where the distributional robust optimization reduces to the empirical risk minimization or stochastic optimization as the ambiguity set shrinks toward zero (Mohajerin Esfahani & Kuhn, 2018; Sun & Xu, 2016).

However, just like how regularization works in machine learning, a biased estimator may perform better in out-of-sample predictions due to lower estimated variances. Rigorously proving the relationship of variance between two estimators is difficult. In Proposition 2, we show that the robust estimator yields a larger trace of the Fisher information matrix than the original MLE estimator, which in general could imply lower variances.

Proposition 2. *Let $\mathcal{I}_{RF}(\cdot)$ and $\mathcal{I}_{MLE}(\cdot)$ be the Fisher information matrix of the robust-feature MNL estimator and the original MLE estimator, respectively. If ρ_n is large enough for all $n \in \mathcal{N}$, we have*

$$Tr\left(\mathcal{I}_{RF}(\hat{\boldsymbol{\beta}}^{RF})\right) \ge Tr\left(\mathcal{I}_{MLE}(\hat{\boldsymbol{\beta}}^{MLE})\right). \tag{43}$$

where $\hat{\boldsymbol{\beta}}^{RF}$ and $\hat{\boldsymbol{\beta}}^{MLE}$ are corresponding estimated parameters. $Tr(\cdot)$ is the trace of a matrix. The specific conditions for ρ_n are provided in the proof.

Proof. Using the same transformation of $\boldsymbol{\beta}$ as in the proof of Proposition 1, we define $\hat{\tilde{\boldsymbol{\beta}}}^{\text{RF}}$ and $\hat{\tilde{\boldsymbol{\beta}}}^{\text{MLE}}$, where $\hat{\tilde{\boldsymbol{\beta}}}_j^{\text{RF}} = \hat{\boldsymbol{\beta}}_j^{\text{RF}} - \hat{\boldsymbol{\beta}}_I^{\text{RF}}$ and $\hat{\tilde{\boldsymbol{\beta}}}_j^{\text{MLE}} =$

$\hat{\boldsymbol{\beta}}_j^{\text{MLE}} - \hat{\boldsymbol{\beta}}_I^{\text{MLE}}$ (for all $j \in C$). Notice that this transformation is an orthogonal linear mapping: $\hat{\tilde{\boldsymbol{\beta}}}^{\text{RF}} = \boldsymbol{A} \cdot \hat{\boldsymbol{\beta}}^{\text{RF}}$, where $\boldsymbol{A}$ is an orthogonal matrix and $\boldsymbol{A}^\top \cdot \boldsymbol{A} = \boldsymbol{I}$. Therefore, to show $\text{Tr}\left(\mathcal{I}_{\text{RF}}(\hat{\boldsymbol{\beta}}^{\text{RF}})\right) \geq \text{Tr}\left(\mathcal{I}_{\text{MLE}}(\hat{\boldsymbol{\beta}}^{\text{MLE}})\right)$, we only need to show $\text{Tr}\left(\mathcal{I}_{\text{RF}}(\hat{\tilde{\boldsymbol{\beta}}}^{\text{RF}})\right) \geq \text{Tr}\left(\mathcal{I}_{\text{MLE}}(\hat{\tilde{\boldsymbol{\beta}}}^{\text{MLE}})\right)$ because the Fisher information matrix is essentially the negative Hessian matrix, and orthogonal linear mapping does not change the eigenvalues of the Hessian matrix, thus the trace (sum of eigenvalues) are still the same.

Similar to Eq. (41), let $\mathcal{LL}_n^{\text{MLE}}$ be the log-likelihood of the original MNL model for sample n, we can also transfer it to

$$\mathcal{LL}_n^{\text{MLE}}(\tilde{\boldsymbol{\beta}}) = -\log \sum_{j \in C_n} \exp\left(\tilde{\boldsymbol{\beta}}_j^\top \boldsymbol{x}_n\right). \tag{44}$$

Calculating the first derivatives of Eqs. (41) and (44) gives:

$$\begin{aligned}\frac{\partial \mathcal{LL}_n^{\text{RF}}}{\partial \tilde{\boldsymbol{\beta}}_i} &= -\frac{\exp\left(\tilde{\boldsymbol{\beta}}_i^\top \boldsymbol{x}_n + \rho_n \|\tilde{\boldsymbol{\beta}}_i\|_q\right) \cdot \left(\boldsymbol{x}_n + \rho_n \cdot \nabla_{\tilde{\boldsymbol{\beta}}_i} \|\tilde{\boldsymbol{\beta}}_i\|_q\right)}{\sum_{j \in C_n} \exp\left(\tilde{\boldsymbol{\beta}}_j^\top \boldsymbol{x}_n + \rho_n \|\tilde{\boldsymbol{\beta}}_j\|_q\right)} \\ &= -\tilde{P}_{i,n}^{\text{Norm}} \cdot \left(\boldsymbol{x}_n + \rho_n \cdot \nabla_{\tilde{\boldsymbol{\beta}}_i} \|\tilde{\boldsymbol{\beta}}_i\|_q\right),\end{aligned} \tag{45}$$

$$\frac{\partial \mathcal{LL}_n^{\text{MLE}}}{\partial \tilde{\boldsymbol{\beta}}_i} = -\frac{\exp\left(\tilde{\boldsymbol{\beta}}_i^\top \boldsymbol{x}_n\right) \cdot \boldsymbol{x}_n}{\sum_{j \in C_n} \exp\left(\tilde{\boldsymbol{\beta}}_j^\top \boldsymbol{x}_n\right)} = -\tilde{P}_{i,n} \cdot \boldsymbol{x}_n, \tag{46}$$

where $\tilde{P}_{i,n}^{\text{Norm}} := \frac{\exp\left(\tilde{\boldsymbol{\beta}}_i^\top \boldsymbol{x}_n + \rho_n \|\tilde{\boldsymbol{\beta}}_i\|_q\right)}{\sum_{j \in C_n} \exp\left(\tilde{\boldsymbol{\beta}}_j^\top \boldsymbol{x}_n + \rho_n \|\tilde{\boldsymbol{\beta}}_j\|_q\right)}$ and $\tilde{P}_{i,n} := \frac{\exp\left(\tilde{\boldsymbol{\beta}}_i^\top \boldsymbol{x}_n\right)}{\sum_{j \in C_n} \exp\left(\tilde{\boldsymbol{\beta}}_j^\top \boldsymbol{x}_n\right)}$ are auxiliary variables. For the second derivatives, we first consider the diagonal terms:

$$\begin{aligned}\frac{\partial^2 \mathcal{LL}_n^{\text{RF}}}{\partial (\tilde{\boldsymbol{\beta}}_i)^2} =& -\tilde{P}_{i,n}^{\text{Norm}} \cdot (1 - \tilde{P}_{i,n}^{\text{Norm}}) \cdot \left(\boldsymbol{x}_n + \rho_n \cdot \nabla_{\tilde{\boldsymbol{\beta}}_i} \|\tilde{\boldsymbol{\beta}}_i\|_q\right) \cdot \left(\boldsymbol{x}_n + \rho_n \cdot \nabla_{\tilde{\boldsymbol{\beta}}_i} \|\tilde{\boldsymbol{\beta}}_i\|_q\right)^\top \\ & - \tilde{P}_{i,n}^{\text{Norm}} \cdot \rho_n \cdot \nabla^2_{\tilde{\boldsymbol{\beta}}_i} \|\tilde{\boldsymbol{\beta}}_i\|_q,\end{aligned} \tag{47}$$

$$\frac{\partial^2 \mathcal{LL}_n^{\text{MLE}}}{\partial (\tilde{\boldsymbol{\beta}}_i)^2} = -\tilde{P}_{i,n} \cdot (1 - \tilde{P}_{i,n}) \cdot \boldsymbol{x}_n \cdot \boldsymbol{x}_n^\top. \tag{48}$$

For off-diagonal terms:

$$\frac{\partial^2 \mathcal{LL}_n^{\text{RF}}}{\partial \tilde{\boldsymbol{\beta}}_i \partial \tilde{\boldsymbol{\beta}}_j} = \tilde{P}_{i,n}^{\text{Norm}} \cdot \tilde{P}_{j,n}^{\text{Norm}} \cdot \left(\boldsymbol{x}_n + \rho_n \cdot \nabla_{\tilde{\boldsymbol{\beta}}_j} \|\tilde{\boldsymbol{\beta}}_j\|_q\right) \cdot \left(\boldsymbol{x}_n + \rho_n \cdot \nabla_{\tilde{\boldsymbol{\beta}}_i} \|\tilde{\boldsymbol{\beta}}_i\|_q\right)^\top, \tag{49}$$

$$\frac{\partial^2 \mathcal{LL}_n^{\text{MLE}}}{\partial \tilde{\boldsymbol{\beta}}_i \partial \tilde{\boldsymbol{\beta}}_j} = \tilde{P}_{i,n} \cdot \tilde{P}_{j,n} \cdot \boldsymbol{x}_n \cdot \boldsymbol{x}_n^\top. \tag{50}$$

Consider the kth feature in alternative i. Let of corresponding diagonal term in $\mathcal{I}_{\text{RF}}(\hat{\tilde{\boldsymbol{\beta}}}^{\text{RF}})$ be $\mathcal{I}_{\text{RF}}(\hat{\tilde{\boldsymbol{\beta}}}^{\text{RF}})_{(i,k),(i,k)} \in \mathbb{R}$, we have:

$$\begin{aligned}\mathcal{I}_{\text{RF}}(\hat{\tilde{\boldsymbol{\beta}}}^{\text{RF}})_{(i,k),(i,k)} &= \sum_{n \in \mathcal{N}} -\left(\frac{\partial^2 \mathcal{LL}_n^{\text{RF}}}{\partial (\tilde{\boldsymbol{\beta}}_i)^2}\right)_{k,k} \\ &= \sum_{n \in \mathcal{N}} \tilde{P}_{i,n}^{\text{Norm}} \cdot (1 - \tilde{P}_{i,n}^{\text{Norm}}) \cdot \left(\left(\boldsymbol{x}_n^{(k)}\right)^2 + 2\rho_n \cdot \left(\boldsymbol{x}_n^{(k)}\right) \cdot \left(\nabla_{\tilde{\boldsymbol{\beta}}_i} \|\tilde{\boldsymbol{\beta}}_i\|_q\right)_k \right. \\ &\quad \left. + (\rho_n)^2 \cdot \left(\nabla_{\tilde{\boldsymbol{\beta}}_i} \|\tilde{\boldsymbol{\beta}}_i\|_q\right)_k^2\right) + \tilde{P}_{i,n}^{\text{Norm}} \cdot \rho_n \cdot \left(\nabla^2_{\tilde{\boldsymbol{\beta}}_i} \|\tilde{\boldsymbol{\beta}}_i\|_q\right)_{k,k}.\end{aligned} \tag{51}$$

Similarly, for the MLE estimator:

$$\mathcal{I}_{\text{MLE}}(\hat{\tilde{\boldsymbol{\beta}}}^{\text{MLE}})_{(i,k),(i,k)} = \sum_{n \in \mathcal{N}} -\left(\frac{\partial^2 \mathcal{LL}_n^{\text{MLE}}}{\partial (\tilde{\boldsymbol{\beta}}_i)^2}\right)_{k,k} = \sum_{n \in \mathcal{N}} \tilde{P}_{i,n} \cdot (1 - \tilde{P}_{i,n}) \cdot \left(\boldsymbol{x}_n^{(k)}\right)^2. \tag{52}$$

Therefore,

$$\text{Tr}(\mathcal{I}_{\text{RF}}(\hat{\tilde{\boldsymbol{\beta}}}^{\text{RF}})) = \sum_{i \in C, k \in \mathcal{K}} \mathcal{I}_{\text{RF}}(\hat{\tilde{\boldsymbol{\beta}}}^{\text{RF}})_{(i,k),(i,k)}, \tag{53}$$

$$\text{Tr}(\mathcal{I}_{\text{MLE}}(\hat{\tilde{\boldsymbol{\beta}}}^{\text{MLE}})) = \sum_{i \in C, k \in \mathcal{K}} \mathcal{I}_{\text{MLE}}(\hat{\tilde{\boldsymbol{\beta}}}^{\text{MLE}})_{(i,k),(i,k)}. \tag{54}$$

Comparing Eqs. (53) and (54), to argue $\text{Tr}(\mathcal{I}_{\text{RF}}(\hat{\tilde{\boldsymbol{\beta}}}^{\text{RF}})) \geq \text{Tr}(\mathcal{I}_{\text{MLE}}(\hat{\tilde{\boldsymbol{\beta}}}^{\text{MLE}}))$, we need to satisfy two requirements:

- **Requirement 1**: For the shared term $\left(\boldsymbol{x}_n^{(k)}\right)^2$, the coefficients are $\sum_{i \in C} \tilde{P}_{i,n}^{\text{Norm}} \cdot (1 - \tilde{P}_{i,n}^{\text{Norm}})$ and $\sum_{i \in C} \tilde{P}_{i,n} \cdot (1 - \tilde{P}_{i,n})$ for Eqs. (53) and (54), respectively. We need to show $\sum_{i \in C} \tilde{P}_{i,n}^{\text{Norm}}(\hat{\tilde{\boldsymbol{\beta}}}^{\text{RF}}) \cdot (1 - \tilde{P}_{i,n}^{\text{Norm}}(\hat{\tilde{\boldsymbol{\beta}}}^{\text{RF}})) \geq \sum_{i \in C} \tilde{P}_{i,n}(\hat{\tilde{\boldsymbol{\beta}}}^{\text{MLE}}) \cdot (1 - \tilde{P}_{i,n}(\hat{\tilde{\boldsymbol{\beta}}}^{\text{MLE}}))$
- **Requirement 2**: The remaining parts in Eq. (53) (excluding the shared term $\left(\boldsymbol{x}_n^{(k)}\right)^2$) are $\sum_{i \in C, k \in \mathcal{K}} \sum_{n \in \mathcal{N}} 2 \cdot \rho_n \cdot \left(\boldsymbol{x}_n^{(k)}\right) \cdot \left(\nabla_{\tilde{\boldsymbol{\beta}}_i} \|\tilde{\boldsymbol{\beta}}_i\|_q\right)_k + (\rho_n)^2 \cdot \left(\nabla_{\tilde{\boldsymbol{\beta}}_i} \|\tilde{\boldsymbol{\beta}}_i\|_q\right)_k^2$ and $\sum_{i \in C, k \in \mathcal{K}} \sum_{n \in \mathcal{N}} \tilde{P}_{i,n}^{\text{Norm}} \cdot \rho_n \cdot \left(\nabla^2_{\tilde{\boldsymbol{\beta}}_i} \|\tilde{\boldsymbol{\beta}}_i\|_q\right)_{k,k}$, respectively. We need to show they are both non-negative. As ℓ_q norm is a convex function, its Hessian matrix is positive semi-definite. Therefore, $\left(\nabla^2_{\tilde{\boldsymbol{\beta}}_i} \|\tilde{\boldsymbol{\beta}}_i\|_q\right)_{k,k} \geq 0$. So the second remaining part is non-negative. We only need to show the first part is non-negative.

Recall our preliminary for the proposition is ρ_n is large enough. Its specific conditions are as follows:

- **Condition 1**: ρ_n is large enough such that $\sum_{i \in C} \tilde{P}_{i,n}^{\text{Norm}}(\hat{\tilde{\boldsymbol{\beta}}}^{\text{RF}}) \cdot (1 - \tilde{P}_{i,n}^{\text{Norm}}(\hat{\tilde{\boldsymbol{\beta}}}^{\text{RF}})) \geq \sum_{i \in C} \tilde{P}_{i,n}(\hat{\tilde{\boldsymbol{\beta}}}^{\text{MLE}}) \cdot (1 - \tilde{P}_{i,n}(\hat{\tilde{\boldsymbol{\beta}}}^{\text{MLE}}))$. We will explain later why this condition is related to the scale of ρ_n.
- **Condition 2**: ρ_n is large enough such that

$$\rho_n \geq 2 \left| \frac{\boldsymbol{x}_n^{(k)}}{\left(\nabla_{\tilde{\boldsymbol{\beta}}_i = \hat{\tilde{\boldsymbol{\beta}}}_i^{\text{RF}}} \|\tilde{\boldsymbol{\beta}}_i\|_q\right)_k} \right|. \tag{55}$$

The explanation for Condition 1 is as follows. For function with form $\sum_i p_i(1 - p_i)$ and $\sum_i p_i = 1$. The maximum value will be achieved when every p_i is the same. When ρ_n is large enough, based on Proposition 1, $\hat{\tilde{\boldsymbol{\beta}}}^{\text{RF}}$ will shrink toward 0, which will make $\tilde{P}_{i,n}^{\text{Norm}}$ closer to $\frac{1}{|C|}$ (i.e., more similar to each other) than $\tilde{P}_{i,n}$. Therefore, Condition 1 can be achieved when ρ_n is large enough. Based on condition 2, we can easily derive that:

$$2\rho_n \cdot \left(\nabla_{\tilde{\boldsymbol{\beta}}_i = \hat{\tilde{\boldsymbol{\beta}}}_i^{\text{RF}}} \|\tilde{\boldsymbol{\beta}}_i\|_q\right)_k + (\rho_n)^2 \cdot \left(\nabla_{\tilde{\boldsymbol{\beta}}_i = \hat{\tilde{\boldsymbol{\beta}}}_i^{\text{RF}}} \|\tilde{\boldsymbol{\beta}}_i\|_q\right)_k^2 \geq 0. \tag{56}$$

Eq. (56) implies the first remaining part listed in Requirement 2 is non-negative. Since Condition 1 directly implies Requirement 1, we have satisfied all requirements needed for claiming $\text{Tr}(\mathcal{I}_{\text{RF}}(\hat{\tilde{\boldsymbol{\beta}}}^{\text{RF}})) \geq \text{Tr}(\mathcal{I}_{\text{MLE}}(\hat{\tilde{\boldsymbol{\beta}}}^{\text{MLE}}))$ □

Proposition 2 implies that, though the robust-feature MNL estimator is biased, its Fisher information matrix is larger (in terms of traces) than the original MLE estimator. Using the generalized Cramer–Rao bound (Cramér, 2016; Rao, 1992), we have:

$$\text{Var}[\hat{\tilde{\boldsymbol{\beta}}}^{\text{RF}}] \geq \text{Diag}\left(\mathcal{I}_{\text{RF}}(\hat{\tilde{\boldsymbol{\beta}}}^{\text{RF}})^{-1} \cdot \left(\boldsymbol{I} + \boldsymbol{b}\boldsymbol{b}^\top\right)\right), \tag{57}$$

$$\text{Var}[\hat{\tilde{\boldsymbol{\beta}}}^{\text{MLE}}] \geq \text{Diag}\left(\mathcal{I}_{\text{RF}}(\hat{\tilde{\boldsymbol{\beta}}}^{\text{MLE}})^{-1}\right), \tag{58}$$

where $\boldsymbol{b}$ is the gradient of the bias term (usually non-tractable). Therefore, a larger trace of the Fisher information matrix usually implies lower estimated variance (Pukelsheim, 2006). However, the rigorous proof is difficult given the complex form of the Hessian matrix and the non-tractable bias term.

5.2. Statistical properties of the robust label estimators

The analysis of robust label estimators is similar to robust feature estimators. We will show that the estimated parameter $\hat{\boldsymbol{\beta}}^{\text{RL}}$ depends on external hyper-parameter Γ, thus is inconsistent, and also show that it has a larger Fisher information matrix.

Proposition 3. *Let $\hat{\boldsymbol{\beta}}^{RL}$ be as the solution of Eq. (40) (robust label MNL problem). $\hat{\boldsymbol{\beta}}^{RL}$ is an inconsistent estimate of the actual parameter. The value of $\hat{\boldsymbol{\beta}}^{RL}$ depends on external hyper-parameter Γ. A larger value of Γ will shrink the scale of $\hat{\boldsymbol{\beta}}^{RL}$.*

Proof. Let us decompose the original robust-label MNL problem (Eq. (36)) as:

$$\max_{\beta\in\mathcal{B}}\left(\sum_{n\in\mathcal{N}}\sum_{i\in C_n} y_{n,i}\cdot\log(P_{n,i}(\boldsymbol{\beta}))+\min_{\Delta\boldsymbol{y}\in\mathcal{U}(\Gamma)}\left(\sum_{n\in\mathcal{N}}\sum_{i\in C_n}\Delta y_{n,i}\cdot\log(P_{n,i}(\boldsymbol{\beta}))\right)\right). \quad (59)$$

From the definition of $\mathcal{U}(\Gamma)$, we need either have $(\Delta y_{n,j}=0,\forall j\in C_n)$ or $(\Delta y_{n,I_n}=-1$, one of the $\Delta y_{n,j}=1, j\neq I_n)$. To minimize $\sum_{i\in C_n}\Delta y_{n,i}\cdot\log(P_{n,i}(\boldsymbol{\beta}))$, we need to consider the relationship of $P_{n,j}$ for $j\in C_n$. For this single sample n, one could easily derive the optimal selection of $\Delta y_{n,j}$ is:

$$\Delta y_{n,J_n^*}=1,\ \Delta y_{n,I_n}=-1 \quad \text{if } P_{n,I_n}(\boldsymbol{\beta})>P_{n,j^*}(\boldsymbol{\beta}),\ \text{where } J_n^*=\arg\min_j\{P_{n,j}: j\in C_n\setminus\{I_n\}\} \quad (60a)$$

$$\Delta y_{n,j}=0,\forall j\in C_n \quad \text{Otherwise} \quad (60b)$$

For $|\mathcal{N}|$ samples, the optimal solution will be selecting top Γ samples with the lowest $\left(\log(P_{n,J_n^*})-\log(P_{n,I_n})\right)$ to assign $\Delta y_{n,I_n}=-1, \Delta y_{n,J_n^*}=1$, and others to 0. Define the set of these selected top Γ samples as $\mathcal{N}^{\text{Top}}(\boldsymbol{\beta};\Gamma)$. The optimal objective function of the inner minimization problem will be

$$\begin{aligned}R(\boldsymbol{\beta};\Gamma)&=\min_{\Delta\boldsymbol{y}\in\mathcal{U}(\Gamma)}\sum_{n\in\mathcal{N}}\sum_{i\in C_n}\Delta y_{n,i}\cdot\log(P_{n,i}(\boldsymbol{\beta}))\\&=\sum_{n\in\mathcal{N}^{\text{Top}}(\boldsymbol{\beta};\Gamma)}\log(P^*_{n,J_n})-\log(P_{n,I_n}).\end{aligned} \quad (61)$$

Then Eq. (59) can be rewritten as:

$$\max_{\beta\in\mathcal{B}}\left(\sum_{n\in\mathcal{N}}\sum_{i\in C_n} y_{n,i}\cdot\log(P_{n,i}(\boldsymbol{\beta}))\right)+R(\boldsymbol{\beta};\Gamma). \quad (62)$$

Therefore, $R(\boldsymbol{\beta};\Gamma)$ can be treated as a regularization term. It will penalize extreme value of $\boldsymbol{\beta}$ because that could make $P_{n,J_n^*}\to 0$ and $R(\boldsymbol{\beta};\Gamma)\to-\infty$. The maximal value of $R(\boldsymbol{\beta};\Gamma)$ is achieved at $\boldsymbol{\beta}=0$ because, in this case, $P_{n,i}=P_{n,j},\forall i,j\in C_n$ and $R(\boldsymbol{\beta};\Gamma)=0$. Therefore, this regularization term tends to shrink $\boldsymbol{\beta}$ toward 0. Thus, the robust-label estimator is biased. More importantly, the bias will not diminish with an increase in samples. Reversely, it will increase because more samples indicate the selection of top Γ samples could have a lower value of $\log(P_{n,J_n^*})-\log(P_{n,I_n})$. Therefore, the robust-label MNL estimator is inconsistent. It is worth noting that when $\Gamma\to\infty$, $R(\boldsymbol{\beta};\Gamma)$ will not go to infinity (thus $\boldsymbol{\beta}$ will not be 0), because the physical meaning of Γ (i.e., the maximum number of label errors) implies it is only effective when $\Gamma\le|\mathcal{N}|$. □

Proposition 4. *Let $\mathcal{I}_{RL}(\cdot)$ and $\mathcal{I}_{MLE}(\cdot)$ be the Fisher information matrix of the robust-label MNL estimator and the original MLE estimator, respectively. We have*

$$Tr\left(\mathcal{I}_{RL}(\hat{\boldsymbol{\beta}}^{RL})\right)\ge Tr\left(\mathcal{I}_{MLE}(\hat{\boldsymbol{\beta}}^{MLE})\right). \quad (63)$$

where $\hat{\boldsymbol{\beta}}^{RL}$ and $\hat{\boldsymbol{\beta}}^{MLE}$ are corresponding estimated parameters. $Tr(\cdot)$ is the trace of a matrix.

Proof. Looking at the rewritten form of the robust-label MNL problem in Eq. (62). We have:

$$\mathcal{I}_{\text{RL}}(\hat{\boldsymbol{\beta}}^{\text{RL}})=\mathcal{I}_{\text{MLE}}(\hat{\boldsymbol{\beta}}^{\text{MLE}})-\nabla^2_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}^{\text{RL}}}R(\boldsymbol{\beta};\Gamma). \quad (64)$$

Notice that $\sum_{n\in\mathcal{N}}\sum_{i\in C_n}\Delta y_{n,i}\cdot\log(P_{n,i}(\boldsymbol{\beta}))$ is jointly concave for both $\Delta\boldsymbol{y}$ and $\boldsymbol{\beta}$. The minimization operator over $\text{Conv}(\mathcal{U}(\Gamma))$ preserve concavity. Therefore, $R(\boldsymbol{\beta};\Gamma)$ is concave. And $-\nabla^2_{\boldsymbol{\beta}}R(\boldsymbol{\beta};\Gamma)$ is positive semi-definite with $\text{Tr}(-\nabla^2_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}^{\text{RL}}}R(\boldsymbol{\beta};\Gamma))\ge 0$. Therefore, $\text{Tr}\left(\mathcal{I}_{\text{RL}}(\hat{\boldsymbol{\beta}}^{\text{RL}})\right)\ge\text{Tr}\left(\mathcal{I}_{\text{MLE}}(\hat{\boldsymbol{\beta}}^{\text{MLE}})\right)$ □

It is worth noting that, compared to the proof of the robust-feature MNL model (Proposition 2), there is no requirement for regularization parameters.

5.3. Out-of-sample prediction performance analysis

We have shown that both robust-feature and label estimators tend to shrink the scale of the estimated $\boldsymbol{\beta}$, and may have lower variance. In this section, we show that when testing data have uncertainties, the introduced errors are larger for a larger scale of $\boldsymbol{\beta}$.

Consider a probabilistic classification problem for the bias–variance decomposition. Define $\sigma(\boldsymbol{x}|\boldsymbol{\beta})_i:=\frac{\exp(\boldsymbol{\beta}_i^\top\boldsymbol{x})}{\sum_{j\in C}\exp(\boldsymbol{\beta}_j^\top\boldsymbol{x})}$. Given a testing sample $\boldsymbol{x}$, its true probability of choosing alternative i is P_i, where $P_i(\boldsymbol{x})=\sigma(\boldsymbol{x}|\boldsymbol{\beta}^{\text{True}})_i+\epsilon_i$. ϵ_i are random errors with zero mean. For a given estimated parameter $\hat{\boldsymbol{\beta}}$, define the estimated probabilities as $\hat{P}_i(\boldsymbol{x}|\hat{\boldsymbol{\beta}}):=\sigma(\boldsymbol{x}|\hat{\boldsymbol{\beta}})_i$. Then one could quantify the estimation error of the probabilistic classification problem as $\mathbb{E}\left[\sum_{i\in C}|P_i(\boldsymbol{x})-\hat{P}_i(\boldsymbol{x}\mid\hat{\boldsymbol{\beta}})|\right]=\mathbb{E}\left[\|\boldsymbol{P}(\boldsymbol{x})-\hat{\boldsymbol{P}}(\boldsymbol{x}\mid\hat{\boldsymbol{\beta}})\|_1\right]$, where $\boldsymbol{P}:=(P_i)_{i\in C}$ and $\hat{\boldsymbol{P}}:=(\hat{P}_i)_{i\in C}$.

Proposition 5. *Given a testing sample $\boldsymbol{x}$, assume we can only observe the contaminated data $\tilde{\boldsymbol{x}}=\boldsymbol{x}+\Delta\boldsymbol{x}$, where $\Delta\boldsymbol{x}$ is a random variable. Then, the prediction error for a given estimated $\hat{\boldsymbol{\beta}}$ satisfies:*

$$\mathbb{E}\left[\|\boldsymbol{P}(\boldsymbol{x})-\hat{\boldsymbol{P}}(\tilde{\boldsymbol{x}}\mid\hat{\boldsymbol{\beta}})\|_1\right]\le\underbrace{\mathbb{E}\left[\|\boldsymbol{P}(\boldsymbol{x})-\hat{\boldsymbol{P}}(\boldsymbol{x}\mid\hat{\boldsymbol{\beta}})\|_1\right]}_{\text{Errors without data perturbations}}+L\cdot\mathbb{E}\left[\|\Delta\boldsymbol{x}\|_2\right]\cdot\|\hat{\boldsymbol{\beta}}\|_2, \quad (65)$$

where L is the Lipschitz constant of the softmax function under ℓ_1 norm.

Proof.

$$\begin{aligned}&\mathbb{E}\left[\|\boldsymbol{P}(\boldsymbol{x})-\hat{\boldsymbol{P}}(\tilde{\boldsymbol{x}}\mid\hat{\boldsymbol{\beta}})\|_1\right]\\&\le\mathbb{E}\left[\|\boldsymbol{P}(\boldsymbol{x})-\hat{\boldsymbol{P}}(\boldsymbol{x}\mid\hat{\boldsymbol{\beta}})\|_1\right]+\mathbb{E}\left[\|\hat{\boldsymbol{P}}(\boldsymbol{x}\mid\hat{\boldsymbol{\beta}})-\hat{\boldsymbol{P}}(\tilde{\boldsymbol{x}}\mid\hat{\boldsymbol{\beta}})\|_1\right]\\&=\mathbb{E}\left[\|\boldsymbol{P}(\boldsymbol{x})-\hat{\boldsymbol{P}}(\boldsymbol{x}\mid\hat{\boldsymbol{\beta}})\|_1\right]+\mathbb{E}\left[\|\sigma(\boldsymbol{x}|\hat{\boldsymbol{\beta}})-\sigma(\boldsymbol{x}+\Delta\boldsymbol{x}|\hat{\boldsymbol{\beta}})\|_1\right]\\&\le\mathbb{E}\left[\|\boldsymbol{P}(\boldsymbol{x})-\hat{\boldsymbol{P}}(\boldsymbol{x}\mid\hat{\boldsymbol{\beta}})\|_1\right]+L\cdot\mathbb{E}\left[\|(\hat{\boldsymbol{\beta}}_i^\top\Delta\boldsymbol{x})_{i\in C}\|_1\right]\\&\le\mathbb{E}\left[\|\boldsymbol{P}(\boldsymbol{x})-\hat{\boldsymbol{P}}(\boldsymbol{x}\mid\hat{\boldsymbol{\beta}})\|_1\right]+L\cdot\mathbb{E}\left[\|\Delta\boldsymbol{x}\|_2\right]\cdot\|\hat{\boldsymbol{\beta}}\|_2,\end{aligned} \quad (66)$$

where the first inequality is from the triangle inequality. The second inequality is from the Lipschitz continuity of the softmax function under the ℓ_1 norm. L is the corresponding constant. The last inequality is from the Cauchy–Schwarz inequality. □

From Proposition 5, we see that the upper bound of prediction errors when data has uncertainties has two parts: one is the error without data perturbations and another one is proportional to the scale of estimated $\boldsymbol{\beta}$. Since robust estimators tend to shrink the scale of $\boldsymbol{\beta}$, they are preferred for cases with data perturbations.

Note that the error without data perturbations follows the typical bias–variances trade-off (Kohavi et al., 1996; Vapnik, 1999). The robust estimators have higher bias and usually smaller variances. It could potentially improve the model's generalizability with proper selection of ρ_n like regularization in machine learning. Details on the bias–variances trade-off analysis can be found in Appendix C

6. Numerical experiments

6.1. Experiments design

6.1.1. Binary case study

The robust BNL is evaluated on the Singapore first- and last-mile travel mode choice data set (Mo et al., 2018). The data set is part of Singapore's Household Interview Travel Survey (HITS) in 2012. The survey collects information on travel characteristics as well as individual sociodemographics. The first and last-mile trips are extracted

from the whole trip train in the HITS. Besides travel characteristics, the built environment information is also included as they highly impact the mode choices. Data collection details can be found in Mo et al. (2018). The alternative travel modes for the first/last mile trips are walk and bus.

The whole data set contains a total of more than 24,000 observations. In the case study, we randomly select 1000 samples as the training data set and another 1000 samples as the testing set. Denote the raw training and testing data set as D^{Train} and D^{Test}, respectively. In order to simulate data uncertainties, we generate the synthetic testing data set with errors as the following:

- Step 1: Train a conventional BNL model in D^{Test} and assume that the obtained parameters $\boldsymbol{\beta}^{\text{Test}}$ are the "true" behavior mechanism that individuals will follow.
- Step 2: For each individual n in the testing data set, generate the synthetic choice $\hat{\boldsymbol{y}}_n$ based on $\boldsymbol{x}_n$ and $\boldsymbol{\beta}^{\text{Test}}$ (i.e., calculate the choice probabilities and randomly select one alternative based on the probabilities).
- Step 3: Add artificial errors to the generated data to get the final synthetic testing set with errors: $\tilde{\boldsymbol{y}}_n = \hat{\boldsymbol{y}}_n + \boldsymbol{\Delta y}$ and $\tilde{\boldsymbol{x}}_n = \boldsymbol{x}_n + \boldsymbol{\Delta x}$. Let the synthetic testing data set with errors be $\tilde{D}^{\text{Test}}$.

Specifically, the random errors $\boldsymbol{\Delta x}$ and $\boldsymbol{\Delta y}$ are generated as follows. $\boldsymbol{\Delta x}$ are drawn from a uniform distribution $\mathbb{U}[-0.3\bar{\boldsymbol{x}}, 0.3\bar{\boldsymbol{x}}]$, where $\bar{\boldsymbol{x}} = \sum_{n\in\mathcal{N}} \boldsymbol{x}_n/|\mathcal{N}|$ is the average value of features (i.e., we perturb the features by maximally 30%). $\boldsymbol{\Delta y}$ is a perturbation to the labels such that, with 10% probability, the label $\boldsymbol{y}_n$ is replaced by a randomly-selected alternative in C_n.

"Walk" is set as the base mode. All models will be trained or estimated in the training data set D^{Train}, and tested in the synthetic testing data $\tilde{D}^{\text{Test}}$ to evaluate their performances. The data generation process is replicated 30 times and all models are trained and evaluated in those 30 replications to reduce the impact of randomness.

6.1.2. Multinomial case study

The robust MNL is evaluated on the Swissmetro stated preference survey data set (Bierlaire et al., 2001). The survey aims to analyze the impact of travel modal innovation in transportation, represented by the Swissmetro, a revolutionary maglev underground system, against the usual transport modes represented by car and train. The data contains 1004 individuals with 9036 responses. Users are asked to select from three travel modes (train, car, and Swissmetro) given the corresponding travel attributes. "Train" is set as the base mode. The training and testing data set are generated in the same way as the binary case study with 1000 randomly selected samples for both training and testing data sets.

6.1.3. Definition of metrics

Two major metrics are reported in the case study: log-likelihood (LL) and accuracy. Given a data set $D = \{(\boldsymbol{x}_n, \boldsymbol{y}_n) : \forall n \in \mathcal{N}_D\}$, $\mathcal{N}_D$ is the set of sample index of the data set D. We have:

$$\text{Accuracy}(D) = \frac{\mathbb{1}\left(\arg\max_{i\in C_n}\{\sigma(\boldsymbol{x}_n|\hat{\boldsymbol{\beta}})_i\} = \arg\max_{i\in C_n}\{y_{n,i}\}\right)}{|\mathcal{N}_D|} \tag{67}$$

$$\text{Log-likelihood}(D) = \sum_{n\in\mathcal{N}}\sum_{i\in C_n} y_{n,i}\cdot\log\left(\sigma(\boldsymbol{x}_n|\hat{\boldsymbol{\beta}})_i\right) \tag{68}$$

where $\mathbb{1}(\cdot)$ is the indicator function (equal to 1 if true, otherwise 0). We will evaluate the accuracy and log-likelihood for both D^{Train} and $\tilde{D}^{\text{Test}}$. Note that the robust formulations are only used to estimate parameters. Once the parameters are obtained. Log-likelihood is evaluated on its original definition.

6.2. Experiment results

6.2.1. Binary case study

In the case study, we set $p = 2$, thus q also equals 2. The final robustness term for robust BNL model (Eq. (18)) is $\rho_n\|\boldsymbol{\beta}_{I_n} - \boldsymbol{\beta}_{J_n}\|_2$. We also assume all individuals have the same uncertainty budget ρ_n (i.e., $\rho_n = \rho_m$ for all $n, m \in \mathcal{N}$). Table 2 shows the training and testing accuracy and log-likelihood (LL) with respect to different values of ρ_n and Γ.

We find that, with larger values of ρ_n and Γ, the training accuracy keeps decreasing and training LL becomes smaller, showing worse goodness of fit in the training data set. This is as expected, because the objective functions are weighted more on the robustness term. However, in the testing set, the both robust feature and label models perform better than the typical binary logit model. The best robust feature BNL ($\rho_n = 0.1$) has a testing accuracy of 0.890 and the best robust label BNL ($\Gamma = 1.5$) has a testing accuracy of 0.882, while the binary logit model's testing accuracy is 0.879. The improvement of testing LL is even higher for robust models.

As shown in Section 5, the prediction errors of the robust model can be decomposed into 3 components: (1) bias, (2) variances, and (3) errors from data. The robust models have higher bias, (likely) lower variance, and lower errors from data. Therefore, its performance will be inherently determined by the value of ρ_n and Γ. If ρ_n and Γ are too large, we will have too conservative uncertainty sets, and the bias term will be too large, resulting in higher prediction errors of the robust models. With proper values of ρ_n and Γ, we could leverage the lower variances and lower errors from data to offset the higher bias and get better prediction performance.

The estimated $\boldsymbol{\beta}$ values of selected features are compared in Table 3. Walk time and bus in-vehicle time (IVT) are alternative-specific variables. Walking is set as the base mode. Hence, the alternative specific constant (ASC bus), distance to subway station (Dis. to sub.), and bus station accessibility (Bus access.) are only put in the utility function of the bus. We find that the estimated $\boldsymbol{\beta}$ of the robust-label DCM works like "regularization", which shrinks the estimated $\boldsymbol{\beta}$ value toward 0 (compared to the binary logit model) as we expected in Proposition 1. This provides another explanation of the good performance for out-of-sample prediction for the robust BNL model: To achieve robustness under data uncertainties, the model tends to estimate "smaller" (absolute) values of $\boldsymbol{\beta}$. The "smaller" $\boldsymbol{\beta}$ has better generalizability toward predicting samples that have different patterns with the training data set (as discussed in Proposition 5). This mechanism is similar to how regularization works in machine learning studies.

6.2.2. Multinomial case study

Similar to the binary case study, we set $p = 2$ (thus $q = 2$) and make the robustness term for the robust MNL model (Eq. (33)) be $\rho_n\|\boldsymbol{\beta}_j - \boldsymbol{\beta}_{I_n}\|_2$. The results in terms of different values of ρ_n and Γ are shown in Table 4. Note that the parameter sets are different from the binary case study as the data sets are different.

The results are similar to what we observe in the binary cases. In general, with reasonable settings of the robust budget parameter ρ_n and Γ, we observe better testing accuracy and LL. The best robust feature model ($\rho_n = 0.1$) has a testing accuracy of 0.565 and the best robust label model ($\Gamma = 100$) has a testing accuracy of 0.558. While the conventional MNL model's testing accuracy is 0.540. It is worth noting that, compared to the binary case study, the best hyper-parameter for robust label models can be quite different. This implies that hyper-parameter tuning may be necessary for implementing the robust MNL model.

The estimated $\boldsymbol{\beta}$ values of alternative specific features (i.e., cost and travel time) are shown in Table 5. Similar to binary cases, we observe the shrinkage of parameter scales with the increasing robust budget parameters. This is consistent with our analysis in Remark 4. Note that

Table 2
Results for robust BNL models.

Model	Parameters	Training accuracy	Training LL	Testing accuracy	Testing LL
Binary Logit	–	0.957 (±0.010)	−152.6 (±22.6)	0.879 (±0.014)	−358.0 (±77.1)
Robust-feature	$\rho_n = 0.001$	0.957 (±0.010)	−152.6 (±22.6)	0.880 (±0.013)	−353.8 (±75.5)
	$\rho_n = 0.01$	0.957 (±0.010)	−153.1 (±22.5)	0.884 (±0.014)	−337.5 (±67.6)
	$\rho_n = 0.1$	0.951 (±0.009)	−163.3 (±22.0)	0.890 (±0.012)	−289.3 (±33.0)
	$\rho_n = 0.2$	0.940 (±0.008)	−188.8 (±21.8)	0.888 (±0.011)	−273.2 (±21.9)
	$\rho_n = 0.3$	0.932 (±0.008)	−230.7 (±21.7)	0.883 (±0.012)	−287.3 (±19.0)
Robust-label	$\Gamma = 1$	0.954 (±0.009)	−157.4 (±21.6)	0.881 (±0.012)	−313.7 (±43.8)
	$\Gamma = 1.5$	0.954 (±0.009)	−160.3 (±21.2)	0.882 (±0.012)	−307.4 (±38.4)
	$\Gamma = 2$	0.953 (±0.009)	−163.6 (±20.9)	0.881 (±0.011)	−303.8 (±34.1)
	$\Gamma = 2.5$	0.953 (±0.009)	−167.0 (±20.6)	0.879 (±0.011)	−302.3 (±31.9)
	$\Gamma = 3$	0.952 (±0.009)	−170.5 (±20.4)	0.878 (±0.012)	−301.7 (±30.5)

– All results are the average values of 30 replications. Values in the parentheses are standard deviations.
– Best models are highlighted with gray cells.

Table 3
Selected estimated β for robust BNL models.

Model	Parameters	Walk time	Bus IVT	Dist. to sub.	Bus access.	ASC bus
Binary Logit	–	−3.21 (±0.56)	−5.29 (±1.11)	6.17 (±1.61)	0.95 (±0.91)	−6.97 (±2.36)
Robust-feature	$\rho_n = 0.001$	−3.21 (±0.55)	−5.29 (±1.11)	6.15 (±1.60)	0.92 (±0.89)	−6.72 (±2.25)
	$\rho_n = 0.01$	−3.19 (±0.52)	−5.19 (±1.06)	5.98 (±1.54)	0.69 (±0.78)	−5.35 (±1.58)
	$\rho_n = 0.1$	−3.10 (±0.30)	−3.79 (±0.65)	3.90 (±0.84)	−0.48 (±0.25)	−2.54 (±0.57)
	$\rho_n = 0.2$	−2.62 (±0.23)	−2.32 (±0.33)	2.19 (±0.34)	−0.70 (±0.11)	−1.39 (±0.25)
	$\rho_n = 0.3$	−1.93 (±0.16)	−1.33 (±0.17)	1.32 (±0.15)	−0.56 (±0.07)	−0.80 (±0.12)
Robust-label	$\Gamma = 1$	−2.79 (±0.45)	−3.99 (±0.73)	4.52 (±1.03)	0.80 (±0.61)	−6.05 (±1.69)
	$\Gamma = 1.5$	−2.63 (±0.41)	−3.68 (±0.66)	4.11 (±0.93)	0.78 (±0.58)	−5.88 (±1.64)
	$\Gamma = 2$	−2.50 (±0.38)	−3.44 (±0.60)	3.77 (±0.86)	0.77 (±0.58)	−5.72 (±1.61)
	$\Gamma = 2.5$	−2.39 (±0.36)	−3.23 (±0.56)	3.49 (±0.80)	0.77 (±0.57)	−5.69 (±1.51)
	$\Gamma = 3$	−2.30 (±0.35)	−3.06 (±0.52)	3.25 (±0.76)	0.76 (±0.57)	−5.61 (±1.44)

– All results are the average values of 30 replications. Values in the parentheses are standard deviations.

Table 4
Results for robust MNL models.

Model	Parameters	Training accuracy	Training LL	Testing accuracy	Testing LL
MNL	–	0.591 (±0.014)	−871.5 (±15.3)	0.540 (±0.021)	−988.5 (±41.5)
Robust-feature	$\rho_n = 0.001$	0.590 (±0.014)	−871.6 (±15.3)	0.542 (±0.022)	−981.5 (±39.5)
	$\rho_n = 0.01$	0.588 (±0.014)	−874.7 (±15.3)	0.552 (±0.021)	−952.2 (±33.2)
	$\rho_n = 0.1$	0.585 (±0.013)	−923.7 (±15.2)	0.565 (±0.019)	−942.0 (±16.8)
	$\rho_n = 0.15$	0.578 (±0.018)	−966.7 (±13.5)	0.558 (±0.018)	−975.8 (±15.0)
	$\rho_n = 0.2$	0.503 (±0.037)	−1001.6 (±10.6)	0.519 (±0.029)	−1007.8 (±14.0)
Robust-label	$\Gamma = 1$	0.589 (±0.013)	−873.9 (±15.1)	0.544 (±0.022)	−972.6 (±36.7)
	$\Gamma = 10$	0.583 (±0.014)	−886.5 (±15.5)	0.555 (±0.022)	−937.5 (±26.2)
	$\Gamma = 100$	0.579 (±0.013)	−941.5 (±13.1)	0.558 (±0.020)	−956.1 (±13.3)
	$\Gamma = 200$	0.582 (±0.015)	−979.7 (±11.0)	0.557 (±0.022)	−987.3 (±9.9)
	$\Gamma = 300$	0.586 (±0.015)	−1007.2 (±9.5)	0.556 (±0.021)	−1012.9 (±8.7)

– All results are the average values of 30 replications. Values in the parentheses are standard deviations.
– Best models are highlighted with gray cells.

for the robust-label optimization, when Γ goes to $+\infty$, β will not be 0 (see Remark 4). The final values will depend on the data set.

6.3. Impact of error generation

The model is derived based on the assumption that all perturbations $\boldsymbol{\Delta x}_n$ are independent across individuals. It is worth exploring model performance if $\boldsymbol{\Delta x}_n$ are correlated given some system errors. In this section, we consider the following system errors to generate the testing data set.

- **Over-reporting of travel time**: For features of travel time, their corresponding $\boldsymbol{\Delta x}$ are drawn from a uniform distribution $\mathbb{U}[0, 0.3\bar{\boldsymbol{x}}]$. Other perturbations follow the same settings as in Section 6.1.1.
- **Under-reporting of travel time**. For features of travel time, their corresponding $\boldsymbol{\Delta x}$ are drawn from a uniform distribution $\mathbb{U}[-0.3\bar{\boldsymbol{x}}, 0]$. Other perturbations follow the same settings as in Section 6.1.1.
- **Social desirability bias** (i.e., respondents might over-report environmentally friendly choices like walking or cycling). With 10% probability, the actual selections of non-walk and non-bike modes are replaced by walk or bike modes. Other perturbations follow the same settings as in Section 6.1.1

Since only the Singapore HITS dataset is from a revealed preference survey with user-reported travel time, the HITS dataset is used for testing. The robust-feature models use ℓ_2 norm with $\rho_n = 0.1$ and robust-label models use $\Gamma = 1.5$ as they are the best hyper-parameters in the previous experiment. The prediction performance is shown in Table 6. We saw that though the robust models are derived from independent error assumptions, they also outperform the nominal model when there are systematic errors in the system (i.e., error terms are correlated).

Table 5
Selected estimated β for robust MNL models.

Model	Parameters	Train time	Train cost	Car time	Car cost	SM time	SM cost
MNL	–	−1.34 (±0.15)	−0.13 (±0.04)	−1.27 (±0.33)	−0.10 (±0.03)	−1.08 (±0.25)	−0.50 (±0.28)
Robust-feature	$\rho_n = 0.001$	−1.34 (±0.15)	−0.12 (±0.03)	−1.26 (±0.33)	−0.09 (±0.03)	−1.07 (±0.25)	−0.49 (±0.28)
	$\rho_n = 0.01$	−1.26 (±0.14)	−0.06 (±0.02)	−1.19 (±0.30)	−0.04 (±0.01)	−1.00 (±0.23)	−0.42 (±0.25)
	$\rho_n = 0.1$	−0.69 (±0.08)	0.04 (±0.01)	−0.44 (±0.12)	0.01 (±0.01)	−0.33 (±0.09)	−0.09 (±0.07)
	$\rho_n = 0.15$	−0.45 (±0.06)	0.06 (±0.01)	−0.16 (±0.06)	0.02 (±0.00)	−0.11 (±0.04)	−0.01 (±0.02)
	$\rho_n = 0.2$	−0.29 (±0.04)	0.07 (±0.01)	−0.03 (±0.02)	0.02 (±0.00)	−0.02 (±0.01)	0.00 (±0.00)
Robust-label	$\Gamma = 1$	−1.20 (±0.17)	−0.13 (±0.04)	−1.07 (±0.34)	−0.10 (±0.03)	−0.74 (±0.21)	−0.74 (±0.26)
	$\Gamma = 10$	−0.89 (±0.19)	−0.06 (±0.02)	−0.80 (±0.33)	−0.05 (±0.02)	−0.44 (±0.23)	−0.67 (±0.24)
	$\Gamma = 100$	−0.36 (±0.08)	−0.03 (±0.01)	−0.37 (±0.15)	−0.02 (±0.01)	−0.21 (±0.10)	−0.29 (±0.11)
	$\Gamma = 200$	−0.22 (±0.06)	−0.03 (±0.01)	−0.24 (±0.10)	−0.02 (±0.01)	−0.13 (±0.06)	−0.17 (±0.06)
	$\Gamma = 300$	−0.16 (±0.06)	−0.03 (±0.01)	−0.18 (±0.08)	−0.01 (±0.01)	−0.10 (±0.04)	−0.10 (±0.04)

– All results are the average values of 30 replications. Values in the parentheses are standard deviations.

Table 6
Impact of different data error generations.

Scenarios	Models	Training accuracy	Training LL	Testing accuracy	Testing LL
Independent errors	BNL	0.957 (±0.010)	−152.6 (±22.6)	0.879 (±0.014)	−358.0 (±77.1)
	Robust-feature	0.951 (±0.009)	−163.3 (±22.0)	0.890 (±0.012)	−289.3 (±33.0)
	Robust-label	0.954 (±0.009)	−160.3 (±21.2)	0.882 (±0.012)	−307.4 (±38.4)
Over-reporting	BNL	0.957 (±0.010)	−152.6 (±22.6)	0.906 (±0.013)	−279.8 (±58.6)
	Robust-feature	0.951 (±0.009)	−163.3 (±22.0)	0.917 (±0.011)	−225.6 (±28.6)
	Robust-label	0.954 (±0.009)	−160.3 (±21.2)	0.909 (±0.013)	−247.4 (±37.8)
Under-reporting	BNL	0.957 (±0.010)	−152.6 (±22.6)	0.881 (±0.018)	−326.2 (±70.3)
	Robust-feature	0.951 (±0.009)	−163.3 (±22.0)	0.897 (±0.012)	−256.4 (±27.2)
	Robust-label	0.954 (±0.009)	−160.3 (±21.2)	0.886 (±0.016)	−284.9 (±42.2)
Social desirability	BNL	0.957 (±0.010)	−152.6 (±22.6)	0.860 (±0.015)	−514.0 (±111.2)
	Robust-feature	0.951 (±0.009)	−163.3 (±22.0)	0.871 (±0.013)	−415.3 (±67.2)
	Robust-label	0.954 (±0.009)	−160.3 (±21.2)	0.863 (±0.013)	−425.5 (±63.8)

– All results are the average values of 30 replications. Values in the parentheses are standard deviations.
– "Independent errors" represent the same setting as Section 6.1.1.

Table 7
Impact of different norms on model performance.

Model	Norms	Training accuracy	Training LL	Testing accuracy	Testing LL
MNL	–	0.591 (±0.014)	−871.5 (±15.3)	0.540 (±0.021)	−988.5 (±41.5)
Robust-feature	$q = 2$	0.585 (±0.013)	−923.7 (±15.2)	0.565 (±0.019)	−942.0 (±16.8)
	$q = 3$	0.586 (±0.012)	−907.8 (±15.7)	0.563 (±0.020)	−935.8 (±19.0)
	$q = 4$	0.586 (±0.012)	−902.4 (±15.8)	0.560 (±0.019)	−935.5 (±20.3)
	$q = 10$	0.586 (±0.013)	−896.7 (±15.9)	0.557 (±0.016)	−936.9 (±21.9)
	$q = 100$	0.588 (±0.014)	−894.2 (±14.7)	0.556 (±0.017)	−936.9 (±22.9)
	$q = \infty$	0.587 (±0.014)	−894.7 (±16.0)	0.556 (±0.017)	−937.7 (±22.8)

– All results are the average values of 30 replications. Values in the parentheses are standard deviations.
– Best models are highlighted with gray cells.

6.4. Impact of norms in uncertainty sets

The above case study is based on the ℓ_2 norm for the robust-feature MNL model. In this section, we test how different norms will affect the prediction results. Only the MNL model with Swissmetro data is tested. We fix $\rho_n = 0.1$ for all testing cases as 0.1 is the best hyper-parameter for the ℓ_2 norm. The implementation of ℓ_∞ is to define a new auxiliary decision variable $u_{n,i}$ to replace $\|\boldsymbol{\beta}_j - \boldsymbol{\beta}_{I_n}\|_\infty$, and add $u_{n,i} \geq (\boldsymbol{\beta}_i - \boldsymbol{\beta}_{I_n})^{(k)}$ and $u_{n,i} \geq -(\boldsymbol{\beta}_i - \boldsymbol{\beta}_{I_n})^{(k)}$ $\forall k \in \mathcal{K}$ to the constraints. The prediction performance is shown in Table 7. From the results, we see that the robust-feature models always outperform the nominal MNL model in out-of-sample prediction accuracy, no matter what norm we use. Since $\rho_n = 0.1$ is the best hyper-parameter for the ℓ_2 norm. ℓ_2 norm still gives the best testing accuracy. Note that for other norms, hyper-parameter tuning is needed to obtain better performance than current results.

6.5. Implementation of robust DCMs in pricing problem

DCMs are used to for many transportation decision-making problems. Pricing or fare policy design is one of the most important decision-making problems. In this section, we will test how the robust-feature MNL model performs in a typical pricing problem. Consider the Swissmetro dataset, assuming the government wishes to use the survey results to decide the best fare policy for the Swissmetro so as to maximize the revenue (i.e., price × demand). Given the same definition as Section 6.1.1, assuming we wish to maximize the revenue of the 1000 samples in $\mathcal{D}^{\text{Test}}$. Denote the corresponding sample index set as $\mathcal{N}_{\mathcal{D}^{\text{Test}}}$. But we only observe their perturbed attributes $\tilde{\boldsymbol{x}}_{\boldsymbol{n}}(\alpha) = (\tilde{\boldsymbol{x}}_{\boldsymbol{n}}^{\text{NoSMCost}}, \alpha \cdot x_n^{\text{SMCost}})$, where $\tilde{\boldsymbol{x}}_{\boldsymbol{n}}^{\text{NoSMCost}} \in \mathbb{R}^{|\mathcal{K}|-1}$ is the perturbed attributes vector excluding the Swissmetro cost. $x_n^{\text{SMCost}} \in \mathbb{R}$ is the original cost of Swissmetro (not perturbed). α is the scaler for the price of the ticket, which is our decision variable. Given an estimated parameter $\hat{\boldsymbol{\beta}}$, we wish to solve the following problem to maximize the revenue:

$$\max_{\alpha \geq 0} \sum_{n \in \mathcal{N}_{\mathcal{D}^{\text{Test}}}} \alpha \cdot x_n^{\text{SMCost}} \cdot \frac{\exp((\hat{\boldsymbol{\beta}}_{\text{SM}})^\top \tilde{\boldsymbol{x}}_{\boldsymbol{n}}(\alpha))}{\sum_{j \in \mathcal{C}_n} \exp\left(\hat{\boldsymbol{\beta}}_j^\top \tilde{\boldsymbol{x}}_{\boldsymbol{n}}(\alpha)\right)} \tag{69}$$

where $\hat{\boldsymbol{\beta}}_{\text{SM}}$ is the estimated parameters for Swissmetro. The problem is a non-convex optimization. However, since there is one single decision variable α, we can solve it through brute-force searching within $0 \leq$

Table 8
Impact of different norms on model performance.

Model	Parameters	Objective	Optimal solution (α^*)	Revenue
MNL	–	2135.70 (±247.54)	2.51 (±3.90)	1931.37 (±273.12)
Robust-feature	$\rho_n = 0.01$	2105.58 (±230.14)	4.35 (±4.78)	1985.99 (±276.94)
Robust-label	$\Gamma = 3$	2076.27 (±166.01)	3.44 (±4.46)	1978.73 (±275.83)
Oracle	–	2067.53 (±208.25)	3.40 (±4.49)	2067.53 (±208.25)

– All results are the average values of 30 replications. Values in the parentheses are standard deviations.

$\alpha \leq 10$. Assuming α^* is the optimal solution of Eq. (69). The final actual revenue will be evaluated using the "true" behavior mechanism $\boldsymbol{\beta}^{\text{Test}}$ (see details in Section 6.1.1) and the non-perpetuated attributes $\boldsymbol{x}_n$.

$$\text{Revenue} = \sum_{n \in \mathcal{N}_{D^{\text{Test}}}} \alpha \cdot x_n^{\text{SMCost}} \cdot \frac{\exp((\boldsymbol{\beta}_{\text{SM}}^{\text{Test}})^\top \boldsymbol{x}_n(\alpha))}{\sum_{j \in C_n} \exp\left((\boldsymbol{\beta}_j^{\text{Test}})^\top \boldsymbol{x}_n(\alpha)\right)} \tag{70}$$

The final results are summarized in Table 8. "Oracle" scenario is obtained by feeding $\boldsymbol{\beta}^{\text{Test}}$ and $\boldsymbol{x}_n(\alpha)$ to Eq. (69), which represents the upper bound of the revenue. The results show that given better predictability, the robust models also outperform the MNL model in decision-making problems.

7. Conclusion

In this paper, we propose a robust BNL and MNL model framework that accommodates testing data uncertainties. The goal is to enhance the model's prediction accuracy in new samples as a classification problem. Our model is rooted in the theory of robust optimization. Specifically, we address feature uncertainties by assuming the ℓ_p-norm of measurement errors are below a predetermined threshold. For label uncertainties, we assume there are at most Γ mislabeled choices as the uncertainty set. Under these assumptions, we derive tractable robust counterparts for both robust-feature and robust-label DCMs. The proposed models are validated in both binary and multinomial choice data sets. The results demonstrate that the robust models outperform the conventional BNL and MNL models in terms of prediction accuracy and log-likelihood when there are measurement errors in testing data. Our findings suggest that the robustness component functions as "regularization", leading to improved generalizability of the models.

There are several future extensions for the current model. First, it is possible to combine the robust-label and the robust-feature models by assuming an uncertainty set with both feature and label measurement errors. Future research may work on deriving the closed-form formulations for the integrated model. Second, The performance of robust models may depend on the hyper-parameters. Future studies may develop methods to automatically tune the hyper-parameters. Third, the uncertainty set defined in this study is decomposable for each individual n (i.e., $\mathcal{Z}(\boldsymbol{\rho}) = \prod_{n \in \mathcal{N}} \mathcal{Z}_n(\rho_n)$). One may also add constraints for the total errors $\|(\boldsymbol{\Delta x}_n)_{n \in \mathcal{N}}\|_p$. However, this will break the tractability for deriving the closed-form formulation (but still solvable). Future studies may explore how different definitions of the uncertainty set impact the prediction performance of robust DCMs.

CRediT authorship contribution statement

Baichuan Mo: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Yunhan Zheng:** Writing – original draft, Methodology, Formal analysis, Conceptualization. **Xiaotong Guo:** Methodology, Conceptualization. **Ruoyun Ma:** Software, Data curation. **Jinhua Zhao:** Project administration, Funding acquisition.

Acknowledgment

This research is supported by the National Research Foundation (NRF), Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme. The Mens, Manus, and Machina (M3S) is an interdisciplinary research group (IRG) of the Singapore-MIT Alliance for Research and Technology (SMART) centre.

Appendix A. Inequality gap for the robust-feature MNL approximation

Define $\boldsymbol{\Delta x}_n^* := \arg\max_{\boldsymbol{\Delta x}_n \in \mathcal{Z}_n} \log \sum_{j \in C_n} \exp\big((\boldsymbol{\beta}_j - \boldsymbol{\beta}_{I_n})^\top (\boldsymbol{x}_n + \boldsymbol{\Delta x}_n)\big)$. Then the gap of the Jensen's inequality in Eq. (30) is :

$$\text{Gap}_n = (\text{RHS - LFS}) = \log \frac{\sum_{j \in C_n} \exp\big((\boldsymbol{\beta}_j - \boldsymbol{\beta}_{I_n})^\top \boldsymbol{x}_n + \rho_n \cdot \|\boldsymbol{\beta}_j - \boldsymbol{\beta}_{I_n}\|_q\big)}{\sum_{j \in C_n} \exp\big((\boldsymbol{\beta}_j - \boldsymbol{\beta}_{I_n})^\top (\boldsymbol{x}_n + \boldsymbol{\Delta x}_n^*)\big)} \tag{A.1}$$

So the gap will be zero if $\rho_n \|\boldsymbol{\beta}_j - \boldsymbol{\beta}_{I_n}\|_q = (\boldsymbol{\beta}_j - \boldsymbol{\beta}_{I_n})^\top \boldsymbol{\Delta x}_n^* \quad \forall j \in C_n$, which implies $\boldsymbol{\beta}_j = \boldsymbol{\beta}_i$ for all $i, j \in C_n$. In the specification of DCM, one of the alternatives will have a fixed coefficient of feature $k \in \mathcal{K}$ to be zero (to avoid perfect co-linearity). Therefore, $\exists i \in C$ such that $\boldsymbol{\beta}_i, (k) = 0, \ \forall k \in \mathcal{K}$. Then $\boldsymbol{\beta}_j = \boldsymbol{\beta}_i$ is equivalent to $\boldsymbol{\beta} = 0$. Since larger ρ_n will shrink $\boldsymbol{\beta}$ toward zero, we know that the gap tends to be small when ρ_n is large. Another condition where the gap is close to zero is that one $\boldsymbol{\beta}_i - \boldsymbol{\beta}_{I_n}$ is significantly larger than others. (i.e., $\exists i^*$ such that $(\boldsymbol{\beta}_{i^*} - \boldsymbol{\beta}_{I_n})^\top (\boldsymbol{x}_n + \boldsymbol{\Delta x}_n) \gg (\boldsymbol{\beta}_j - \boldsymbol{\beta}_{I_n})^\top (\boldsymbol{x}_n + \boldsymbol{\Delta x}_n), j \neq i^*$. In this case, we can ignore other terms in the summation and the RHS and LHS will be the same.

Now let us find an upper bound for the gap. Notice that:

$$\rho_n \cdot \|\boldsymbol{\beta}_j - \boldsymbol{\beta}_{I_n}\|_q \leq \rho_n \cdot \max_{i \in C_n} \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_{I_n}\|_q, \quad \forall i \in C_n \tag{A.2}$$

$$\sum_{j \in C_n} \exp\big((\boldsymbol{\beta}_j - \boldsymbol{\beta}_{I_n})^\top (\boldsymbol{x}_n + \boldsymbol{\Delta x}_n^*)\big) \geq \sum_{j \in C_n} \exp\big((\boldsymbol{\beta}_j - \boldsymbol{\beta}_{I_n})^\top \boldsymbol{x}_n\big) \tag{A.3}$$

where the second inequality is obtained by setting $\boldsymbol{\Delta x}_n = 0$. Substituting Eqs. (A.2) and (A.3) to Eq. (A.1), we have

$$\text{Gap}_n \leq |C_n| \cdot \rho_n \cdot \max_{i \in C_n} \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_{I_n}\|_q \tag{A.4}$$

Therefore, we have the property that when ρ_n is small, the gap is also small. Combining the analysis above, the gap for the approximation will not become extreme no matter what the values of ρ_n are. When ρ_n is large, $\boldsymbol{\beta}$ will be close to 0, which helps to lower the gap. When ρ_n is small, Eq. (A.4) implies the gap is small as well.

Appendix B. Upper bound for the robust-feature MNL problem

The upper bound of the robust MNL problem can be obtained by getting the lower bound of the LHS in Eq. (30). Notice that

$$\sum_{j \in C_n} \exp\big((\boldsymbol{\beta}_j - \boldsymbol{\beta}_{I_n})^\top (\boldsymbol{x}_n + \boldsymbol{\Delta x}_n)\big) \geq \max_{i \in C_n} \exp\big((\boldsymbol{\beta}_i - \boldsymbol{\beta}_{I_n})^\top (\boldsymbol{x}_n + \boldsymbol{\Delta x}_n)\big) \tag{B.1}$$

The inequality is because the sum of all positive terms is at least as large as any single term (including the maximum one). Take the logarithm of both sides:

$$\log \sum_{j \in C_n} \exp\big((\boldsymbol{\beta}_j - \boldsymbol{\beta}_{I_n})^\top (\boldsymbol{x}_n + \boldsymbol{\Delta x}_n)\big) \geq \max_{i \in C_n} \big((\boldsymbol{\beta}_i - \boldsymbol{\beta}_{I_n})^\top (\boldsymbol{x}_n + \boldsymbol{\Delta x}_n)\big) \tag{B.2}$$

Then we maximize over $\boldsymbol{\Delta x}_n$ on both sides:

$$\max_{\boldsymbol{\Delta x}_n \in \mathcal{Z}_n} \log \sum_{j \in C_n} \exp\big((\boldsymbol{\beta}_j - \boldsymbol{\beta}_{I_n})^\top (\boldsymbol{x}_n + \boldsymbol{\Delta x}_n)\big)$$

$$\geq \max_{i\in C_n}\Big(\max_{\Delta \boldsymbol{x}_n\in Z_n}(\boldsymbol{\beta}_i-\boldsymbol{\beta}_{I_n})^\top(\boldsymbol{x}_n+\Delta \boldsymbol{x}_n)\Big)$$
$$= \max_{i\in C_n}\left((\boldsymbol{\beta}_i-\boldsymbol{\beta}_{I_n})^\top \boldsymbol{x}_n+\rho_n\|\boldsymbol{\beta}_i-\boldsymbol{\beta}_{I_n}\|_q\right) \tag{B.3}$$

Therefore, Eq. (29) can be approximated as:

$$\sum_{n\in\mathcal{N}}\left((\boldsymbol{\beta}_i-\boldsymbol{\beta}_{I_n})^\top \boldsymbol{x}_n+\rho_n\|\boldsymbol{\beta}_i-\boldsymbol{\beta}_{I_n}\|_q\right)\leq -t \quad \forall i\in C \tag{B.4}$$

Therefore, an upper bound of the robust-feature MNL problem is:

$$\max_{\boldsymbol{\beta}\in\mathcal{B},\, t} \quad t \tag{B.5a}$$
$$\text{s.t.} \sum_{n\in\mathcal{N}}\left((\boldsymbol{\beta}_i-\boldsymbol{\beta}_{I_n})^\top \boldsymbol{x}_n+\rho_n\|\boldsymbol{\beta}_i-\boldsymbol{\beta}_{I_n}\|_q\right)\leq -t \quad \forall i\in C \tag{B.5b}$$

Appendix C. Bias–variance trade off for errors without perturbations

The expected errors without data perturbation can be decomposed as bias and variances:

$$\mathbb{E}\left[\|\boldsymbol{P}(\boldsymbol{x})-\hat{\boldsymbol{P}}(\boldsymbol{x}\mid\hat{\boldsymbol{\beta}})\|_1\right]=\mathbb{E}\left[\|\boldsymbol{P}(\boldsymbol{x})-\mathbb{E}[\hat{\boldsymbol{P}}(\boldsymbol{x}\mid\hat{\boldsymbol{\beta}})]+\mathbb{E}[\hat{\boldsymbol{P}}(\boldsymbol{x}\mid\hat{\boldsymbol{\beta}})]-\hat{\boldsymbol{P}}(\boldsymbol{x}\mid\hat{\boldsymbol{\beta}})\|_1\right]$$
$$\leq \underbrace{\|\boldsymbol{P}(\boldsymbol{x})-\mathbb{E}[\hat{\boldsymbol{P}}(\boldsymbol{x}\mid\hat{\boldsymbol{\beta}})]\|_1}_{\text{Bias term}}+\underbrace{\mathbb{E}\left[\|\hat{\boldsymbol{P}}(\boldsymbol{x}\mid\hat{\boldsymbol{\beta}})-\mathbb{E}[\hat{\boldsymbol{P}}(\boldsymbol{x}\mid\hat{\boldsymbol{\beta}})]\|_1\right]}_{\text{Variance term}}, \tag{C.1}$$

where the second inequality is derived from the triangle inequality. Note that typical bias–variance trade off are derived from the squared errors. For the ℓ_1 norm, the "variance" term here is essentially a measure of dispersion or variability, often referred to as expected deviation or mean absolute deviation from the predictor. We keep the name of "variance" here to be consistent with the literature.

For the proposed robust MNL estimator, as it is biased (Propositions 1 and 3). The bias term will be larger than the nominal MLE estimator. However, as we show it has higher trace of the Fisher information matrix (Propositions 2 and 4), the variance term will be smaller than the MLE estimator. Therefore, the final expected errors without data perturbation will depend on the selection of ρ_n. A larger value of ρ_n will increase the bias and decrease the variances.

References

Baiocchi, M., Cheng, J., & Small, D. S. (2014). Instrumental variable methods for causal inference. *Statistics in Medicine*, *33*(13), 2297–2340.

Ben-Akiva, M., & Bierlaire, M. (1999). Discrete choice methods and their applications to short term travel decisions. In *Handbook of transportation science* (pp. 5–33). Springer.

Ben-Akiva, M. E., & Lerman, S. R. (1985). *vol. 9, Discrete choice analysis: theory and application to travel demand*. MIT Press.

Ben-Moshe, D. (2014). Identification of dependent nonparametric distributions in a system of linear equations. *Maurice Falk Institute for Economic Research in Israel. Discussion Paper Series*, (3), 2A.

Ben-Tal, A., Den Hertog, D., & Vial, J.-P. (2015). Deriving robust counterparts of nonlinear uncertain inequalities. *Mathematical Programming*, *149*(1), 265–299.

Ben-Tal, A., & Nemirovski, A. (1998). Robust convex optimization. *Mathematics of Operations Research*, *23*(4), 769–805.

Ben-Tal, A., & Nemirovski, A. (1999). Robust solutions of uncertain linear programs. *Operations Research Letters*, *25*(1), 1–13.

Bertsimas, D., Dunn, J., Pawlowski, C., & Zhuo, Y. D. (2019). Robust classification. *INFORMS Journal on Optimization*, *1*(1), 2–34.

Bertsimas, D., & Hertog, D. d. (2022). *Robust and adaptive optimization*. Dynamic Ideas LLC.

Bertsimas, D., Iancu, D. A., & Parrilo, P. A. (2010). Optimality of affine policies in multistage robust optimization. *Mathematics of Operations Research*, *35*(2), 363–394.

Bierlaire, M., Axhausen, K., & Abay, G. (2001). The acceptance of modal innovation: The case of swissmetro. In *Swiss transport research conference*.

Bowman, J. L., & Ben-Akiva, M. E. (2001). Activity-based disaggregate travel demand model system with activity schedules. *Transportation Research Part A: Policy and Practice*, *35*(1), 1–28.

Campanelli, P. C., Martin, E. A., & Rothgeb, J. M. (1991). The use of respondent and interviewer debriefing studies as a way to study response error in survey data. *Journal of the Royal Statistical Society: Series D (the Statistician)*, *40*(3), 253–264.

Cramér, H. (2016). *Mathematical methods of statistics (PMS-9), volume 9*. Princeton University Press.

Duchi, J. C., Glynn, P. W., & Namkoong, H. (2021). Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, *46*(3), 946–969.

Duchi, J., & Namkoong, H. (2019). Variance-based regularization with convex objectives. *Journal of Machine Learning Research*, *20*(68), 1–55.

Fernandes, B., Street, A., Valladão, D., & Fernandes, C. (2016). An adaptive robust portfolio optimization model with loss constraints based on data-driven polyhedral uncertainty sets. *European Journal of Operational Research*, *255*(3), 961–970.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.

Gorissen, B. L., Yanıkoğlu, İ., & den Hertog, D. (2015). A practical guide to robust optimization. *Omega*, *53*, 124–137.

Guo, X., Mo, B., Koutsopoulos, H. N., Wang, S., & Zhao, J. (2024). Robust transit frequency setting problem with demand uncertainty. *IEEE Transactions on Intelligent Transportation Systems*.

Hausman, J. (2001). Mismeasured variables in econometric analysis: problems from the right and problems from the left. *Journal of Economic Perspectives*, *15*(4), 57–67.

Hausman, J. A., Abrevaya, J., & Scott-Morton, F. M. (1998). Misclassification of the dependent variable in a discrete-response setting. *Journal of Econometrics*, *87*(2), 239–269.

Hölder, O. (1889). Ueber einen mittelwerthabsatz. *Nachrichten Von der Königl. Gesellschaft der Wissenschaften Und der Georg-Augusts-Universität Zu Göttingen*, *1889*, 38–47.

Hu, Y. (2008). Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution. *Journal of Econometrics*, *144*(1), 27–61.

Hurst, E., Li, G., & Pugsley, B. (2014). Are household surveys like tax forms? Evidence from income underreporting of the self-employed. *Review of Economics and Statistics*, *96*(1), 19–33.

Kohavi, R., Wolpert, D. H., et al. (1996). Bias plus variance decomposition for zero-one loss functions. *96*, In *ICML* (pp. 275–283). Citeseer.

Kreuter, F., Presser, S., & Tourangeau, R. (2008). Social desirability bias in CATI, IVR, and web surveys: The effects of mode and question sensitivity. *Public Opinion Quarterly*, *72*(5), 847–865.

Lanckriet, G. R., Ghaoui, L. E., Bhattacharyya, C., & Jordan, M. I. (2002). A robust minimax approach to classification. *Journal of Machine Learning Research*, *3*(Dec), 555–582.

Li, M., & Xu, H. (2023). A modified late arrival penalised user equilibrium model and robustness in data perturbation. arXiv preprint arXiv:2401.00380.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083.

Mo, Z., Fu, Y., & Di, X. (2022). Quantifying uncertainty in traffic state estimation using generative adversarial networks. In *2022 IEEE 25th international conference on intelligent transportation systems* (pp. 2769–2774). IEEE.

Mo, B., Koutsopoulos, H. N., Shen, Z.-J. M., & Zhao, J. (2023). Robust path recommendations during public transit disruptions under demand uncertainty. *Transportation Research Part B: Methodological*, *169*, 82–107.

Mo, B., Shen, Y., & Zhao, J. (2018). Impact of built environment on first-and last-mile travel mode choice. *Transportation Research Record*, *2672*(6), 40–51.

Mohajerin Esfahani, P., & Kuhn, D. (2018). Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, *171*(1), 115–166.

Musunuru, A., & Porter, R. J. (2019). Applications of measurement error correction approaches in statistical road safety modeling. *Transportation Research Record*, *2673*(8), 125–135.

Paleti, R., & Balan, L. (2019). Misclassification in travel surveys and implications to choice modeling: application to household auto ownership decisions. *Transportation*, *46*, 1467–1485.

Pukelsheim, F. (2006). *Optimal design of experiments*. SIAM.

Rao, C. R. (1992). Information and the accuracy attainable in the estimation of statistical parameters. In *Breakthroughs in statistics* (pp. 235–247). Springer.

Schennach, S. M. (2004). Estimation of nonlinear models with measurement error. *Econometrica*, *72*(1), 33–75.

Schennach, S. M. (2016). Recent advances in the measurement error literature. *Annual Review of Economics*, *8*, 341–377.

Schennach, S. M. (2019). Convolution without independence. *Journal of Econometrics*, *211*(1), 308–318.

Shao, M., Xie, C., & Sun, L. (2021). Optimization of network sensor location for full link flow observability considering sensor measurement error. *Transportation Research Part C: Emerging Technologies*, *133*, Article 103460.

Shi, Y., Boudouh, T., & Grunder, O. (2019). A robust optimization for a home health care routing and scheduling problem with consideration of uncertain travel and service times. *Transportation Research Part E: Logistics and Transportation Review*, *128*, 52–95.

Shivaswamy, P. K., Bhattacharyya, C., & Smola, A. J. (2006). Second order cone programming approaches for handling missing and uncertain data. *Journal of Machine Learning Research*, 1283–1314.

Sinha, A., Namkoong, H., Volpi, R., & Duchi, J. (2017). Certifying some distributional robustness with principled adversarial training. arXiv preprint arXiv:1710.10571.

Stopher, P., FitzGerald, C., & Xu, M. (2007). Assessing the accuracy of the sydney household travel survey with GPS. *Transportation*, *34*, 723–741.

Sun, H., & Xu, H. (2016). Convergence analysis for distributionally robust optimization and equilibrium problems. *Mathematics of Operations Research*, *41*(2), 377–401.

Sungur, I., Ordónez, F., & Dessouky, M. (2008). A robust optimization approach for the capacitated vehicle routing problem with demand uncertainty. *Iie Transactions*, *40*(5), 509–523.

Train, K. E. (2009). *Discrete choice methods with simulation*. Cambridge University Press.

Vapnik, V. (1999). *The nature of statistical learning theory*. Springer science & business media.

Wang, X.-F., & Wang, B. (2011). Deconvolution estimation in measurement error models: the r package decon. *Journal of Statistical Software*, *39*(10).

Wang, Y., Zhang, Y., & Tang, J. (2019). A distributionally robust optimization approach for surgery block allocation. *European Journal of Operational Research*, *273*(2), 740–753.

Xu, H., Caramanis, C., & Mannor, S. (2009). Robustness and regularization of support vector machines.. *Journal of Machine Learning Research*, *10*(7).

Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., & Jordan, M. (2019). Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning* (pp. 7472–7482). PMLR.